Model Uncertainty and Measures of Inequality of Opportunity*

Giovanni Bernardo[†] **Andros Kourtellos**[‡] **Chih Ming Tan**[§]

December 3, 2024

*We thank Steven Durlauf, Pablo Mitnik, Le Wang, Ömer Özak, Elena Falcettoni, Matheus Bandeira, Xuejuan Su, and participants of the 2023 Stone Center for Research on Wealth Inequality and Mobility Conference on New Methods to Measure Intergenerational Mobility, 2022 Southern Economic Association conference, CRETE 2023, the Canadian Economic Association conference for numerous helpful comments and suggestions. Tan thanks the Greg and Cindy Page Faculty Distribution Fund for financial support. Bernardo acknowledges the support of the National Operational Programme of the Italian Ministry of Education, University, and Research under the project REACT-EU (18-G-15333-5). The authors are also grateful to Max Propst for providing superb research assistantship.

[†]Department of Law, University of Naples "Federico II", Via Marina Nuova 22, 80133 Naples, Italy, email: giovanni.bernardo@unina.it.

[‡]Department of Economics, University of Cyprus, P.O. Box 537, CY 1678 Nicosia, Cyprus, email: andros@ucy.ac.cy.

[§]Department of Economics and Finance, Nistler College of Business and Public Administration, University of North Dakota, 3125 University Ave, Stop 8369, Grand Forks, ND 58202-8369, USA, e-mail: chihming.tan@und.edu.

Abstract

We argue that normative disagreements about which factors constitute circumstances in the equality of opportunity literature lead to substantial model uncertainty in empirical analyses. To systematically address this issue, we employ linear Bayesian Model Averaging (BMA) methods. Using data from the 2011 and 2019 waves of the European Union Statistics on Income and Living Conditions covering 31 EU countries, we estimate measures of absolute and relative inequality of opportunity. We compare our linear BMA findings to those using popular regression tree-based machine learning models including Bayesian Additive Regression Trees, Random Forest, and Extreme Gradient Boosting. Our findings show that ignoring model uncertainty and nonlinearity can substantially overstate IOp estimates across countries. Notably, linear BMA and Random Forest yield more conservative and robust estimates compared to traditional linear approaches, with Random Forest also emerging as the most reliable tree-based method.

Keywords: Equality of opportunity, Great Gatsby Curve, model uncertainty, Bayesian model averaging, regression trees, machine learning.

JEL Classification Codes: C3, C14, C24, C51, D85

1 Introduction

Following Roemer (1993, 1996, 1998), the concept of inequality of opportunity (IOp) has increasingly become the central ideal of distributive justice. The IOp perspective distinguishes between the determinants of an individual's outcomes; such as income, education, or health; see, a survey by Pignataro (2012), that are viewed as being under the control of the individual ("effort") and those that are not under her control ("circumstances"). The fundamental question posed by the IOp perspective is to what extent individual agency, decision-making, choices, and effort-factors within an individual's control-can overcome the constraints of circumstances in shaping individual outcomes. The societal aggregation of these outcomes then begets notions of inequality. Inequality in outcomes is justifiable when it is the result of individuals who share the same circumstances making different choices (e.g., exerting different levels of effort). However, inequalities that derive from differences in factors that are outside of the individual's control (i.e., circumstances) are not justified and should be mitigated by or compensated for via public policy.

While this approach has deep normative appeal because of its conceptual clarity, in practice, substantial issues arise when taking the concept to data. First, the extent to which the determinants of outcomes are observable may vary across contexts, and that may determine the principles that govern public compensatory efforts. The IOp literature distinguishes between two main perspectives, namely ex-ante and ex-post.[1]

From the ex-ante perspective, there is equality of opportunity if differences in circumstances do not matter for an individual's opportunity set. The focus of this perspective is therefore centered on inequality that arises from differences in circumstances and 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 (Van De Gaer, 1993; Lefranc et al., 2009; Peragine and Serlenga, 2008). Among the ex-ante perspective, Ferreira and Peragine (2016) distinguish two different approaches, "between-types inequality" (see for instance, Bourguignon et al. (2007); Checchi and Peragine (2010)) and "direct unfairness" (Fleurbaey and Schokkaert, 2009; Hufe et al., 2018).

[1] See, Brunori et al. (2013), Ferreira and Peragine (2016) and Roemer and Trannoy (2016) for excellent surveys of the recent literature.

From the ex-post perspective, there is equality of opportunity if individuals who exert the same level of effort receive the same outcomes. The ex-post approach assumes that compensation occurs after efforts are realized, and hence the requirement is that all individuals exerting the same level of effort receive the same outcomes, regardless of their circumstances (e.g., Roemer (1993); Fleurbaey and Peragine (2013)). Compensatory policies would then need to properly identify different degrees of effort and then be targeted at equalizing the outcomes of individuals exerting the same degree of effort. The ex-post approach can be divided into "within-tranches inequality" that was applied, for instance in Checchi and Peragine (2010) and Aaberge et al. (2011), while the concept of "unfair inequality" was proposed by Fleurbaey and Schokkaert (2009).

It is clear that the ex-ante and ex-post perspectives have different information requirements. The ex-ante perspective only requires the researcher to observe circumstance variables, but the ex-post perspective requires both circumstances and effort variables to be observed. The extent to which either case can be fulfilled depends on the context (and data quality). Like much of the empirical literature, this paper will focus on an applied example from the ex-ante perspective. Nevertheless, the methodological critiques that follow apply equally to both cases.

Perspectives on redistributive justice rely crucially on the proper estimation of measures of IOp. Our main contribution in this paper will be to point out that a standard critique of empirical work appears to be particularly relevant to the IOp context. A key issue with the standard approach to estimating measures of IOp ignores model uncertainty and bases inference on a single model with a particular set of circumstances, assuming that the model was correctly specified. This is a particularly unrealistic assumption in the IOp context because there exists substantive normative disagreement in the literature regarding which factors constitute circumstances and which constitute effort.

As Ferreira and Peragine (2016, p.5) quoting Barry (1992) point out:

> "In general, a conservative (or "right-wing") interpretation of the theory tends to limit the scope of circumstances, and attribute much to individual choice, whereas a progressive (or "left-wing") interpretation tends to recognize the importance of individual responsibility in principle, but to downplay the ability of individuals to make real choices: in this view, individual choices tend to be considered as completely determined by social circumstances."

In fact, a recent display of such normative disagreements about what factors constitute circumstances can be seen in the context of the U.S. Supreme Court's rejection of race-based affirmative action in college admissions:

> "In his concurring opinion, [Justice Clarence] Thomas directly engaged with [Justice Ketanji Brown] Jackson, one of the court's most liberal members, and the only other Black justice. In her view, he wrote, "almost all of life's outcomes may be unhesitatingly ascribed to race (The Washington Post, June 29, 2023)."[2]

It should be noted that it may well be the case that these disagreements are entirely reasonable or even founded in theory. Take, for instance, the following suggestion by Roemer and Trannoy (2016) :

> "...if one views all actions of the child as due to either nature or nurture, both of which are beyond his/her control, by hypothesis, before the age of consent, then one should simply take the child's educational accomplishments at the age of consent as a circumstance with respect to determining outcomes in later life."[3]

They argue that a child should not be able to claim accomplishments (positive or otherwise) below a certain age (18 years being the age of consent) and that all such factors should comprise the child's circumstances. While standard models in the social interactions literature (see, Durlauf and Ioannides (2010) for a recent survey) that seek to identify neighborhood or peer group effects explicitly posit interdependence between circumstances (the choices made by members of one's social group) and individual decision-making, the literature in no way suggests a complete lack of agency on the part of the individual for her outcomes. Instead, the dependencies generate probabilistic outcomes, not deterministic ones driven by either nature or nurture as suggested above. Not all children in disadvantaged environments will choose to drop out of school, engage in teen pregnancy, etc.

It is unclear if disagreements over which factors that affect an individual's outcomes are within their control and which are not (and to what extent) will ever be satisfactorily resolved. Methodologically, what these disagreements highlight is the existence of substantial *model uncertainty* in how the collective literature views the effect of circumstances on

[2]https://www.washingtonpost.com/politics/2023/06/29/affirmative-action-supreme-court-ruling/

[3]Hufe et al. (2017) similarly argue that a child should not be held responsible for any of their accomplishments below 16 years of age.

outcomes. There is therefore a need to obtain estimates for IOp that are sensitive to differences in normative perspectives. Conceptually, what this means is that we need to shift the discussion from focusing on IOp measures that are based on particular models and instead to focus on estimates for IOp that are contingent on a well-constructed model space. Hence, while individual proponents in this discourse may advocate for particular circumstances variables as being the "correct" focal points of analysis, the idea is to expand the discussion to encompass a broad (but potentially contentious) collective viewpoint of what the field would consider the set of applicable circumstances. In this formulation, a specific model within the model space represents a particular subset of circumstance variables. While individual proponents may agree or disagree with any given model, the union of all such models within the model space collectively represents the "consensus" (for lack of a better term) of the field regarding what constitutes circumstances.

The main contribution of this paper is therefore to highlight the need to obtain estimates for inequality of opportunity that are robust to model uncertainty. We propose to address systematically the issue of model uncertainty using Bayesian model averaging (BMA) approaches (e.g., Raftery et al. (1997)).[4] Unlike the standard approach that conditions on a single model, model averaging approaches construct estimates conditional on a model space which is composed of the set of all possible combinations of the universe of predictors suggested by all the relevant theories of equality of opportunity. The idea of model averaging is to "integrate out" the uncertainty over models across the model space by computing the average of model-specific estimates weighted by the posterior model probabilities that capture the evidentiary support of each model given the data.

It is important to state upfront what our approach does not do. Our paper should not be viewed as an attempt to resolve normative disagreements amongst proponents of opposing views. As we alluded to above, such a resolution is not something that is in-principle achievable due to the fact that, in many cases, such disagreements are essentially value-based and may not be amenable to change. Instead, our target is someone who is

[4]Model averaging approaches have been extensively applied in the economic growth literature. In many ways, the economic growth literature in the 1990's and 2000's was struggling with similar issues as the IOp literature. Multiple so-called "fundamental determinants" of growth; e.g., economic institutions, political institutions, geographic conditions, ethnic fractionalization, etc., were proposed, but inference was typically based on specific models where the inclusion or exclusion of alternative perspectives and the employment of particular proxy variables were largely arbitrary. This led growth researchers; see, for example, Brock and Durlauf (2001), Fernández et al. (2001), and Sala-i Martin et al. (2004), to propose model averaging methods that directly addressed such ambiguity or "model uncertainty".

agnostic about which circumstance variables truly are outside of the control of the individual. This constitutes the clearest connection to how we have recast the issue of normative disagreements as a problem of model uncertainty. And, in particular, it provides justification for our choice of non-informative priors under BMA (which we discuss in Section 3.1.2). The issue of model uncertainty shifts the focus of our analysis from individual models that capture the particular normative views of various proponents to the model space of all conceivable circumstance variables. Because the model space encompasses all of these individual models, it essentially houses the basis for forming a consensus viewpoint for estimating measures of IOp from the point of view of the broad literature.

Model averaging allows us to operationally estimate such a consensus measure of IOp under particular prior specifications that do hold substantive content. Our non-informative priors ultimately address a reader who is willing to entertain equally the validity of each model in the model space. This may not be true if the reader herself has a particular ideological bent. In that case, she might place higher prior weights on some models in the model space and lower weights on others. There is, in fact, a large literature on prior elicitation in the Bayesian literature (see, the recent critical review by Falconer et al. (2022)). However, it is beyond the scope of this paper to conduct such elicitations for the range of possible ideological persuasions/political preferences that would typify the population of readers.

In our analysis, we consider both linear and nonlinear model averaging methods. While traditional methods in model averaging predominantly focus on linear models, our analysis extends to incorporate nonlinear models as well.[5] As we discuss in Section 3 below, while the standard approach to the construction of IOp measures relies on a linear model, the focus of IOp measures on variations in the opportunity sets faced by *types* suggest that nonlinear models; specifically, tree-based methods, may be good alternatives for comparison. A type is essentially defined in terms of a particular partition of the support of circumstance variables. Tree-based models employ the data to endogenously partition the support of circumstance variables to form types and are therefore natural extensions to the linear approach. The capability of tree-based approaches to model deep nonlinearities and intricate interactions in a nuanced and data-driven way makes them exceptionally suitable for our study on inequality

[5]Various theoretical models underpinning the Great Gatsby Curve, including those emphasizing family investments, skills, social influences, political economy, and culture, indicate that the relationship between an individual's socio-economic outcomes and their circumstances is not inherently linear; see, Durlauf et al. (2022).

of opportunity.

We consider three tree-based approaches: Bayesian Additive Regression Trees with Model Averaging (BART-BMA), Conditional Random Forests (C-Forest), and Extreme Gradient Boosting (XGBoost). In BART-BMA, a greedy search algorithm is utilized to find predictive splitting rules, concentrating on constructing models from a sum of trees based on these predictive splits. Following this, BMA is applied to average across these models. C-Forest, which was also employed by Brunori et al. (2023) in the IOp context, builds an ensemble of independent trees using bootstrapping and random feature selection, averaging their predictions to reduce variance. In contrast, XGBoost sequentially builds trees, each correcting the residual errors of the previous ensemble (dependent process), focusing on bias reduction through boosting. We describe these tree-based approaches in Section 3.2 with more detailed technical discussion in the Online Appendix; Bernardo et al. (2024).

We investigate the implications of model uncertainty in both the linear and tree regression settings on estimates of IOp using data for 31 countries from both the 2011 and 2019 waves of the European Union Statistics on Income and Living Conditions (EU-SILC). Previous studies on IOp have primarily focused on the 2011 data. Our findings affirm the crucial importance of accounting for model uncertainty in the construction of IOp measures (both absolute and relative). It is tempting when there is a lack of consensus on model specification, to simply throw all available variables into the linear model. In fact, there is very little overall support for the "kitchen sink" model (i.e., the one that includes all the variables in the model space) which possessed (posterior) evidentiary weight that ranked it outside the top 10000 models under BMA. When comparing measures of IOp derived under Linear-BMA with those obtained using a linear "kitchen sink" model, we find that the "kitchen sink" estimates tend to significantly overstate the degree of inequality of opportunity experienced across our sample of countries.

In terms of our tree-based models, we first compared the out-of-sample prediction performance of the three approaches. We found that C-Forest consistently dominated both BART-BMA and XGBoost in terms of mean square prediction error. When compared to Linear-BMA, C-Forest tended to deliver more conservative estimates for both absolute and relative IOp across the set of EU countries for both the 2011 and 2019 waves of the data although the ordering of countries by IOp varies slightly between the two approaches. XGBoost generally tended to produce the largest estimates for IOp measures across the set of EU countries.

Qualitatively, all of our approaches paint a very similar story regarding the IOp experiences of EU residents: (i) there exists wide variation in IOp experiences across the set of EU countries with residents of Nordic countries generally facing lowest IOp while those of Southern and Eastern European countries experiencing the highest, (ii) there were substantive changes in IOp across the set of EU countries from 2011 to 2019 wth some countries showing improvement in terms of lowering IOp and other countries regressing on those measures, and (iii) the strongest and most consistent predictors of IOp were circumstance variables associated with social class; specifically, parental education and occupation.

However, the results for (iii) are nuanced. While social class variables were significant contributors to IOp for the vast majority of EU countries, other variables displayed varying levels of importance for different countries. A notable special case was Sweden which experienced a remarkable deterioration in IOp from 2011 to 2019 going from being one of the lowest IOp countries in the EU to becoming a country that ranked of the middle of distribution (in less than a decade). For Sweden (and some of the other Nordic countries as well) immigrant status of the respondent and/or her parents (described by their place of birth being either outside of Europe or the EU) became increasingly important predictors of IOp by 2019. These findings underscore the absence of a universal "one size fits all" model capable of explaining inequality of opportunity across all countries and reinforces the imperative need to consider model uncertainty when estimating IOp.

The subsequent sections of the paper are structured as follows: Section 2 discusses the conceptual framework underlying inequality of opportunity. This is followed by Section 3, which details our approach to empirical implementation. Section 4 is dedicated to discussing the data used in our analysis. In Section 5, we present our findings. Finally, the paper concludes with the last section, synthesizing our key insights and implications

2 Conceptual Framework

The ex-ante approach to inequality of opportunity (IOp) is rooted in the idea of 'types' that group individuals based on shared circumstances beyond their control. This framework emphasizes the role of circumstances in shaping advantages or outcomes while isolating the role of effort, for which individuals are deemed responsible.

The general framework can be represented as a finite mixture model, which links the distribution of an outcome Y_i to types, T_g, where $g = 1, .., n_g$. In this framework, types are defined by partitioning the population into mutually exclusive groups, T_g, where all individuals in a type share identical circumstances C_i. Specifically, each type, T_g, represents a group of individuals with identical values for a K-dimensional circumstance vectors $C_i = (C_{1i}, C_{2i}, \ldots, C_{Ki})'$, while allowing for variation in outcomes due to differences in individual effort.

The overall distribution of outcomes is therefore modeled as a finite mixture model where

$$F(y) = \sum_{g=1}^{G} \pi_g F_g(y), \tag{1}$$

where π_g is the probability of an individual belonging to type, T_g, and $F_g(y)$ is the outcome distribution for type, T_g. The opportunity set for a type can be summarized by characteristics of $F_g(y)$, such as the conditional mean $\mu_g = E(Y_i|T_g)$. The type-specific mean outcome Y_i^* is then

$$Y_i^* = \mu_g = E(Y_i|T_g), \tag{2}$$

representing the opportunity set for type, T_g. Deviations of Y_i from Y_i^* capture effort or unobserved individual characteristics. From this perspective, inequality of opportunity is captured by differences in opportunity sets across types. Perfect equality of opportunity implies $F_g(y) = F(y)$ for all g, whereas deviations indicate the presence of IOp.

To quantify IOp, we use inequality indices applied to the counterfactual distribution of Y_i^*, which represents the outcome if only circumstances determined Y_i. IOp can be defined in absolute or relative terms. The absolute inequality of opportunity index (absolute IOp) is defined as the inequality in Y_i^*

$$\text{Absolute IOp} = I(Y_i^*), \tag{3}$$

where $I(\cdot)$ is an inequality measure, such as the Gini coefficient[6]. We adopt the Gini coefficient as our benchmark inequality measure due to its status as the typical measure in the

[6]It should be noted that the relationship between circumstances and absolute IOp is one that is definitional in the sense that the construction of absolute IOp (the measure of inequality of opportunity) is based on the definition of types. There is no inherent causal relationship that is being claimed between circumstances and absolute IOp measures in the sense of, for instance, a one standard deviation change in a particular circumstance variable leading to a change in inequality of opportunity of a particular effect size.

inequality of opportunity (IoP) literature and its balanced approach to assessing disparities across the distribution. As highlighted by Brunori et al. (2023), the choice of an inequality index is inherently normative, reflecting varying sensitivities to inequality at different points of the distribution. The Gini coefficient provides a middle ground, offering equal sensitivity across the entire distribution without disproportionately emphasizing any specific segment. In contrast, measures like the Mean Logarithmic Deviation (MLD) focus more on inequalities at the lower end, reflecting a stronger emphasis on the most disadvantaged. Nevertheless, alternative measures are described in Online Appendix E, and their results are discussed in Section 5.4 to ensure robustness.

Relative IOp compares inequality due to circumstances to total inequality in realized outcomes Y_i

$$\text{Relative IOp} = \frac{I(Y_i^*)}{I(Y_i)}. \tag{4}$$

Relative IOp provides a normalized measure of the contribution of circumstances to overall inequality, making it particularly useful for cross-country comparisons, see Brunori et al. (2019).

3 Empirical Implementation

Building on the conceptual framework outlined in Section 2, we now describe the empirical implementation of the ex-ante compensation approach to estimating inequality of opportunity (IOp). To operationalize the finite mixture model in (1), we employ two regression-based methodological approaches: linear methods and tree-based methods.

Let $Z_i = (Y_i, C_i')'$ represent the combined vector of income and circumstances for observation i, and consider a random sample $\mathcal{Z} = (Z_1, \ldots, Z_n)$. Then, the model can be expressed as

$$Y_i = m(C_i) + e_i, \quad i = 1, 2, \ldots, N, \tag{5}$$

where $m(C_i) = E(Y_i|C_i)$ is the regression function mapping the K-dimensional vector of circumstance variables $C_i = (C_{1i}, C_{2i}, \ldots, C_{Ki})'$ to the expected outcome, and e_i represents the residual error, assumed to satisfy $E(e_i|C_i) = 0$.

This framework allows us to predict Y_i^*, the counterfactual outcomes determined solely by circumstances, which are used to estimate IOp measures as discussed in Section 2.

However, the choice of $m(C_i)$ is not straightforward due to model uncertainty, which arises from competing theories about the relevant set of circumstance variables and their interactions.

To address this challenge, we employ two classes of methods: (i) Linear Bayesian Model Averaging (Linear-BMA) and (ii) three advanced tree-based machine learning methods approaches. The following subsections provide a detailed discussion of these approaches each offering complementary strengths in addressing model uncertainty and capturing the complexity of IOp estimation.

3.1 Linear Methods

3.1.1 Linear Regression

Linear regression models are a traditional approach in the literature for estimating the relationship between circumstances and outcomes. However, selecting the appropriate set of circumstance variables and the correct model specification poses challenges due to model uncertainty. One straightforward method to address model uncertainty is the *Kitchen Sink* (KS) model, which includes all available circumstance variables in the regression:

$$Y_i = C_i'\beta + e_i, \tag{6}$$

where β is the vector of coefficients and e_i is the regression error term, $E(e_i|C_i) = 0$.

By including all potential circumstance variables, this model minimizes omitted variable bias. However, the least-squares (LS) coefficient estimators may suffer from high variance due to multicollinearity and overfitting, making it a poor choice for prediction. Overfitting can lead to poor out-of-sample predictive performance, as the model captures noise along with the underlying relationship.

Another approach is to select the model based on information criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria balance model fit and complexity by penalizing the number of parameters to prevent overfitting. Models with lower AIC or BIC values are preferred. By comparing these criteria across different model specifications, we can select a model that balances goodness-of-fit with parsimony. We focus on BIC as it imposes a heavier penalty for complexity, especially

in large samples. As we discuss below, BIC-based model selection effectively identifies the posterior mode model under BMA when using diffuse priors.

However, relying on a single selected model may not fully account for model uncertainty, and the inference might be sensitive. This sensitivity reflects the concept of "fragility," as described by Leamer (1978). In the context of inequality of opportunity, similar concerns arise. Just as Brock and Durlauf (2001) introduced the notion of openness in economic growth–arguing that model uncertainty renders coefficient estimates fragile–the estimation of inequality of opportunity is also subject to fragility due to model uncertainty.

Fragility implies that the estimated effect of a particular circumstance on outcomes can change dramatically in magnitude, lose statistical significance, or even switch signs depending on which other variables are included or excluded in the regression equation. More importantly, in the context of inequality of opportunity, model uncertainty affects not just the estimation of the coefficients in (6) but also the estimation of IOp measures themselves since these measures are derived from the predicted outcomes based on circumstance variables. A consequence of not accounting for model uncertainty would therefore be that the estimates of IOp measures may vary significantly depending on the chosen model, potentially leading to misleading conclusions about the extent of inequality due to circumstances. By systematically accounting for model uncertainty, methods like BMA help mitigate the fragility of both coefficient estimates and IOp predictions, leading to more robust and trustworthy findings.

3.1.2 Linear-BMA

To systematically address model uncertainty in estimating inequality of opportunity, we employ linear Bayesian Model Averaging (Linear-BMA). Unlike traditional methods that rely on a single prespecified model, Linear-BMA considers a comprehensive model space $\mathcal{M}^L$ consisting of all possible combinations of the K circumstance variables.

In Linear-BMA, each model $\mathcal{M}_m^L \in \mathcal{M}^L$ represents a particular subset of circumstance variables included in the regression. Linear-BMA addresses model uncertainty by averaging over all models in the space $\mathcal{M}^L$, weighting each model by its posterior probability. These weights quantify the degree of support for each model based on the data, allowing for a data-driven aggregation of estimates.

Formally, the posterior distribution of the coefficients β given the data $\mathcal{D}$ is

$$\mu(\beta|\mathcal{D}) = \sum_{m \in \mathcal{M}^L} w_m \mu(\beta|\mathcal{M}_m^L, \mathcal{D}), \tag{7}$$

where $w_m = \mu(\mathcal{M}_m^L|\mathcal{D})$ is the posterior probability of model $\mathcal{M}_m^L$ and $\mu(\beta|\mathcal{M}_m^L, \mathcal{D})$ is the posterior distribution of β under model $\mathcal{M}_m^L$.

The model-averaged estimator of β is the posterior mean

$$\widehat{\beta}_{\text{BMA}} = \sum_{m \in \mathcal{M}} w_m \widehat{\beta}_m, \tag{8}$$

where $\widehat{\beta}_m$ is the LS estimator of β under model $\mathcal{M}_m$.

By combining results from multiple models, Linear-BMA effectively mitigates the uncertainties inherent in relying on a single model. The efficacy of this approach lies in its ability to provide more robust estimations, inferences, and predictions compared to traditional model selection methods. As demonstrated in studies by Raftery and Zheng (2003) and Porwal and Raftery (2022), BMA "averages out" model-specific uncertainties inherent in relying on a single model, leading to more robust estimates and predictions.[7]

The standard way to conduct inference in the context of BMA is to rely on the posterior probability of inclusion (PIP) for each regressor. The PIP is computed as the sum of the posterior probabilities of all the models that contain that variable, and is, in that sense, a measure of variable importance.[8]

An important aspect of the linear-BMA is the specification of priors over both parameters and models. As we noted in the Introduction, our aim is to produce model averaged "'consensus" estimates for IOp measures for a reader who is agnostic about the models in the model space; i.e., one who is not taking a firm stance one way or the other

[7]The efficacy of BMA in providing accurate estimations and inferences has been demonstrated in previous studies (e.g., Raftery et al. (1997), Raftery and Zheng (2003)). In particular, under the assumption of correct specification in the sense that the practical distribution is equal to the prior distribution over parameters, Raftery and Zheng (2003) were able to derive a set of optimality theorems for BMA including that the BMA estimator minimizes MSE among point estimators. Further, they show in a simulation that the performance of BMA is robust to even large deviations of the prior specification from the practical distribution.

[8]Following Eicher et al. (2011), a PIP value less than 50% implies insufficient evidence for an effect, while PIP values ranging from 50% to 75% suggest weak evidence for an effect. Similarly, PIP values ranging from 75% to 95% indicate positive evidence for an effect, while PIP values ranging from 95% to 99% imply strong evidence for an effect. Finally, PIP values ranging from 99% to 100% indicate decisive evidence for an effect.

regarding normative disagreements over circumstance variables. To effectively address model uncertainty under this context, we therefore wish to avoid injecting undue prior information in the construction of equations (7) and (8) above. That is, we aim to use priors that are as non-informative as possible.

For the parameter priors, we follow standard practices in the literature in coding non-informativeness. We assume improper uniform priors for the intercept α and the error variance σ^2 $\mu(\alpha) \propto 1$, $\mu(\sigma^2) \propto \sigma^{-2}$. For the coefficients β, we employ Zellner's g-prior $\mu(\beta|\sigma^2) = N\left(0, \sigma^2 g(C_m' C_m)^{-1}\right)$, where g is a hyperparameter controlling the degree of shrinkage. Our benchmark prior is the *Unit Information Prior* (g-UIP), which sets $g = N$, the sample size. This choice represents the information from one observation and is considered relatively non-informative.

For the model priors, we use the *Uniform Model Prior* as our baseline, assigning equal probability to all models $\mu(\mathcal{M}_m) = \frac{1}{|\mathcal{M}^L|}$, where $|\mathcal{M}^L|$ is the total number of models in the model space. This reflects an agnostic stance regarding which variables should be included in the true model for estimating inequality of opportunity.[9] Details on the implementation of Linear-BMA can be found in Online Appendix A.

3.2 Tree-Based Machine Learning Methods

While the standard approach to obtaining IOp estimates relies on linear models, a more flexible approach allows the data to uncover the specific combinations of circumstances that define relevant types, as characterized in equation (1). Tree-based methods partition the observations into subgroups based on the values of the predictors, effectively creating partitions within the support of circumstance variables. These methods naturally handle nonlinearities and high-order interactions without requiring the explicit specification of interaction terms. By recursively splitting the data into subsets based on predictor variables, tree-based models minimize a criterion, typically the residual sum of squares in regression tasks. This data-driven partitioning makes tree-based models particularly suitable for uncovering the distinct opportunity sets that define types in the context of inequality of

[9]However, we acknowledge potential issues with this choice, such as the irrelevance of independent alternatives (IIA) problem highlighted by Brock and Durlauf (2001), which arises when the model prior ignores interdependencies among variables. To address these concerns, we explore alternative model priors that account for such interdependencies, including hierarchical priors and dilution priors in the Online Appendix A.

opportunity.

3.2.1 BART-BMA

We start with the Bayesian Additive Regression Trees with Model Averaging (BART-BMA) approach proposed by Hernández et al. (2018). BART-BMA naturally extends Linear-BMA to tree-based methods by building on the Bayesian Additive Regression Trees (BART) method of Chipman et al. (2010) and incorporating model averaging over an ensemble of trees. This method not only effectively captures nonlinearities and interactions but also systematically addresses model uncertainty by averaging over multiple tree models weighted by their posterior probabilities. In particular, in BART-BMA, each sum of trees is considered a single model $\mathcal{M}_m^T$ within the model space, $\mathcal{M}_m^T \in \mathcal{M}^T$. The model space consists of various sums of trees, and BART-BMA performs BMA over these models. The posterior probability of each model is used to weight its contribution to the final predictions, addressing model uncertainty systematically.

In BART-BMA, the regression function $m(C_i)$ is modeled as a sum of regression trees. Suppressing the subscript on model dependence $\mathcal{M}_m$ for simplicity, the model is expressed as:

$$Y_i = \sum_{t=1}^{T} g_t(C_i; \mathcal{T}_t, \Theta_t) + e_i, \quad e_i \sim N(0, \sigma^2), \tag{9}$$

where each g_t represents a regression tree. The function g_t is defined as:

$$g_t(C_i; \mathcal{T}_t, \Theta_t) = \sum_{\ell=1}^{L_t} \theta_{t\ell} \cdot \mathbb{I}(C_i \in R_\ell), \tag{10}$$

where $\mathbb{I}(C_i \in R_\ell)$ is an indicator function that equals 1 if C_i belongs to terminal node R_ℓ, and 0 otherwise.

In this formulation, $\mathcal{T}_t$ denotes the structure of the t-th tree, defining how the predictor space C_i is partitioned into L_t mutually exclusive regions, or terminal nodes or leafs, R_ℓ ($\ell = 1, 2, \ldots, L_t$). Each tree $\mathcal{T}_t$ is constructed through a sequence of decision rules (splits) based on the values of the predictor variables C_i, recursively dividing the data into subsets that are increasingly homogeneous with respect to the outcome variable Y_i. Θ_t denotes the vector of terminal node parameters, $\Theta_t = \{\theta_{t1}, \theta_{t2}, \ldots, \theta_{tL_t}\}$, representing the node-specific means of the outcome variable within each terminal node of tree t. Specifically, $\theta_{t\ell}$ represents

the mean of Y_i conditional on C_i being in terminal node R_ℓ, $\theta_{t\ell} = E(Y_i|C_i \in R_\ell)$.

Similar to Linear-BMA, BART-BMA provides measures of variable importance through *Posterior Inclusion Probabilities* (PIPs). The PIP for a circumstance variable reflects the probability that the variable is used as a splitting variable in the ensemble of trees across all models. Additionally, variable importance scores are computed based on the proportion of splits attributed to each variable, offering a relative measure of each variable's contribution to the model.

BART-BMA requires priors for the tree structures $\mathcal{T}_t$, the terminal node parameters Θ_t, and the error variance σ^2. For the tree structure priors, we specify that the probability that a node at depth d splits is given by $\kappa(1+d)^{-\lambda}$, where $\kappa \in (0,1)$ and $\lambda \geq 0$ are hyperparameters controlling the depth and shape of the trees. Typical values are $\kappa = 0.95$ and $\lambda = 2$, which encourage shallow trees and prevent overfitting. The terminal node parameters $\theta_{t\ell}$ are assigned conjugate normal priors: $\theta_{t\ell} \sim N(0, \sigma^2/\upsilon)$, where υ is a shrinkage parameter controlling the variance of the terminal node parameters. For the error variance σ^2, we assume an inverse gamma prior on the precision σ^{-2}, specifically $\sigma^{-2} \sim \text{Gamma}(\nu/2, \nu\xi/2)$, where ν and ξ are hyperparameters. Finally, uniform priors are assigned over models to reflect an agnostic stance regarding which tree structures are more probable.

BART-BMA employs an efficient algorithm that constructs the model space by generating multiple sets of trees through a greedy search for optimal splitting rules. It uses Bayesian model averaging to weigh these sets based on their posterior probabilities. The estimation involves starting with single-tree models and identifying optimal splits that minimize the residual sum of squares. Then, it iteratively adds trees to the model, updating residuals and searching for new splits that improve the fit. Occam's Window (Raftery et al., 1997) is applied to retain models within a certain threshold of the best model, based on the BIC. The use of Occam's Window helps to focus on models that offer a good balance between fit and complexity, mitigating overfitting. Finally, the retained models are averaged to obtain predictions and variable importance measures.

Detailed discussions on the priors, hyperparameter choices, and the algorithmic steps of the estimation procedure can be found in Appendix B.

3.2.2 Conditional Random Forest

Building on foundational work such as Classification and Regression Trees (CART) (Breiman et al., 1984) and Random Forests (RF) (Breiman, 2001), Conditional Random Forests (C-Forest) represent an advanced ensemble learning framework tailored for regression and classification tasks. Introduced by Hothorn et al. (2006), C-Forest refines the classical RF methodology by incorporating statistical tests to ensure unbiased variable selection while maintaining robust modeling of interactions. These features are particularly relevant for estimating inequality of opportunity, where intricate relationships among circumstance variables-such as family background and education-are critical for understanding outcomes.

In the RF framework, the regression function is represented as a regression tree: $m(c) = E(Y|C = c) = \mathcal{G}(c; \xi, \mathcal{Z})$, where $\mathcal{G}$ is the regression tree function mapping predictors c to the expected outcome, and ξ is auxiliary randomness drawn from a distribution Ξ, independent of the data $\mathcal{Z} = (Y, C)$, introducing variability into the construction of the tree. A regression tree captures the relationship between circumstances and outcomes by recursively partitioning the data into groups of observations. For any given individual i, the conditional mean of income is determined by the mean income of the group to which they belong.[10]

While individual trees can have low bias if grown sufficiently deep, they often suffer from high variance. Random Forests address this issue by constructing an ensemble of independent regression trees through bagging (Bootstrap Aggregation) with an adjustment to decorrelate the trees. Bagging involves constructing B regression trees using B bootstrapped training sets and averaging their predictions, which reduces variance and improves model stability. The decorrelation of trees is achieved by considering, at each split, only a random subset of S circumstances as split candidates from the full set of K circumstances. As a result, on average, $100(1 - K/S)\%$ of the splits will not include the strongest circumstance, allowing other circumstances a greater chance to influence the splits. This process reduces the dependence among trees, resulting in an average with lower variability and increased reliability. Biau (2012) and Scornet et al. (2015) demonstrated the consistency of the Random Forest estimator, while Wager and Athey (2018) showed it to be asymptotically normal.[11]

[10] For simplicity, the dependence of the tree on its structure $\mathcal{T}$ and terminal node parameters Θ is suppressed.

[11] Unlike BART-BMA, C-Forest does not require the assumption of error normality.

Specifically, RF is an ensemble estimator of B independent regression trees. Each tree is grown on a bootstrapped sample $\mathcal{Z}_b^* = (Z_{b1}^*, \ldots, Z_{bn}^*)$, drawn from the empirical distribution function $F_n(z) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(Z_i \leq z)$. The RF prediction at c_i is computed by averaging the predictions of all trees in the ensemble:

$$\widehat{m}^{RF}(c_i) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathcal{G}}_b^*(c_i; \xi_b^*, \mathcal{Z}_b^*), \tag{11}$$

where $\widehat{\mathcal{G}}_b^*(c_i; \xi_b^*, \mathcal{Z}_b^*)$ is the prediction of the b-th tree, and ξ_b^* denotes the randomness specific to the b-th tree, drawn independently from Ξ. In our case, bootstrap samples are constructed on average based on 60% of the observations for each tree. The remaining observations are considered as the out-of-bag (OOB) sample for the b^{th} tree, denoted by $\overline{\mathcal{Z}}_b^*$.[12]

C-Forests refine the splitting rules in RFs to address known biases and limitations of traditional tree-based methods. As noted by Breiman et al. (1984), splitting rules in traditional RFs are often biased toward covariates with many possible values or categories. C-Forests mitigate this issue by employing Conditional Inference Trees (CITrees), proposed by Hothorn et al. (2006), as the base estimator. This is important because it allows to handle predictors and outcomes on various scales, balancing the capture of complex interactions with model interpretability. This makes them particularly effective for tasks like inequality of opportunity estimation, where understanding intricate predictor relationships is critical. C-Forests improve upon traditional RFs by enhancing robustness and applicability across varied datasets.

Specifically, CITrees replace the heuristic criteria used in CART, such as Gini impurity or variance reduction, with a statistically rigorous approach. Splitting variables are selected through statistical hypothesis tests within a permutation testing framework, where the conditional distribution of test statistics measures associations between predictors and the response. Adjusted p-values guide the selection, with the most significant predictor chosen for splitting if a significant relationship exists; otherwise, the process stops. This ensures unbiased variable selection and reduces overfitting by avoiding splits influenced by noise or overly complex predictors.

The performance of C-Forests relies on careful tuning of key parameters that influence the construction of individual trees and the forest as a whole. A critical parameter is the

[12] When B is sufficiently large, OOB error is approximately equal to leave-one-out cross-validation error

significance level (α^*), which sets the threshold for statistical significance in the independence tests used for splitting. Lower α^* values enforce stricter criteria, leading to shallower trees with fewer splits, while higher α^* values permit deeper trees that capture more complex interactions.

Another important parameter is the number of predictor variables considered at each split ($\bar{K}^*$). Limiting this subset prevents highly correlated variables from dominating splits, ensuring a balanced exploration of predictors. Finally, the number of trees in the forest (T^*) controls the size of the ensemble. Increasing T^* reduces prediction variance by averaging across more trees, improving robustness. These parameters are often tuned to minimize the mean square OOB-error (MSE^{OOB}) using the OOB sample, where the OOB-error for the b^{th} tree is given by $\widehat{\varepsilon}_i^{OOB} = y_i - \widehat{\mathcal{G}}_b^*(c_i; \xi_b^*, \mathcal{Z}_b^*)$.

To evaluate variable importance, CFOREST evaluates variable importance by comparing the predictive accuracy of the model before and after permuting the values of a given variable. Specifically, for each variable, its values are randomly shuffled, breaking the relationship between that variable and the response, while all other variables remain unchanged. The decrease in predictive accuracy after this permutation, averaged across all trees and out-of-bag (OOB) samples, quantifies the importance of the variable. Variables causing larger accuracy drops are considered more important. A detailed explanation of the computation is provided in the online appendix.

Further details on the tree construction process, including the role of key parameters and their tuning, and variable importance scores are provided in Online Appendix C.

3.2.3 Extreme Gradient Boosting

The main idea of gradient boosting, introduced by Friedman (2001), is to enhance predictive accuracy by iteratively improving a model through the sequential addition of simple predictive models (weak learners) such as a small regression tree. Unlike bagging methods such as Random Forests, which build independent trees on bootstrapped samples, gradient boosting constructs each new tree to predict the residuals of the current ensemble, effectively addressing areas where the model performs poorly. This approach focuses on "learning slowly," as each iteration incrementally refines the model by concentrating on instances with larger prediction errors. Instead of relying on a single large tree, which risks overfitting, gradient boosting uses small, shallow trees, or weak learners, that individually have limited

capacity but collectively form a strong predictive model. By adapting to the errors of previous iterations, the algorithm progressively improves its accuracy and captures complex relationships in the data.

Extreme Gradient Boosting (XGBoost) is a highly efficient and scalable implementation of gradient boosting tailored for tree-based models (Chen and Guestrin, 2016). It introduces several algorithmic optimizations and regularization techniques that enhance both the speed and performance of gradient boosting, making it particularly well-suited for large-scale and high-dimensional data. In the context of the current study, XGBoost provides a powerful framework for modeling the complex relationships between circumstances and outcomes.

The XGBoost model predicts the outcome Y_i as a sum of R regression trees:

$$\hat{Y}_i = \sum_{r=1}^{R} \widehat{\mathcal{G}}_r(C_i), \tag{12}$$

where $\widehat{\mathcal{G}}_r(C_i)$ represents the fitted r-th regression tree. Each tree in the model is constructed to minimize a regularized objective function that balances the fit of the model to the data with its complexity, ensuring both accuracy and generalization.

The estimation is based on a penalized mean squared error criterion

$$\mathcal{L} = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 + \sum_{r=1}^{R} \Omega(\widehat{\mathcal{G}}_r), \tag{13}$$

where the term $\sum_{r=1}^{R} \Omega(\widehat{\mathcal{G}}_r)$ represents the regularization component, which penalizes model complexity to prevent overfitting and improve generalization. The regularization term $\Omega(\widehat{\mathcal{G}}_r)$ is defined as $\Omega(f_r) = \gamma L_r + \frac{1}{2}\lambda\|\omega_r\|^2$, where L_r is the number of leaves (terminal nodes) in the r-th tree, ω_r is the vector of predicted weights assigned to the leaves of r-th tree, and γ and λ are global regularization parameters controlling the trade-off between model complexity and training loss. The regularization term penalizes both the number of leaves to control tree depth and the magnitude of leaf weights, ensuring the trees remain simple and interpretable while avoiding overfitting.

The fitting process in XGBoost begins with an initial prediction, typically the mean of Y_i, denoted as $\widehat{\mathcal{G}}_0(C)$. The residuals, representing the errors of the initial model, are computed as $\widehat{\epsilon}_i^{(1)} = Y_i - \widehat{\mathcal{G}}_0(C_i)$. The first tree, $\widehat{\mathcal{G}}_1(C_i)$, is then fitted to these residuals to minimize

the objective function. This optimization leverages second-order approximations of the loss function, which utilize both the gradient and Hessian to determine the optimal adjustments. The gradient captures the direction and magnitude of the error, while the Hessian accounts for the curvature of the loss function, ensuring updates are scaled appropriately for more stable and efficient learning. Once the tree is fitted, the model is updated by adding the predictions from this first tree to the initial prediction, resulting in the updated prediction $\widehat{\mathcal{G}}_i^{(1)} = \widehat{\mathcal{G}}_0(C_i) + \widehat{\mathcal{G}}_1(C_i)$.

For subsequent iterations, the residuals of the current ensemble are recalculated as $\widehat{\epsilon}_i^{(t)} = Y_i - \widehat{\mathcal{G}}_i^{(t-1)}$, where $\widehat{\mathcal{G}}_i^{(t-1)}$ represents the prediction from the ensemble up to iteration $t-1$. The t-th tree, $\widehat{\mathcal{G}}_t(C_i)$, is fitted to these residuals, again using the gradient and Hessian to refine the model's predictions by efficiently targeting areas of greatest error. The updated model prediction at iteration t becomes $\widehat{\mathcal{G}}_i^{(t)} = \widehat{\mathcal{G}}_i^{(t-1)} + \widehat{\mathcal{G}}_t(C_i)$. This iterative process continues until a predefined number of trees, R, is reached, or the improvement in the loss function becomes negligible.

Finally, we employ the Gain metric to assess variable importance. Gain measures the improvement in the objective function that results from splits involving a specific circumstance. Specifically, it quantifies the reduction in the loss function achieved each time a variable is used to partition the data within the regression trees. Variables that accumulate higher Gain across all splits are deemed more influential, as they significantly contribute to minimizing the loss and enhancing the model's predictive accuracy.

Details on the tuning parameters, regularization parameters as well as specifics of the algorithmic implementation, are provided in Appendix D.

4 Data

Our empirical analysis is based on the European Union Statistics on Income and Living Conditions (EU-SILC) dataset. This dataset provides harmonized survey data on income, poverty, and living conditions, making it a valuable resource for statistics that compare income distribution and social inclusion from both cross-sectional and longitudinal perspectives. We specifically used the 2011 and 2019 wave of EU-SILC as them include a form that provides detailed information on the individual characteristics related to the education, employment, and income of the respondent's parents. Wave 2011 covers a representative

sample of households from 31 different European countries while Wave 2019 covers 29 different European countries[13]. Focusing on the data under consideration, the 2011 wave includes a minimum of approximately 4,000 observations in smaller countries such as Iceland and Ireland, and a maximum of 20,000 in larger countries like Italy. In contrast, the 2019 wave records a minimum of around 4,300 observations in countries like Ireland and Norway, and exceeds 17,000 in larger countries such as Italy, Spain, and Poland. These waves serve as a crucial resource for our research, allowing for a comprehensive examination of the impact of parental background and household financial situation on children's income.

Following the literature, we use equivalized disposable household income, adjusted for the square root of household size, as a proxy for the respondent's outcome variable. This measure encompasses various income sources, including wages, bonuses, overtime, commissions, farm and business income, income from professional practice, roomers and boarders, profits from capital investments, interests, dividends, and other income streams. Taxes included in the analysis cover regular taxes on wealth, inter-household cash transfers paid, income taxes, and social insurance contributions.

To ensure cross-country comparability of household income levels, we adjust for purchasing power differences using the Purchasing Power Parities (PPP) index for the relevant wave years (2011 and 2019). The PPP index accounts for price level indices and real expenditures across ESA aggregates, standardized to EU28=100 for each year. For temporal comparability, we deflate the data using the Harmonized Index of Consumer Prices (HICP, base year 2010), which captures changes in all-item price levels. Both PPP and HICP data are sourced from Eurostat.

To address outliers and non-positive incomes, we apply a standard winsorization procedure: non-positive incomes are set to 1, and incomes above the 99.5th percentile of the country-specific income distribution are capped at this threshold. Finally, we take the logarithm of the adjusted disposable household income to normalize its distribution. To ensure the representativeness of our inequality of opportunity estimates, we use individual cross-sectional weights when calculating IOp index.

Given that young individuals often experience income instability and significant year-to-year fluctuations, we restrict our sample to respondents aged 30 to 60 years. Table 2 provides descriptive statistics for equivalized disposable household income in euros.

[13]Wave 2011 accounts for EU27 countries, plus Iceland, UK, Norway, and Switzerland. Wave 2019 accounts for EU27 countries, plus Norway, Serbia, and Switzerland.

Concerning the personal characteristics of the respondents, we include gender and country of birth. Additionally, to account for potential life-cycle effects, we control for age and age square in robustness exercises described in Section 5.5 below. Our analysis examines three distinct sets of circumstances pertaining to the households of the respondents at the age of 14, as detailed in Table 1.

(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, family type, number of adults and children, and number of workers.

To address the contribution of individual characteristics and mitigate issues arising from variables with non-ordered categorical structures, variables with multiple non-ordinal categories-such as employment status or tenancy status-were transformed into binary dummy variables. Specifically, categories converted into binary dummy variables include the respondent's citizenship; the country of birth and citizenship of both parents; the occupational activity status of both parents; the father's type of occupation; and family type. In contrast, ordinal variables, such as education level or the household's financial situation, were retained in their original format. These ordinal variables include the education levels of both parents, the household's financial situation, the household's ability to meet end-of-month needs, the number of adults, the number of children, and the number of employed individuals in the household when the respondent was approximately 14 years old. Gender is captured through a dummy variable equal to 1 if the respondent is female. For tenancy status, we only consider ownership, represented by a dummy variable equal to 1 if the household owns the residence.

Variables related to the mother's managerial position and occupation type were excluded from the analysis due to a high proportion of missing values across several countries. Following Brunori et al. (2023), the father's employment type was classified into three broad categories: (i) high-skilled non-manual (managers, professionals, technicians, and associate professionals); (ii) low-skilled non-manual (clerical support workers, service and

sales workers, armed forces occupations); and (iii) skilled manual and elementary occupations (skilled agricultural, forestry, and fishery workers; craft-related trades workers; plant and machine operators; assemblers; and elementary occupations). Fathers categorized as unemployed, unknown, or deceased were grouped separately.

A specific adjustment was made for parental education levels to account for missing information, particularly when the respondent did not know their mother or father. As detailed in Table 1, including a "missing" category in an ordinal variable-ranging from 1 (lowest education level) to 4 (highest education level)-would compromise the variable's ordinal nature. To address this, missing parental education levels were replaced with the median level of education for the respective country, preserving the variable's ordinal structure. Summary statistics for all variables are provided in Appendix Table A.1.

5 Results

5.1 Performance Assessment

To assess the predictive performance of the models, we employ K-fold cross-validation with $K = 10$, which provides a robust framework for estimating out-of-sample performance. The dataset is randomly split into K roughly equal-sized folds, denoted as I_k $(k = 1, \ldots, K)$. For each fold, one subset is held out as the test set, while the remaining $K - 1$ folds form the training set. The model is then trained on the training folds to estimate the regression function $\widehat{m}^k(C_i)$, which is subsequently used to generate predictions $\widehat{Y}_i^k$ for the observations in the test set I_k. The residuals for the k-th fold are calculated as $\tilde{e}_i^k = Y_i - \widehat{Y}_i^k$, and the mean squared error for the fold is computed as: $\text{MSE}_k = \frac{1}{|I_k|} \sum_{i \in I_k} (\tilde{e}_i^k)^2$. This process is repeated K times, rotating the test set for each fold. The out-of-sample mean squared error (MSE^{OOS}) based on K-fold cross-validation is obtained by averaging the MSEs across all folds: $\text{MSE}^{OOS} = \frac{1}{K} \sum_{k=1}^{K} \text{MSE}_k$.

The results are reported in Table 3.[14] The table focuses on comparing the performance of Linear-BMA with three tree-based machine learning methods–C-Forest, BART-BMA, and

[14]The predictions and evaluations are based on the logarithmic transformation of income. An MSE of 0.25 in the log scale translates to an average prediction error of 0.5, indicating that income is over- or underpredicted by approximately 50% on average. This highlights the practical relevance of even small differences in MSE^{OOS} values across methods.

XGBoost–in predicting the logarithm of income for the 2011 and 2019 waves. To facilitate comparison, the best-performing model (in terms of lowest MSE^{OOS}) for each country and wave is shaded in green. C-Forest consistently emerges as the top performer across most countries and both waves, reflecting its ability to effectively model complex interactions and nonlinear relationships. In contrast, BART-BMA often yields the highest MSE^{OOS}'s across models, suggesting potential issues with overfitting or difficulties in capturing the data structure in this context. While in some cases Linear-BMA performs similarly it is generally outperformed by tree-based methods such as C-Forest and XGBoost. While XGBoost also delivers competitive results, its performance is slightly less dominant compared to C-Forest, highlighting the latter's robustness in this setting. These findings underscore the importance of selecting appropriate models depending on the nature of the data and the desired balance between computational efficiency and predictive accuracy.

5.2 Results for Linear Models

We first discuss our IOp results based on the Gini coefficient using linear models. Figure 1 presents results for absolute IOp and relative IOp for the set of EU countries for the 2011 wave. As discussed, the standard approach in the literature is to employ the linear KS model under any model space.[15] We therefore compare the KS model with two benchmark alternatives. Our benchmark findings are based on Linear-BMA, as discussed above. However, if a researcher had reservations about results derived from model averaging and preferred instead to select a model from the model space, then, the Bayesian approach would suggest the model with the highest posterior probability (which, under our benchmark priors, also corresponds to the model with the lowest BIC); i.e., the posterior mode model. Specifically, therefore, Figure 1 compares results from the KS model with the posterior mode model and with Linear-BMA.

We find that there are substantial variations in absolute and relative IOp across the set of EU countries based on the Gini coefficient. This is visually evident in Figures 3, panel A and panel B, which display the geographic distribution of absolute and relative IOp for the 2011 and 2019 waves, respectively. The results show that Mediterranean and Eastern European countries (e.g., Romania (RO), Bulgaria (BG), Spain (ES), Portugal (PT), and Greece (EL))

[15]The KS model is one of the potential models within the model space considered in the BMA, and it has a negligible posterior model probability for all parameter priors and model prior specifications that we considered.

have the highest levels of both absolute and relative IOp, while the Scandinavian countries (e.g., Sweden (SE), Norway (NO), Iceland (IS), Denmark (DK)) have the lowest levels. We will discuss the results for Sweden in a separate subsection below because of the nature of the data for that country as well as some evidence that there may be a case for "Swedish Exceptionalism" in the broader narrative on inequality of opportunity.

When comparing the KS results with the results from the posterior mode model and Linear-BMA, we find that the KS model tends to overstate the influence of circumstances on outcomes; i.e., report higher levels of both absolute and relative IOp, compared to the latter two approaches. The results for the posterior mode model and Linear-BMA, on the other hand, are relatively close to each other with the Linear-BMA results being typically more conservative for the lower inequality of opportunity countries. In this regard, therefore, accounting for model uncertainty, and not basing conclusions on a single a priori model, potentially mitigates biases and offers a more accurate picture of the state of inequality of opportunity across EU countries.

To get a sense of the cross-country differences in terms of magnitude it may be easiest to focus on the relative IOp results. In this case, we find that absolute IOp accounts for over 40% of income inequality for countries with the highest inequality of opportunity like Romania and Bulgaria (over 50% for the case of Romania). However, for low inequality of opportunity countries like the Nordic states (setting Sweden aside for now), absolute IOp accounts for less than 30% (in most cases, less than 20%) of income inequality.

When we compare our results from the 2011 wave to those for the 2019 wave, we find interesting changes in the levels of inequality of opportunity over time. Figure 1 show results for absolute IOp and relative IOp for the set of EU countries for the 2011 and 2019 wave while Figure 5 provides a comparison of the findings from 2011 to 2019 in terms of the changes in both IOp measures over those two waves. The main message from the 2019 findings in Figure 1 is the same in the aggregate from 2011; i.e., there remains wide variation in inequality of opportunity experiences across EU countries.

However, the range of inequality of opportunity experiences across the EU appears to have widened over time. From Figure 5, we see that some of the lowest inequality of opportunity countries in 2011 like Denmark and Finland have made improvements in terms of of reducing both absolute and relative IOp even further while some of the highest inequality of opportunity countries in 2011 like Romania, Bulgaria, and Spain have seen

their IOp measures worsen over time. There has also been considerable churning in the cross-country distribution of IOp measures over time. Some high absolute and relative IOp countries in 2011 such as Estonia (EE) and Hungary (HU) have made marked improvements in reducing inequality of opportunity over time while others like the Slovak Republic (SK) that had lower levels of both absolute and relative IOp in 2011 appear to have regressed.

5.2.1 Variable Importance

We now characterize the set of circumstance variables that appear to be most important in terms of predicting absolute IOp across the EU taking into account model uncertainty. Table 4 Panel A presents average posterior means and PIP for the Linear-BMA model for circumstance variables with the strongest evidence for variable importance across the EU countries for both the 2011 and 2019 waves. For example, around 60% of countries in 2011 and 57% of countries in 2019 reported a PIP of 0.75 or higher for the variable that described the educational attainment of the respondent's father. The average PIP for that variable across the EU countries was 0.81. For a complete set of results detailing all circumstance variables included in the Linear-BMA model across the EU, see Online Table A.7.

Our results suggest that the most important circumstance variables that determine inequality of opportunity across EU countries have to do with variables associated with social class; in particular, father's education and, to a lesser degree, mother's education. While these two latter variables came out on top for both 2011 and 2019, we also see evidence that other variables that speak to social class; i.e., whether the father's occupation was skilled (manual) and the financial situation of the family, took on greater importance in 2019. There is also some weaker evidence that whether the respondent was born in a non-EU country may play a role in determining inequality of opportunity.

These results are, however, at the aggregate level of the EU. When we dug into the specific circumstances that were of importance to individual countries, a more complex picture emerged. Tables 5 present heatmaps that show the 15 most important variables in terms of PIP for 2011 and 2019, respectively, across the set of countries. In these heatmaps, shades of red indicate the importance of each circumstance. The intensity of the red color increases as the PIP approaches 1, with darker red indicating higher importance. While variables that characterize social class - variables linked to parental educational attainment and occupation, and household financial situation - dominate the set of important

circumstance variables, we also see, especially, in the 2019 wave, that place of birth (within or outside the EU), whether for the respondent or her parents play an increasingly important role in determining inequality of opportunity for certain countries like Denmark, Estonia, Norway, and Sweden.

In summary, our findings expose a complex array of circumstances influencing inequality of opportunity across the EU. Factors related to social class are certainly significant contributors for the vast majority of countries. But, other variables also play varying roles of importance for specific countries as well. The variability of the set of circumstances across different countries suggests that there is no "one size fits all" model that explains IOp across the EU, thus highlighting the critical need to account for model uncertainty when investigating the diverse and complex nature of inequality of opportunity within the EU.

5.2.2 Swedish Exceptionalism

Perhaps no country has seen greater changes from 2011 to 2019 than Sweden. The results we reported for Sweden for 2011 require some caveats. To be consistent with the exercises we ran for the other EU countries, we used the same model space for Sweden. However, for the 2011 wave, Sweden had issues with missing observations that were not observed for the other EU countries. For example, for the variable describing the Occupation of the Father (as well as for parental education variables) a large share of the sample had missing observations. Out of an initial sample of approximately 5,400 respondents, more than 3,000 were excluded due to missing data. When we dropped variables with that suffered most severely from missing observations from the model space for Sweden, however, we obtained absolute and relative IOp estimates that were similar to Norway; i.e., Sweden still belonged to the set of very low inequality of opportunity Nordic countries.

Hence, the change in absolute and relative IOp measures for Sweden from 2011 to 2019 as seen clearly from Figure 5 remains remarkable. Sweden went from being one of the most equal countries in terms of opportunity to becoming a country resting in the middle of the pack (just below Cyprus) in less than a decade. Multiple studies have documented significant adverse changes in income and wealth inequality, poverty rates, and health outcomes since 2011–Pareliussen (2017)–especially when compared to other Nordic countries but these findings may rest crucially on how inequality in capital income is accounted for. There is some evidence that accounting for the treatment of capital income, inequality in Sweden may

actually have stabilized after 2010 after increasing during the Financial Crisis of 2008-09; see, Swedish Fiscal Policy Council (2024). This stabilization is notable especially since the post-2011 period saw large migrant inflows in around 2015.

While trends in income inequality may have stabilized between 2011 and 2019, our findings suggest that inequality of opportunity continued to worsen during this period. Specifically, in Sweden, the analysis of variable importance (see Table 5) reveals a notable shift: by 2019, parental birthplace variables–such as having parents born outside the EU or Europe–emerged as the most significant determinants of absolute IOp, surpassing all other factors. In contrast, these variables were relatively unimportant in 2011. This shift underscores how recent large-scale migrant inflows have reshaped the dynamics of inequality of opportunity within Swedish society.

5.2.3 Robustness to Prior Specifications

Finally, we document the robustness of the above findings to different parameter priors and model prior specifications. Figure A.1 of the Online Appendix shows absolute IOp estimates for the set of EU countries under 10 different prior specifications. The pattern of absolute IOp estimates across the set of EU countries obtained under different prior specifications appear to be largely similar. In Online Appendix Table A.2, we report the correlation matrix between the absolute IOp indices computed using the 10 prior specifications. The results show that the absolute IOp indices are strongly correlated with one another. We further examine the out-of-sample prediction performance of Linear-BMA under these 10 prior specifications. As shown in Tables A.3 and A.4 in Appendix, the UIP prior specification is either the best performing or very close to the best performing in terms of MSE. These results show therefore that Linear-BMA estimates for absolute IOp indices are qualitatively robust to the choice of priors. They also justify our selection of the UIP specification as the benchmark model.

5.3 Results from C-Forest and other Tree-based Approaches

As established in Section 5.1, C-Forest consistently emerges as the best-performing method. We now turn to examining how its IOp estimates compare to those derived from Linear-BMA. By leveraging data-driven sample splits, C-Forest uncovers nuanced heterogeneities

in opportunity sets that linear methods might fail to detect. This section investigates the impact of model choice on the estimated absolute and relative IOp, highlighting the additional insights offered by tree-based methods.

Figure 2 presents a comparison of absolute and relative IOp estimates for EU countries using C-Forest and Linear-BMA for the 2011 and 2019 waves. Across both years, C-Forest generally produces more conservative IOp estimates compared to Linear-BMA, although the ranking of countries varies between the two approaches. Additionally, Panel B of Figure 5, highlights that both methods largely agree on the direction of change in absolute and relative IOp between 2011 and 2019. However, a notable exception is Cyprus (CY), where C-Forest indicates an improvement in reducing inequality of opportunity, while Linear-BMA suggests a slight deterioration.

Table 4 also finds that C-Forest and Linear-BMA are in general agreement in terms of variable importance. Comparing Table 4 panel A to Table 4 to Panel C, we find that the social class variables – primarily father's education, but also mother's education – are the most important determinants of absolute IOp. Overall, the results for C-Forest are relatively similar to those obtained under the benchmark Linear-BMA approach; see Online Appendix Figure A.4. Comparing the results for C-Forest with those from alternative tree-based approaches, we do see some disagreements in terms of the pattern of both absolute and relative IOp indices across the EU countries. The estimates for C-Forest are consistently the most conservative across the different approaches while those for XGBoost tend to report higher levels of IOp.

Since C-Forest demonstrates the best out-of-sample performance among all methods, its results are likely the most reliable. Nonetheless, to provide a comprehensive view, we present the range of absolute and relative IoP estimates across all methods in Online Appendix Figure A.6, panels A and B. Moreover, the most conservative estimates (represented by the green dots) are generally produced by C-Forest. Online Appendix Table A.8 details all circumstance variables analyzed with C-Forest across all countries.[16]

[16]Full results for BART-BMA and XGBoost are also available upon request.

5.4 Alternative Inequality Indices

As noted in Section 2, our analysis thus far has been based on specifying the inequality index $I(\cdot)$ using the Gini coefficient. However, the literature has explored alternative specifications for $I(\cdot)$. A common alternative is the Mean Logarithmic Deviation (MLD; see Checchi and Peragine (2010), Ferreira and Gignoux (2011)). Other widely used measures include the Theil index and Generalized Entropy measures (GE; see Brunori et al. (2023)). As Ferreira and Peragine (2016) highlight, different inequality indices reflect varying levels of aversion to inequality across different segments of the opportunity distribution. Consequently, it is reasonable to expect that the quantitative aspects of our findings may vary depending on the chosen specification for $I(\cdot)$.

In fact, we find quite a substantial variation when we compare IOp measures derived using different specifications for the inequality index $I(\cdot)$. As shown in Figure 4 panel A and B, IOp measures derived using the alternative inequality indices to Gini tend to produce more conservative estimates relative IOp under both Linear-BMA and CRF. Hence, it is possible to argue that the Gini derived inequality index potentially overstates the degree of inequality of opportunity across countries compared to alternative approaches. Nevertheless, the qualitative pattern of findings across countries are virtually identical and our findings are robust in this sense.

In this sense, it reaffirms the basic message that regardless of the choice of specification for the inequality index, accounting for model uncertainty and being cognizant of possible nonlinearities in the estimation of IOp measures is crucial in informing our views on the degree of inequality of opportunity that exists across the set of countries as well as its variation.

5.5 Potential Lifecycle Bias

As expounded in the data section, this study incorporates the age variable, recognizing its pivotal role in assessing potential life-cycle bias. The scholarly discourse on this matter is extensive, with varying perspectives. Noteworthy contributions by scholars such as Checchi and Peragine (2010), Pistolesi (2009) and Suárez Álvarez and Jesús López Menéndez (2018) treat age as a pertinent circumstance, while others, exemplified by Hufe et al. (2018), Brunori et al. (2023) and Ferreira and Peragine (2016)), do not adopt this perspective.

Mitnik et al. (2022) posits cogent arguments against compensating for age, contending that individuals, in principle, traverse all age categories and should naturally experience compensation over time. However, the drawback lies in the potential overestimation of the overall IOp estimates if age exhibits a correlation with circumstances included in the analysis. To mitigate this concern, Mitnik et al. (2022) suggests concentrating on four consecutive cohorts observed in the same year (when they are in their late 30s)), effectively removing age as a meaningful influence.[17]

In this paper, we follow Checchi et al. (2016) and Brunori et al. (2023) by not including age as a circumstance, arguing that cross-sectional income disparities across age groups tend to balance out over an individual's life cycle. To ensure the robustness of our findings, we conduct additional exercises using residualized income data adjusted for age. Online Table A.10 confirms that results for linear models remain consistent, while Online Figure A.11 demonstrates similar robustness for tree-based methods. These analyses reinforce the validity of our conclusions, which remain substantially unaffected, aligning with the findings of Brunori et al. (2023).

5.6 Robust IOp and the Great Gatsby Curve

The recent literature on inequality and economic mobility has focused on the cross-country relationship between the Gini coefficient (a measure of inequality of outcomes) and the intergenerational elasticity of earnings (IGE; a measure of intergenerational persistence in outcomes or economic immobility); i.e., the GGC (Krueger, 2012; Corak, 2013).[18] We show the GGC for our set of countries in Figure 6. Like for the larger set of countries plotted in Corak (2013), we find a positive relationship between the Gini coefficient and the IGE.[19]

The fact that GGC shows a positive relationship between the Gini coefficient and IGE, suggests that countries that pursue greater equality of opportunity (measured by higher

[17]Nonetheless, the practical application of Mitnik et al. (2022)'s methodology to our dataset encounters challenges, primarily due to its potential to significantly reduce the pool of observations, resulting in a loss of representativeness. As illustrated in Online Appendix Figure A.9, if we assume to select a sample of people aged between 37 and 40, the number of observations would sharply decline, especially in smaller countries such as Sweden, Denmark, Norway or Lithuania.

[18]Durlauf et al. (2022) provide a survey of work surrounding the GGC and the reasons for its prominence in recent empirical work.

[19]The data for IGE comes from Corak (2013) which only has IGE data for 10 EU countries: Denmark, Finland, France, Italy, Germany, Norway, Spain, Sweden, Switzerland, United Kingdom)

mobility or lower IGE) experience lower levels of inequality of outcomes (lower Gini coefficient values). It is of interest, therefore, to see how our measures of absolute IOp which measure ex ante inequality of opportunity correlate with the GGC variables; i.e., the Gini coefficient and IGE.[20]

We start by characterizing the pseudo-GGC relationship between our robust IOp measures with the Gini coefficient. Figure 7 shows the relationship between the Gini coefficient and absolute IOp measures for both linear as well as tree-based models. We find a positive relationship between the Gini coefficient and our measures of absolute IOp suggesting that increasing income inequality is correlated with higher levels of inequality of opportunity. In terms of the strength of the association, our benchmark Linear-BMA and Posterior Mode model results are very similar and they describe a stronger (steeper) relationship between the Gini coefficient and absolute IOp than that for either the KS model or C-Forest.

Figure A.8 in the Online Appendix illustrates the relationship between absolute IOp and the IGE for both linear and tree-based models. Similar to the above, when we compare absolute IOp with the IGE, we find that the relationship is positive across all approaches suggesting that countries with higher levels of inequality of opportunity also experience lower income mobility. Like with the relationship between absolute IOp and the Gini coefficient, the association here also appears stronger (i.e., steeper) for our benchmark Linear-BMA and Posterior Mode models compared to the KS model or C-Forest.

These findings suggest that different modeling approaches yield varying sensitivities in capturing the link between absolute IOp and GGC variables. Nevertheless, the magnitude of the differences appears to be relatively small. Overall, while our results consistently indicate a positive relationship between inequality of opportunity and both income inequality and intergenerational persistence of earnings, the limited number of observations poses significant challenges for drawing robust and generalizable conclusions about the GGC within the EU context. Consequently, while our study supports the notion that higher levels of income inequality are intricately linked to higher levels of both inequality of opportunity and immobility, definitive claims about the strength and nature of the GGC should be made judiciously. Future research incorporating a larger and more diverse set of countries, as well as longitudinal data, as Brunori et al. (2013) posit, would be instrumental in elucidating the relationship between income inequality and economic mobility.

[20]Here, we follow Brunori et al. (2013) who, to our knowledge, were the first to do so.

6 Conclusion

This study has rigorously examined the impact of model uncertainty on the estimation of IOp across 31 EU countries from 2011 to 2019. By employing both Linear-BMA and C-Forest, alongside other tree-based methods such as BART-BMA and XGBoost, we have elucidated the complex determinants and evolving patterns of IOp within diverse European contexts.

Our analysis reveals that social class factors, particularly parental education and occupation, consistently emerge as the most significant contributors to IOp. These findings are robust across different modeling approaches and remain pivotal in both the Linear-BMA and C-Forest frameworks. Additionally, variables related to household financial status and parental birthplace have gained prominence over time, highlighting the dynamic nature of socio-economic influences on IOp.

A critical insight from our study is the superior performance of C-Forest among tree-based methods (with Linear-BMA coming in as a close second), demonstrating its enhanced ability to capture complex interactions and nonlinear relationships compared to BART-BMA and XGBoost. This methodological advancement is particularly salient in cross-country analyses, where traditional linear models like the "kitchen sink" approach tend to overstate IOp measures. The kitchen sink model, which includes all available circumstance variables, often lacks robustness and fails to account for model uncertainty, leading to inflated estimates of IOp. In contrast, Linear-BMA and C-Forest provide more conservative and reliable estimates, underscoring the importance of advanced modeling techniques in accurately capturing the multifaceted nature of inequality.

A particularly noteworthy finding pertains to Sweden, which has undergone a remarkable shift from being one of the most egalitarian societies in terms of IOp to occupying a middle-ranking position within less than a decade. This transition is likely influenced by large-scale migration and changing societal dynamics, as evidenced by the increasing importance of parental birthplace variables in 2019. Such country-specific insights emphasize the absence of a universal model for IOp, reinforcing the necessity of accounting for model uncertainty and heterogeneity in IOp estimation.

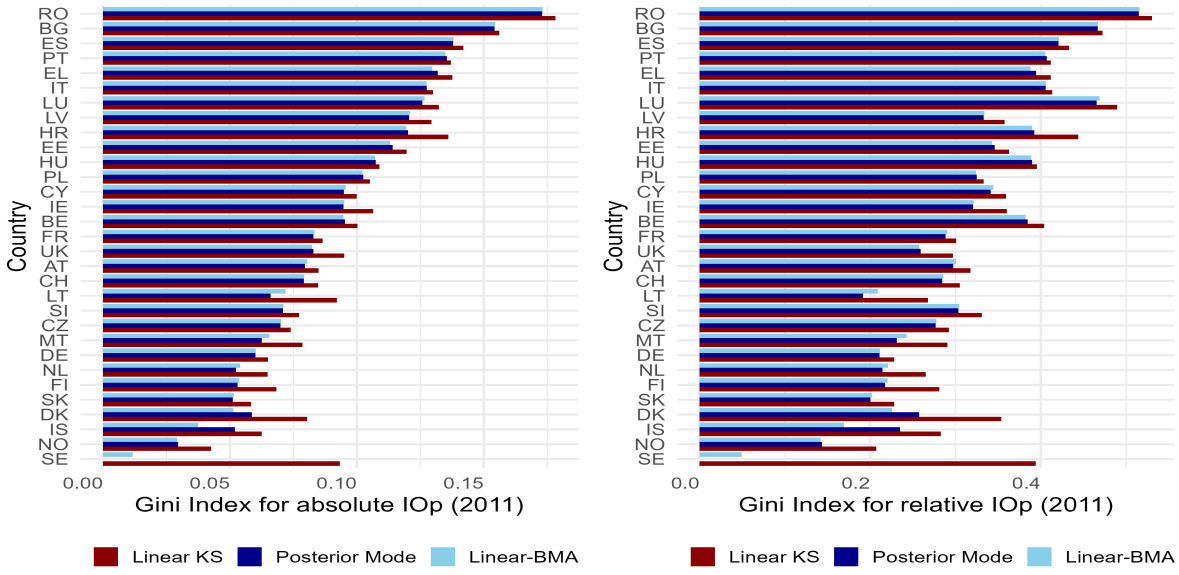
In conclusion, our research highlights the intricate interplay of socio-economic variables influencing inequality of opportunity and demonstrates the critical role of advanced modeling techniques in capturing these complexities. By addressing model uncertainty through

Linear-BMA and leveraging the robustness of C-Forest, we offer a more comprehensive and accurate estimation of IOp across diverse European settings. These methodological advancements not only enhance academic understanding but also provide valuable guidance for policymakers aiming to design effective interventions to mitigate unjust inequalities rooted in circumstances beyond individual control.

Figure 1: IOp Estimates - Linear Models

This set of figures presents Inequality of Opportunity estimates derived from the Linear-KS, Posterior Mode, and Linear-BMA approaches for Waves 2011 and 2019. Panel (A) refers to Wave 2011, while panel (B) corresponds to Wave 2019. The figures on the left show IOp in absolute terms, and the figures on the right display IOp in relative terms. All estimates are based on the Gini coefficient as the measure of inequality.

(A) Wave 2011



(B) Wave 2019

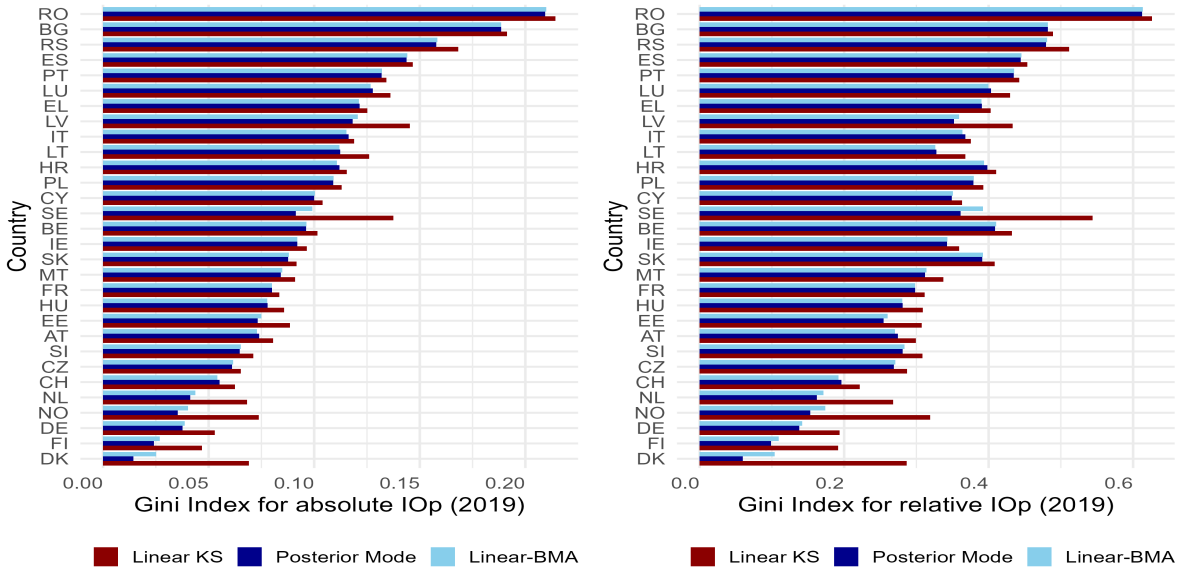
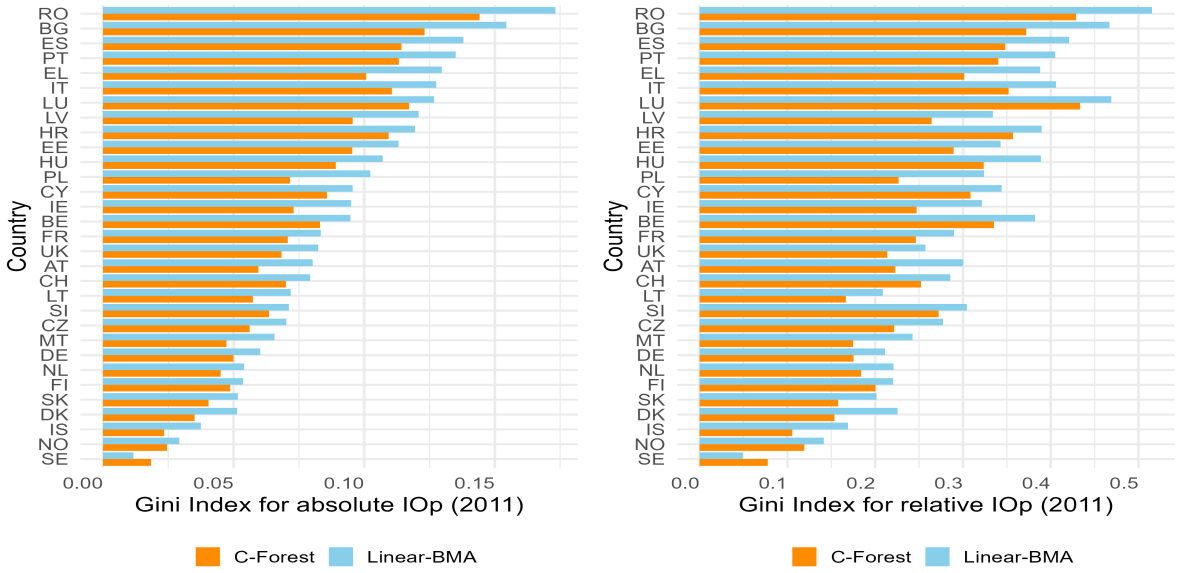


Figure 2: IOp Estimates - Linear-BMA and C-Forest

This set of figures presents Inequality of Opportunity (IOp) estimates derived from the Linear-BMA and C-Forest approaches. Panel (A) refers to Wave 2011, while panel (B) corresponds to Wave 2019. The figures on the left show IOp in absolute terms, while the figures on the right display IOp in relative terms. All estimates are based on the Gini coefficient as the measure of inequality.

(A) Wave 2011



(B) Wave 2019

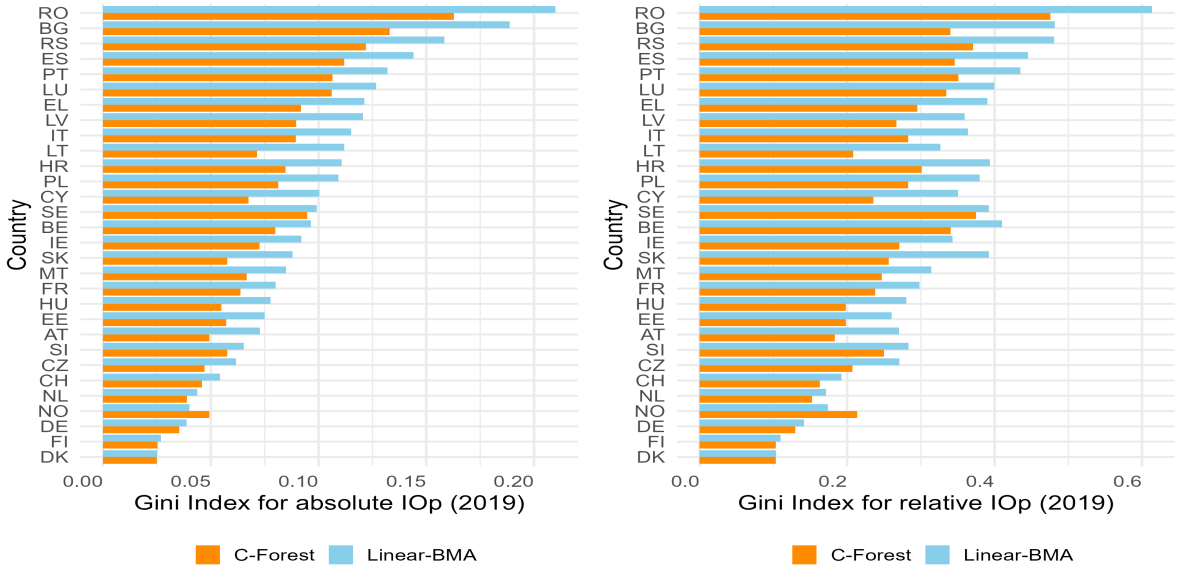
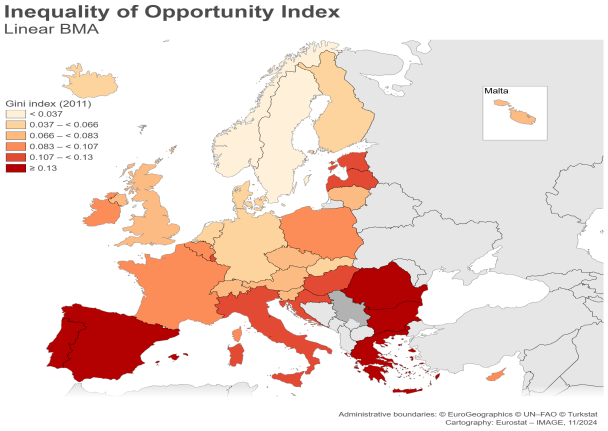


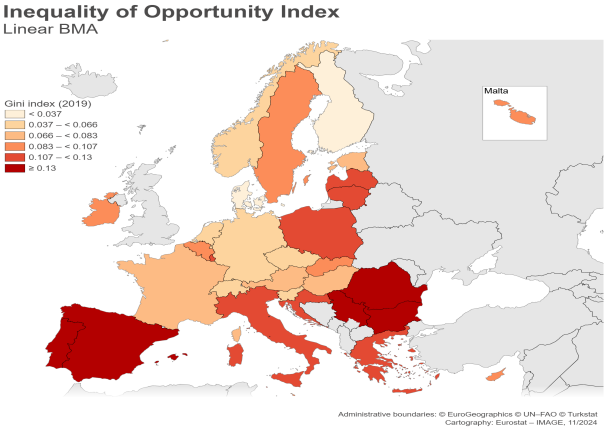
Figure 3: Inequality of Opportunity Estimates Maps

This set of figures presents the geographical distribution based on the Linear-BMA and C-Forest relative IOp estimates based on the Gini coefficient across Europe.

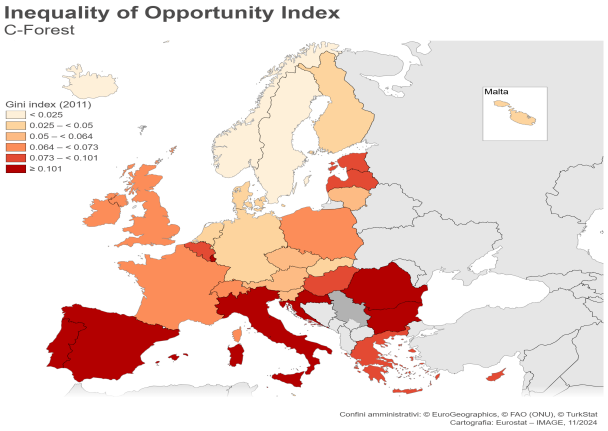
(A) Linear-BMA: Wave 2011



(B) Linear-BMA: Wave 2019



(C) C-Forest: Wave 2011



(D) C-Forest: Wave 2019

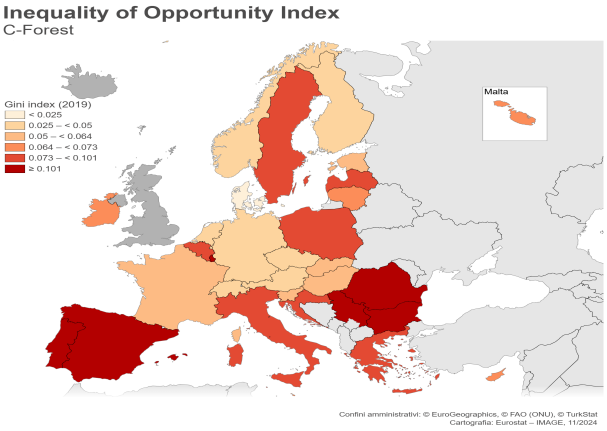
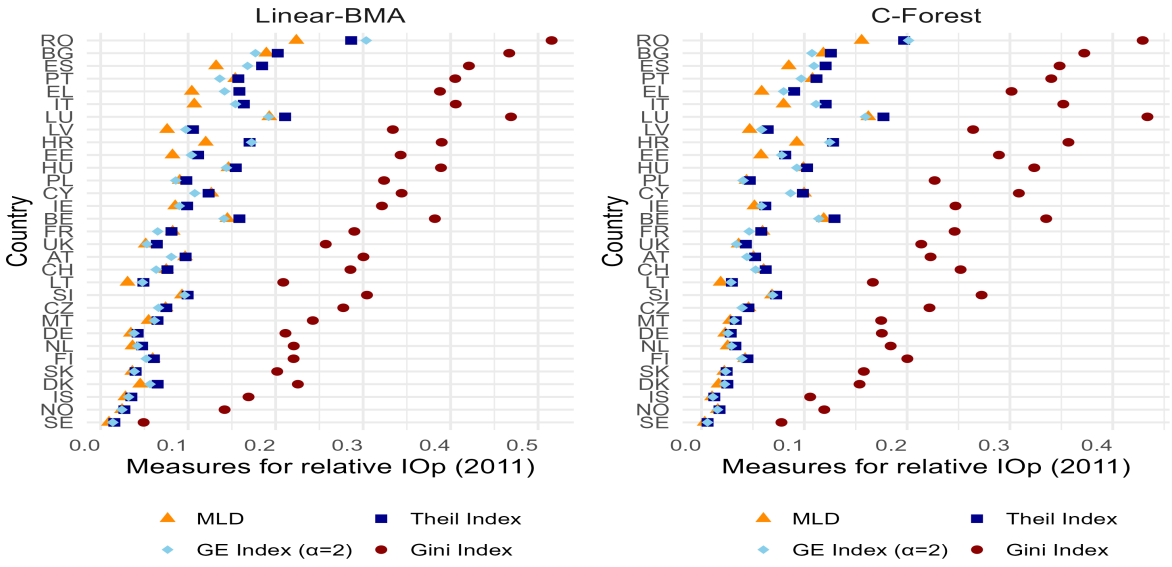


Figure 4: IOp Estimates - Various Inequality Indices

This panel of figures presents a comparative analysis of various measures of Inequality of Opportunity) using the Mean Log Deviation (MLD), Theil Index, Generalized Entropy with $\alpha = 2$, and the Gini index. Panel (A) refers to Wave 2011, while panel (B) corresponds to Wave 2019. The figures on the left show IOp in absolute terms, while the figures on the right display IOp in relative terms. All indicators are shown in relative terms.

(A) Wave 2011



(B) Wave 2019

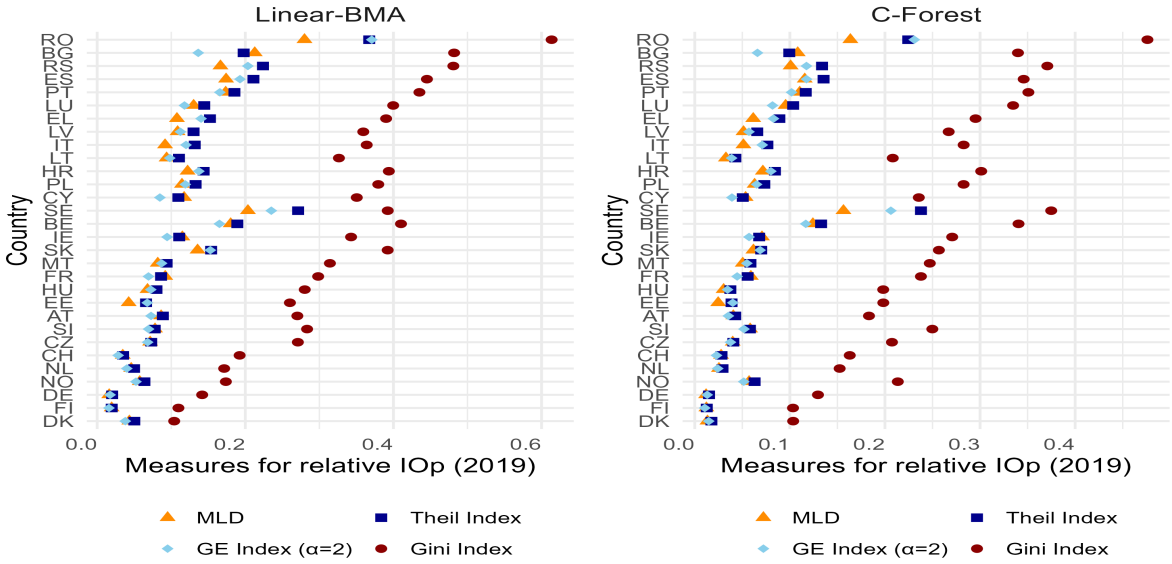
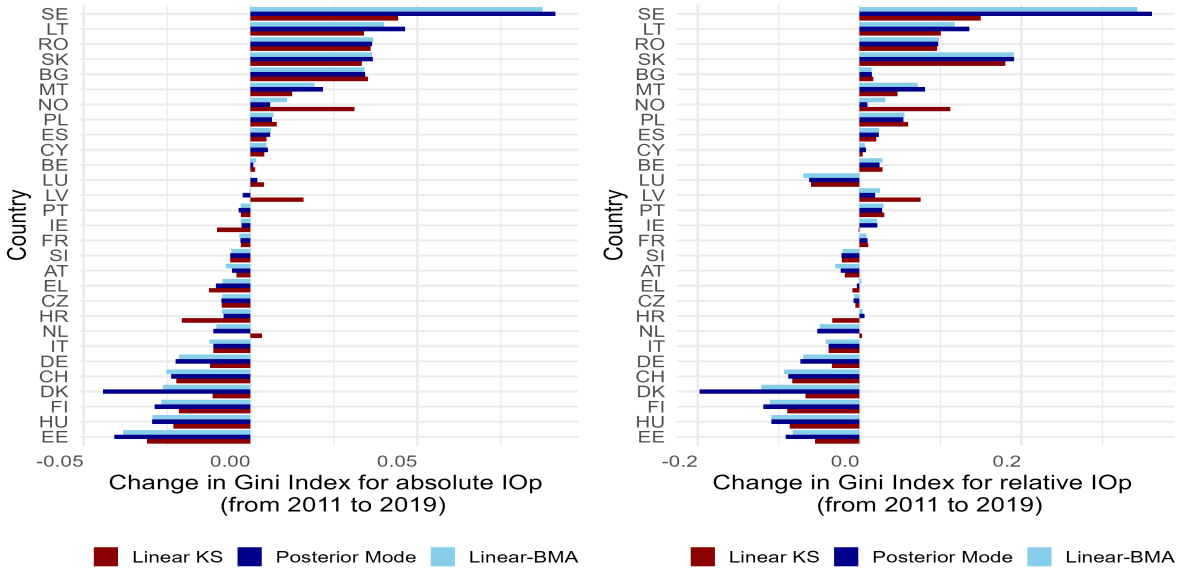


Figure 5: Change in IOp Between Waves 2011 and 2019

This panel of figures illustrates the changes in IOp estimates based on the Gini coefficient, with linear models displayed in panel (A) and Linear-BMA and C-Forest in panel (B).

(a) Linear Models



(b) Linear-BMA and C-Forest

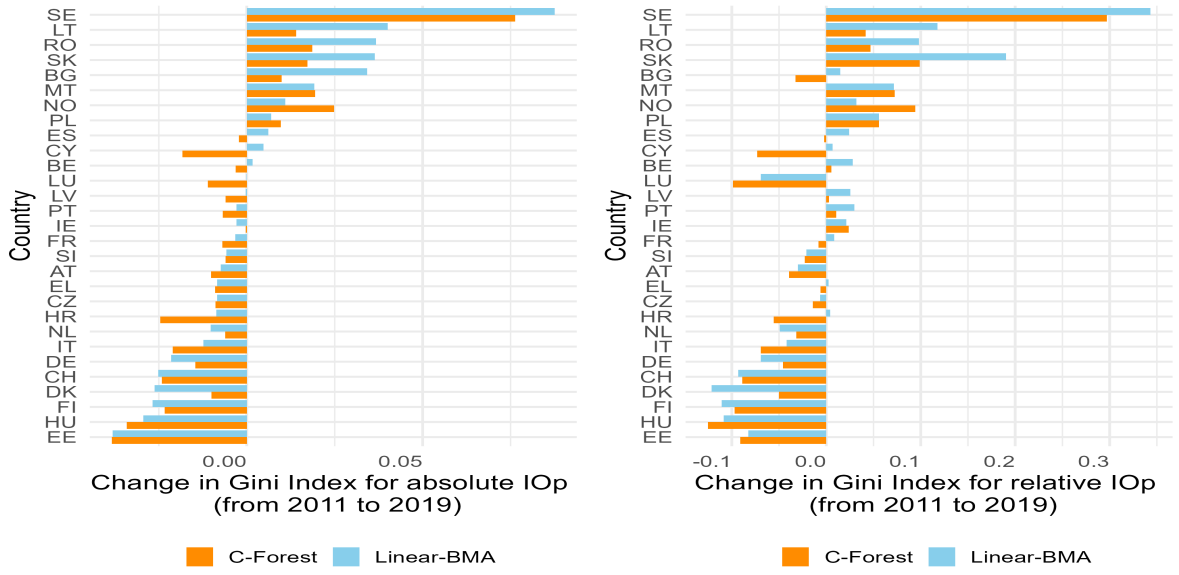


Figure 6: Great Gatsby Curve

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

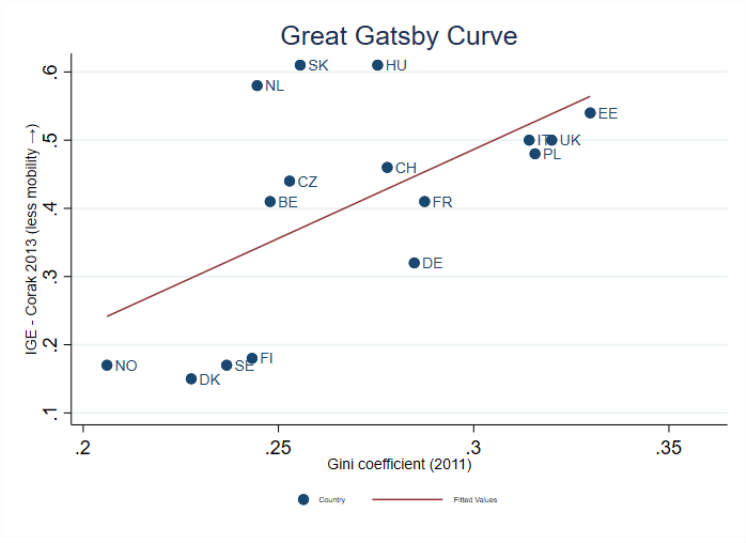
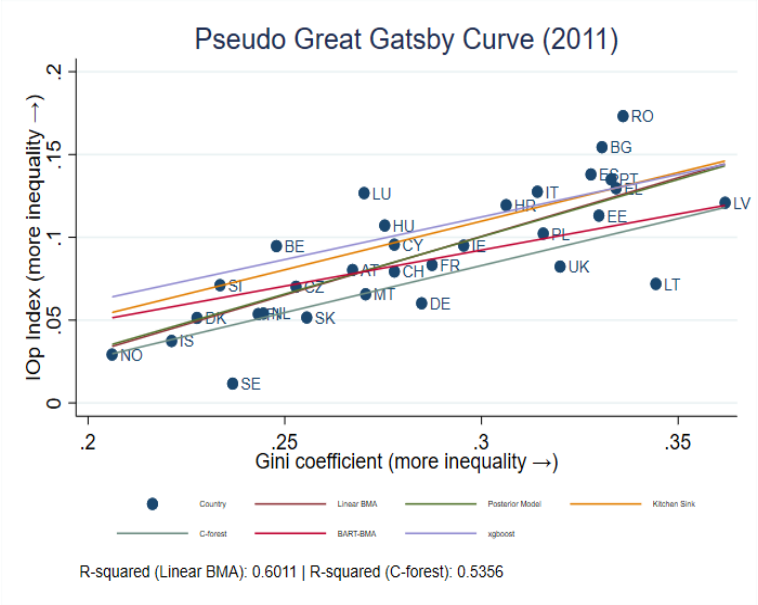


Figure 7: Pseudo Great Gatsby Curves

This figure provides the relationship between Inequality of Opportunity and Income Inequality.IOp estimates derived from Linear models and Tree-based Methods. Panel (A) refers to Wave 2011, while panel (B) corresponds to Wave 2019.

(A) Wave 2011



(B) Wave 2019

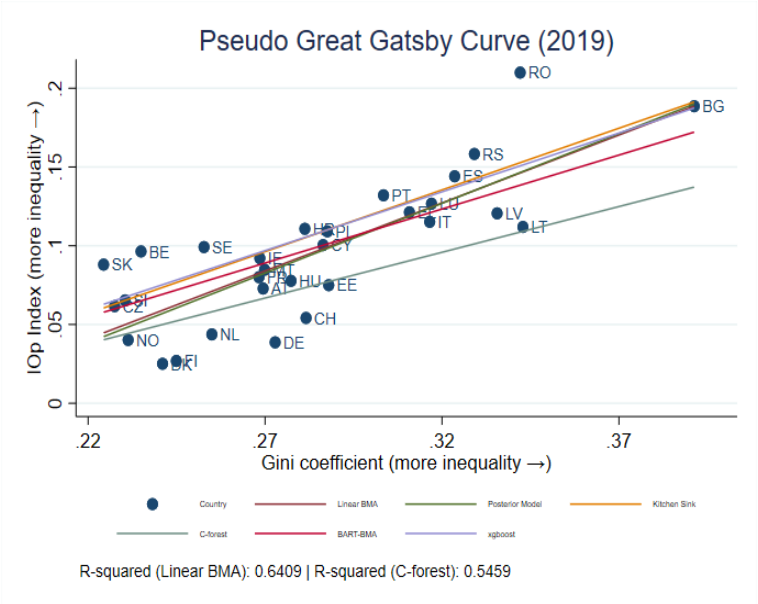


Table 1: Set of Circumstances

Individual characteristics	
(1) Citizenship	**(2) Gender**
- Citizen of the present country of residence	- Sex of the responder
- Citizen of another EU-27 country	
- Born/citizen of a country outside Europe	

Father's/Mother's characteristics	**Household characteristics**
(1) Country of birth and citizenship	**(1) The financial status of household**
- Born/citizen of the present country of residence	- Don't know
- Born/citizen of another EU-27 country	- Very bad
- Born/citizen of another European country	- Bad
- Born/citizen of a country outside Europe	- Moderately bad
(2) Highest level of education attained	- Moderately good
- Don't know/Dead	- Good
- Neither read nor write in any language†	- Very good
- Primary or lower secondary education	**(2) Ability to meet the end of the month†**
- Upper secondary and post-secondary education	- Don't know
- First and second stage of tertiary education	- With great difficulty
(3)Occupational activity status‡	- With difficulty
- Don't know/Dead	- With some difficulty
- Employed	- Fairly easily
- Self-employed (including family worker)	- Easily
- Unemployed	- Very easily
- In retirement or early retirement	**(3) Tenancy status of the house**
- Fulfilling domestic tasks and care responsibilities	- Don't know
- Other inactive person	- Owner
(4) Type of occupation	- Tenant
- Don't know/Dead/Unemployed	- Public housing
- Managers	**(4) Family type**
- Professional	- Lived with both parents (or persons considered as parents)
- Technicians and associate professionals	- Lived with father only (or person considered as a father)
- Clerical support workers	- Lived with mother only (or person considered as a mother)
- Service and sales workers	- Lived in a private household without any parent
- Skilled agricultural, forestry and fishery workers	- Lived in a collective household or institution
- Craft-related trades workers	
- Plant and machine operators, and assemblers	**(5) Number of adults**
- Elementary occupations	**(6) Number of children**
- Armed forces occupations	**(7) Number of working adults**

‡ Not included in Wave 2011.
† Not included in Wave 2019.

Table 2: Summary statistics for Income

This table provides summary statistics by country in 2011 and 2019. The column provides the total number of observations and the remaining five columns summarize the distribution of equivalized disposable household income PPP adjusted: mean, standard deviation and Gini index, respectively.

	2011				2019			
Country	**N**	**Mean**	**SD**	**Gini**	**N**	**Mean**	**SD**	**Gini**
AT	6220	25822.59	13559.18	0.267	5398	25772.31	13329.88	0.269
BE	6011	21.989	10.210	0.248	6212	23174.09	10385.38	0.235
BG	7146	8044.97	5379.53	0.331	6715	11702.34	10339.4	0.391
CH	7583	26536.39	14853.23	0.278	7096	31667.28	17161.06	0.282
CY	4589	22296.34	12.128	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	10.301	0.228	4367	28372.49	11811.74	0.241
EE	5338	10355.61	6288.5	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	13.798	0.287	10255	23601.26	12914.76	0.268
HR	6945	9.828	5.764	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	20704.1	11.970	0.295	4238	22999.76	12703.41	0.269
IS	3682	22628.31	9242.7	0.221	N/A	N/A	N/A	N/A
IT	21070	20.420	12.263	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	28.799	15.451	0.270	4719	28842.45	18281.41	0.317
LV	6423	8188.93	5.747	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	10.868	0.206	6207	28909.88	12922.14	0.231
PL	15545	11235.09	7.128	0.316	19743	13957.88	7991.95	0.288
PT	5899	12.731	8.849	0.333	13767	13198.72	8683.95	0.303
RO	7820	5636.55	3.530	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	8.342	0.237	5206	22387.92	9585.13	0.253
SI	13183	10796.54	4.630	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

Table 3: Out-of-Sample Mean Squared Error

This table reports the out-of-sample Mean Squared Error (MSE) for predicting the logarithm of income across various countries using Linear-BMA, BART-BMA, C-Forest, and XGBoost methods. Results are shown for the 2011 and 2019 waves. The values represent the average MSE computed through 10-fold cross-validation. Lower values indicate better predictive performance, with the best-performing model for each country and wave highlighted in green.

	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
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

Table 4: Top Circumstances

This table presents the top circumstances for Linear-BMA in Panel A, BART-BMA in Panel B, C-Forest in Panel C, and XGBoost in Panel D. It shows the average PIP or Variable Importance scores, as well as the share of countries with PIP or Variable Importance greater than 50%, 75%, 95%, and 99%. For Linear-BMA, the average posterior mean is also provided.

	Panel A: LINEAR-BMA											
	2011						2019					
	Post. mean	Avg. PIP	Share of countries with PIP Greater				Post. mean	Avg. PIP	Share of countries with PIP greater			
Variable			50%	75%	95%	99%			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 in Non-EU Country (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
No. of 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 manual 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 situation (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
	Panel B: BART-BMA											
	2011						2019					
		Avg. PIP	Share of countries with PIP greater					Avg. PIP	Share of countries with PIP greater			
Variable			50%	75%	95%	99%			50%	75%	95%	99%
No. of adults (H)		0.50	0.71	0.42	0.42	0.42		0.52	0.77	0.57	0.33	0.33
No. of children (H)		0.31	0.68	0.19	0.19	0.13		0.42	0.60	0.40	0.30	0.27
No. of workers (H)		0.42	0.61	0.35	0.35	0.29		0.24	0.47	0.20	0.13	0.13
Financial situation (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
	Panel C: C-Forest											
	2011						2019					
		Avg. VarImp	Share of countries with VarImp greater					Avg. VarImp	Share of countries with VarImp greater			
Variable			50%	75%	95%	99%			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 manual occ. (F)		0.41	0.35	0.13	0.10	0.06		0.35	0.20	0.10	0.00	0.00
Born in Non-EU Country (I)		0.30	0.29	0.13	0.06	0.06		0.38	0.40	0.27	0.17	0.17
High-skilled non-manual occ. (F)		0.28	0.23	0.10	0.06	0.06		0.25	0.27	0.10	0.10	0.10
Born outside Europe (F)		0.11	0.10	0.03	0.00	0.00		0.22	0.23	0.13	0.07	0.03
	Panel D: XGBoost											
	2011						2019					
		Avg. VarImp	Share of countries with VarImp greater					Avg. VarImp	Share of countries with VarImp greater			
Variable			50%	75%	95%	99%			50%	75%	95%	99%
Financial situation (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
No. of children (H)		0.60	0.97	0.55	0.35	0.19		0.69	1.00	0.70	0.47	0.27
No. of adults (H)		0.49	0.94	0.48	0.16	0.10		0.45	0.90	0.43	0.17	0.03
No. of 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

Table 5: Top 15 Circumstances - Linear-BMA

This table presents the top 15 variables with the highest Posterior Inclusion Probabilities (PIPs) from Linear-BMA for Waves 2011 and 2019. Panel A corresponds to Wave 2011, while Panel B corresponds to Wave 2019. Variables are sorted by their average PIP values, and a heatmap is used to visually represent the magnitude of the PIPs, with darker shades indicating higher probabilities. (F) denotes Father, (M) denotes Mother, (H) represents Household characteristics, and (I) represents Individual characteristics.

	Panel A: Wave 2011																													
	AT	BE	BG	CH	CY	CZ	DE	DK	EE	EL	ES	FI	FR	HR	HU	IE	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	SE	SI	SK	UK
Education (F)	0.50	0.34	1.00	1.00	1.00	0.99	1.00	0.03	0.85	1.00	1.00	0.99	1.00	0.95	1.00	0.94	0.75	1.00	1.00	1.00	1.00	1.00	0.67	0.03	1.00	1.00	1.00	0.04	1.00	0.65	0.32
Education (M)	1.00	1.00	1.00	1.00	1.00	1.00	0.04	0.02	1.00	0.00	1.00	0.02	1.00	1.00	1.00	0.98	0.64	0.00	0.04	1.00	1.00	0.00	0.62	0.02	0.00	0.00	1.00	0.05	1.00	0.96	1.00
Born in Non-EU Country (I)	1.00	0.27	0.02	0.96	0.02	0.01	0.55	0.03	0.01	1.00	1.00	0.62	1.00	0.01	0.44	0.25	0.03	1.00	0.01	1.00	0.02	0.66	0.16	0.52	0.02	0.06	0.01	0.08	1.00	0.04	0.68
No. of children (H)	0.99	0.73	1.00	1.00	0.04	1.00	0.06	0.03	0.02	0.63	0.22	0.03	1.00	0.14	1.00	0.00	0.04	1.00	0.17	0.81	0.80	0.17	0.14	0.04	1.00	1.00	0.17	0.04	0.04	0.01	0.01
Skilled manual occ. (F)	0.23	0.22	0.98	1.00	0.04	0.20	1.00	0.57	0.02	1.00	1.00	0.03	1.00	0.01	0.05	0.35	0.05	0.94	0.03	0.15	0.01	0.36	0.62	0.09	1.00	1.00	0.04	0.04	0.08	0.04	0.26
High-skilled non-manual occ. (F)	0.88	0.98	0.03	0.00	1.00	0.39	0.00	0.24	0.98	0.00	0.00	0.00	0.00	0.44	1.00	0.00	0.00	0.00	0.03	0.88	0.00	0.00	0.43	0.92	0.00	0.00	0.03	0.00	0.00	0.99	0.72
Financial situation (H)	0.02	0.04	0.98	0.02	0.91	0.01	0.02	0.04	0.55	0.10	1.00	0.02	0.08	0.38	0.01	0.11	0.03	1.00	0.26	0.03	0.01	0.04	0.01	0.03	0.99	1.00	0.97	0.08	0.07	0.21	0.01
Self-employed (F)	0.01	0.31	0.02	0.01	0.16	0.64	0.28	0.79	0.13	0.05	1.00	0.02	0.98	0.01	0.01	0.02	0.02	0.01	0.03	0.01	0.02	0.02	0.01	0.05	1.00	0.02	1.00	0.04	0.35	0.01	0.01
House worker (M)	0.01	0.00	0.03	0.04	0.15	0.31	0.01	0.02	0.78	0.00	0.02	1.00	0.01	0.11	0.99	0.27	0.06	0.99	0.01	0.01	0.03	0.27	0.02	0.02	0.01	0.02	1.00	0.09	0.20	0.72	0.01
Born in other EU Country (I)	1.00	0.02	0.01	0.03	0.02	0.01	0.00	0.13	0.00	0.01	1.00	0.02	0.01	0.01	0.09	1.00	0.03	1.00	0.01	0.93	0.00	0.00	0.02	0.02	0.01	0.03	0.00	0.07	0.00	0.01	0.08
No. of adults (H)	0.07	0.00	1.00	0.06	0.01	0.06	0.01	0.03	0.02	0.01	0.23	0.39	0.44	0.08	0.08	0.90	0.03	1.00	0.01	0.01	0.01	0.02	0.01	0.02	1.00	0.80	0.01	0.04	0.02	0.27	0.02
No. of workers (H)	0.05	0.00	0.97	0.11	0.76	0.01	1.00	0.10	0.02	0.03	0.01	0.03	0.01	0.16	0.07	0.03	0.03	0.04	0.02	0.95	0.01	0.00	0.01	0.08	0.08	0.00	0.01	0.04	0.68	0.02	0.01
Unemployed (F)	0.31	0.00	0.01	0.01	0.01	0.03	0.01	0.02	0.02	0.01	0.99	0.19	0.03	1.00	0.02	0.90	0.02	1.00	0.01	0.01	0.01	0.05	0.04	0.02	0.02	0.05	0.42	0.04	0.34	0.03	0.09
Homeowner (H)	0.15	0.44	0.28	0.02	0.01	0.02	1.00	0.02	0.16	0.01	0.01	0.03	0.06	0.02	0.06	0.99	0.48	0.08	0.01	0.32	0.06	0.31	0.01	0.02	0.01	0.03	0.93	0.07	0.01	0.47	0.76
Female (I)	0.06	0.04	0.21	0.53	0.01	0.84	0.46	0.01	0.03	0.03	0.01	0.03	0.82	0.02	0.03	0.07	0.03	0.18	0.02	0.11	0.15	0.34	0.09	0.02	0.03	0.01	0.01	0.17	0.01	0.04	0.01
	Panel B: Wave 2019																													
	AT	BE	BG	CH	CY	CZ	DE	DK	EE	EL	ES	FI	FR	HR	HU	IE	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	SE	SI	SK	-
Education (F)	0.08	1.00	1.00	1.00	1.00	1.00	0.12	0.03	0.04	1.00	1.00	0.17	1.00	1.00	1.00	1.00	0.92	0.92	1.00	0.99	0.95	0.04	0.01	0.88	1.00	1.00	1.00	0.10	0.37	0.96	
Education (M)	0.70	0.00	1.00	0.96	0.00	1.00	0.04	0.02	1.00	0.82	1.00	0.08	0.00	0.70	0.30	0.00	0.16	0.78	0.99	1.00	0.84	0.01	0.16	1.00	1.00	0.02	0.13	0.05	1.00	0.99	
Financial situation (H)	0.03	1.00	1.00	0.05	0.04	0.02	0.02	0.02	0.01	1.00	1.00	0.02	0.02	1.00	0.87	0.85	0.99	0.01	0.02	0.13	1.00	0.01	0.22	1.00	1.00	1.00	1.00	0.41	0.10	1.00	
No. of children (H)	0.17	0.99	1.00	0.01	1.00	0.53	0.01	0.05	0.04	0.56	1.00	0.03	1.00	0.01	0.97	0.06	0.11	0.14	1.00	0.02	0.02	0.01	0.11	0.88	1.00	1.00	0.05	0.11	0.02	1.00	
Born in Non-EU country (I)	1.00	1.00	0.01	0.12	1.00	0.02	0.03	0.03	0.01	1.00	0.06	1.00	1.00	0.01	0.03	1.00	1.00	0.01	1.00	0.03	0.03	0.72	0.29	0.01	0.01	0.01	0.05	1.00	0.78	0.01	
Skilled manual occ. (F)	0.91	0.38	0.18	0.00	1.00	0.05	0.97	0.29	0.03	1.00	0.79	0.81	1.00	0.43	0.02	0.14	0.98	0.66	0.08	0.15	0.17	0.81	0.02	1.00	1.00	0.22	0.01	0.02	0.03	0.10	
Homeowner (H)	0.03	0.62	0.09	0.02	0.24	0.31	0.99	0.02	0.04	0.06	0.96	0.02	0.99	0.18	0.76	1.00	1.00	0.01	0.06	0.02	0.94	0.04	0.22	0.01	0.02	0.02	0.33	0.11	0.10	1.00	
High-skilled non-manual occ. (F)	0.00	0.00	0.98	0.00	0.00	0.91	0.00	0.00	1.00	0.01	0.25	0.00	0.00	0.00	0.00	1.00	0.00	0.34	0.10	0.00	0.95	0.21	0.13	0.00	0.00	0.86	0.01	0.03	0.97	0.03	
Unknown country of birth (M)	0.02	0.17	0.03	0.56	0.02	0.94	0.10	0.02	0.09	0.03	0.84	0.10	0.44	0.40	0.25	0.00	0.11	0.12	0.47	0.38	0.07	0.01	0.01	1.00	1.00	0.01	0.01	0.03	0.78	0.01	
Self-employed (F)	0.01	0.07	0.02	0.01	0.03	0.02	0.02	0.02	0.01	0.03	0.01	0.02	0.01	0.98	0.01	1.00	0.44	0.33	0.63	0.05	0.08	0.04	0.02	0.99	1.00	1.00	0.00	0.02	0.09	0.41	
Born in Europe, Non-EU (F)	0.00	0.01	0.01	0.02	0.00	0.04	0.00	0.02	0.78	0.00	0.06	0.01	0.26	0.00	0.00	0.10	0.02	0.01	0.02	0.04	0.95	0.01	0.04	0.01	1.00	0.02	0.69	1.00	0.00	0.01	
Born outside Europe (F)	1.00	0.00	0.01	0.04	0.00	0.02	0.04	1.00	0.02	0.02	0.11	0.01	0.00	0.01	0.91	0.00	0.00	0.08	0.18	0.02	0.12	0.21	0.50	0.01	0.44	0.07	0.02	1.00	0.12	0.01	
Part-time worker (F)	0.00	0.01	0.11	0.01	0.05	0.01	0.01	0.02	0.01	0.53	0.10	0.02	0.01	0.01	0.01	0.05	0.79	0.03	0.02	0.02	0.15	0.01	0.01	0.01	0.02	1.00	1.00	0.00	0.01	0.79	
Oth. activity status (F)	0.01	0.01	0.33	0.07	0.03	0.01	0.01	0.03	0.01	0.01	0.97	0.20	0.02	0.02	1.00	0.03	1.00	0.02	0.02	0.00	0.01	0.01	0.01	0.01	0.89	0.13	0.53	0.02	0.03	0.02	
Oth. activity status (M)	0.10	1.00	0.75	0.01	0.03	0.03	0.03	0.02	0.02	0.01	0.03	0.03	0.01	0.02	0.01	0.02	0.57	1.00	0.01	0.01	0.01	0.01	0.02	0.05	0.08	1.00	0.15	0.02	0.01	0.20	

Table 6: Top 15 Circumstances - C-Forest

This table presents the top 15 variables with the highest variable importance scores (VIS) from C-Forest for Waves 2011 and 2019. Panel A corresponds to Wave 2011, while Panel B corresponds to Wave 2019. Variables are sorted by their average VIS values, and a heatmap is used to visually represent the magnitude of the VIS, with darker shades indicating higher scores. (F) denotes Father, (M) denotes Mother, (H) represents Household characteristics, and (I) represents Individual characteristics.

	Panel A: Wave 2011																													
	AT	BE	BG	CH	CY	CZ	DE	DK	EE	EL	ES	FI	FR	HR	HU	IE	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	SE	SI	SK	UK
Education (F)	0.11	0.32	0.83	1.00	0.78	0.67	0.59	0.21	0.80	1.00	0.38	0.27	1.00	0.69	1.00	1.00	0.86	1.00	1.00	0.86	0.83	1.00	0.81	0.66	1.00	1.00	1.00	-0.22	0.70	1.00	0.64
Education (M)	0.23	0.25	1.00	0.58	0.43	1.00	0.05	0.23	1.00	0.00	0.26	0.02	0.77	1.00	0.93	0.86	1.00	0.00	0.74	1.00	1.00	0.00	0.73	0.06	0.00	0.00	0.98	-0.28	1.00	0.88	1.00
Skilled manual occ. (F)	0.10	0.16	0.42	0.99	0.28	0.35	1.00	1.00	0.32	0.56	0.29	0.41	0.67	0.12	0.29	0.50	-0.12	0.31	0.33	0.21	0.15	0.24	0.90	0.57	0.72	0.69	0.16	-0.29	0.23	0.65	0.55
Born in Non-EU Country (I)	1.00	0.46	0.01	0.65	0.83	-0.07	0.09	0.21	0.10	0.80	1.00	0.29	0.64	0.06	0.01	0.09	0.14	0.26	-0.01	0.57	0.06	0.01	0.25	0.69	-0.02	0.01	0.00	0.19	0.53	0.00	0.46
High-skilled non-manual occ. (F)	0.10	0.16	0.32	0.00	0.49	0.40	0.00	0.71	0.71	0.00	0.00	0.00	0.00	0.36	0.59	0.00	0.00	0.00	0.47	0.45	0.00	0.00	1.00	1.00	0.00	0.00	0.31	0.00	0.00	0.79	0.66
Born outside Europe (F)	0.00	0.57	0.00	0.17	0.74	0.00	0.00	0.00	0.00	0.13	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.06	0.82	0.00	0.00	0.00	0.00	0.00	0.00	0.44
Born outside Europe (M)	0.00	0.57	0.00	0.00	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.07	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.56
Cit. outside Europe (F)	0.00	1.00	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.26	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.13
Cit. outside Europe (M)	0.00	0.99	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.12	0.28	0.00	0.00	0.00	0.00	0.00	0.00	0.11
Self-employed (F)	0.02	0.03	0.01	0.09	0.05	0.04	0.18	0.51	0.11	0.09	0.06	0.02	0.10	0.00	0.00	0.00	-0.12	0.00	0.04	0.03	0.03	0.01	-0.01	0.01	0.34	0.03	0.85	-0.22	0.11	-0.03	0.04
Financial situation (H)	0.03	0.13	0.21	0.00	0.31	0.00	0.13	0.00	0.18	0.09	0.15	0.06	0.08	0.34	0.12	0.33	0.26	0.30	0.55	0.07	0.07	0.07	0.08	0.03	0.22	0.90	0.32	0.18	0.03	0.16	0.08
Homeowner (H)	0.03	0.06	0.03	0.06	0.07	0.05	0.38	0.10	0.05	0.00	0.01	0.01	0.09	0.00	0.02	0.72	0.95	0.01	0.07	0.07	0.08	0.05	-0.05	-0.03	0.09	0.02	0.08	0.22	-0.01	0.18	0.27
No. of children (H)	0.06	0.05	0.52	0.22	0.02	0.25	0.05	0.01	0.05	0.16	0.04	-0.01	0.49	0.22	0.64	0.00	0.13	0.45	0.22	0.08	0.33	0.04	0.16	0.03	0.14	0.22	0.17	-0.33	0.06	0.02	0.00
Female (I)	0.03	0.00	0.00	0.04	-0.01	0.08	0.03	-0.14	0.03	-0.01	0.00	-0.02	0.02	-0.06	0.00	0.06	-0.09	0.01	-0.08	0.01	0.12	0.04	0.14	-0.09	0.00	-0.01	-0.03	0.85	0.00	0.03	-0.04
Cit. EU (F)	0.00	0.02	0.00	0.00	0.36	0.03	0.02	0.53	0.45	0.00	0.00	0.00	0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.01	-0.02	0.00	0.00	0.00	0.00	0.00	-0.09	-0.01
	Panel B: Wave 2019																													
	AT	BE	BG	CH	CY	CZ	DE	DK	EE	EL	ES	FI	FR	HR	HU	IE	IS	IT	LT	LU	LV	MT	NL	NO	PL	PT	RO	SE	SI	SK	-
Education (F)	0.03	0.25	1.00	0.84	1.00	1.00	1.00	0.00	0.30	1.00	0.46	0.11	0.56	1.00	1.00	1.00	0.59	1.00	1.00	0.96	0.86	0.11	0.02	0.77	0.55	0.77	1.00	-0.03	0.86	0.60	
Education (M)	0.13	0.00	0.97	1.00	0.00	0.73	0.20	-0.03	0.50	0.45	0.20	0.23	0.00	0.82	0.82	0.00	0.40	0.81	0.59	1.00	0.57	-0.01	0.02	1.00	0.33	0.32	0.41	0.09	0.90	0.57	
Born in Non-EU country (I)	1.00	1.00	0.00	0.25	0.92	-0.02	0.12	0.13	0.02	0.55	0.49	1.00	1.00	0.01	-0.04	0.14	1.00	-0.06	0.90	0.07	0.14	0.58	0.66	-0.05	0.04	0.00	0.10	0.54	0.84	0.00	
High-skilled non-manual occ. (F)	0.00	0.00	0.58	0.00	0.00	0.64	0.00	0.00	1.00	0.14	0.19	0.00	0.00	0.00	0.00	0.52	0.00	0.66	0.30	0.00	1.00	0.58	0.10	0.00	0.00	0.49	0.17	0.00	1.00	0.30	
Born outside Europe (F)	0.44	0.00	0.00	0.11	0.00	0.00	0.95	1.00	0.00	-0.01	0.56	0.31	0.00	0.00	0.11	0.00	0.00	0.00	0.56	0.00	0.09	0.20	0.99	0.00	0.01	0.02	0.00	0.75	0.54	0.00	
Skilled manual occ. (F)	0.12	0.08	0.46	0.00	0.77	0.36	0.81	0.36	0.30	0.45	0.17	0.41	0.36	0.39	0.12	0.16	0.27	0.69	0.25	0.48	0.44	0.27	0.03	0.71	0.81	0.37	0.12	-0.01	0.50	0.24	
Born outside Europe (M)	0.00	0.00	0.00	0.21	0.00	-0.01	0.04	0.00	0.01	-0.01	0.73	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.38	0.05	0.02	0.05	0.60	0.00	0.00	0.00	0.00	0.80	0.75	0.00	
Cit. outside Europe (F)	0.00	0.00	0.00	0.14	0.00	0.00	0.04	0.00	0.02	0.00	0.82	0.00	0.00	0.00	0.11	0.00	0.00	0.20	0.47	0.06	0.09	1.00	1.00	0.00	0.00	0.03	0.00	1.00	0.46	0.00	
Cit. outside Europe (M)	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.08	0.00	1.00	0.00	0.00	0.00	0.02	0.00	0.00	0.06	0.34	-0.05	0.03	0.50	0.83	0.00	0.00	0.01	0.05	0.89	0.58	0.00	
Financial situation (H)	0.02	0.13	0.41	0.06	0.23	0.06	0.07	0.05	-0.04	0.36	0.32	-0.06	0.05	0.76	0.35	0.18	0.72	0.11	0.09	0.11	0.40	0.02	0.14	0.23	1.00	0.36	0.46	0.03	0.07	0.56	
Born in EU country (I)	0.21	0.02	0.00	0.36	0.46	0.04	0.00	0.63	0.00	0.03	0.09	-0.06	0.02	-0.02	-0.03	0.31	0.39	0.00	0.27	0.00	0.00	0.60	0.19	0.00	0.00	0.00	-0.02	0.04	0.00	-0.01	
Self-employed (M)	-0.01	0.01	0.01	-0.01	0.03	0.01	-0.14	0.05	-0.04	0.05	0.01	0.21	0.00	0.25	-0.05	0.01	0.05	-0.01	-0.01	0.05	0.04	-0.09	-0.03	0.64	0.02	0.70	0.00	-0.01	0.02	0.00	
Homeowner (H)	0.03	0.06	0.02	0.14	0.06	0.06	0.79	0.03	-0.04	0.09	0.05	0.01	0.14	0.06	0.13	0.88	0.30	-0.01	0.07	0.06	0.30	-0.02	0.26	0.01	0.02	0.02	0.13	0.02	0.13	0.19	
No. of workers (H)	0.02	0.04	0.02	0.05	0.00	0.07	-0.06	0.42	0.17	0.02	0.02	0.07	0.03	0.28	0.27	0.00	0.00	-0.03	0.00	0.29	-0.04	0.07	0.08	-0.01	0.06	0.25	0.64	-0.01	0.35	0.51	
No. of children (H)	0.06	0.12	0.38	0.06	0.30	0.18	0.17	0.08	0.00	0.15	0.06	0.09	0.21	0.04	0.58	0.05	0.14	0.10	0.37	0.15	0.04	0.10	0.22	0.13	0.39	0.20	0.07	0.01	0.13	1.00	

References

Aaberge, R., M. Mogstad, and V. Peragine (2011). Measuring long-term inequality of opportunity". Journal of Public Economics 95, 193–204.

Barry, B. (1992, July). Theories of justice: Treatise on social justice v. 1. California Series on Social Choice & Political Economy. Berkeley, CA: University of California Press.

Bernardo, G., A. Kourtellos, and C. M. Tan (2024). Model uncertainty and measures of inequality of opportunity: Supplementary online appendix. Unpublished Working Paper, University of Cyprus.

Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research 13, 1063–1095.

Bourguignon, F., F. H. G. Ferreira, and M. Menéndez (2007). Inequality of opportunity in brazil. Review of Income and Wealth 53(4), 585–618.

Breiman, L. (2001). Random forests. Machine Learning 45(1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). Classification and Regression Trees (1st ed.). Chapman and Hall/CRC.

Brock, W. A. and S. N. Durlauf (2001). Growth empirics and reality. The World Bank Economic Review 15(2), 229–272.

Brunori, P., F. H. G. Ferreira, and V. Peragine (2013). Inequality of opportunity, income inequality, and economic mobility: Some international comparisons. In E. Paus (Ed.), Getting Development Right: Structural Transformation, Inclusion, and Sustainability in the Post-Crisis Era, pp. 85–115. New York: Palgrave Macmillan US.

Brunori, P., P. Hufe, and D. Mahler (2023). The roots of inequality: estimating inequality of opportunity from regression trees and forests. The Scandinavian Journal of Economics 125(4), 900–932.

Brunori, P., V. Peragine, and L. Serlenga (2019). Upward and downward bias when measuring inequality of opportunity. Social Choice and Welfare 52(4), 635–661.

Checchi, D. and V. Peragine (2010). Inequality of opportunity in italy. J. Econ. Inequal. 8(4), 429–450.

Checchi, D., V. Peragine, and L. Serlenga (2016). Inequality of opportunity in europe: Is there a role for institutions? In L. Cappellari, S. W. Polachek, and K. Tatsiramos (Eds.), Inequality: Causes and Consequences, Volume 43, Chapter 1, pp. 1–44. Bingley: Emerald Insight.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 785–794. ACM.

Chipman, H. A., E. I. George, and R. E. McCulloch (2010). Bart: Bayesian additive regression trees. The Annals of Applied Statistics 4(1), 266–298.

Corak, M. (2013, August). Income inequality, equality of opportunity, and intergenerational mobility. J. Econ. Perspect. 27(3), 79–102.

Durlauf, S. and Y. Ioannides (2010). Social interactions. Annual Review of Economics 2(1), 451–478.

Durlauf, S. N., A. Kourtellos, and C. M. Tan (2022, August). The great gatsby curve. Annu. Rev. Econom. 14(1), 571–605.

Eicher, T. S., C. Papageorgiou, and A. E. Raftery (2011, January). Default priors and predictive performance in bayesian model averaging, with application to growth determinants. J. Appl. Econ. (Chichester Engl.) 26(1), 30–55.

Falconer, J. R., E. Frank, D. L. L. Polaschek, and C. Joshi (2022). Methods for eliciting informative prior distributions: A critical review. Decision Analysis 19(3), 189–254.

Fernández, C., E. Ley, and M. F. J. Steel (2001, February). Benchmark priors for bayesian model averaging. J. Econom. 100(2), 381–427.

Ferreira, F. H. G. and J. Gignoux (2011, December). The measurement of inequality of opportunity: Theory and an application to latin america. Rev. Income Wealth 57(4), 622–657.

Ferreira, F. H. G. and V. Peragine (2016, November). Individual responsibility and equality of opportunity. In M. D. Adler and M. Fleurbaey (Eds.), The Oxford Handbook of Well-Being and Public Policy. Oxford University Press.

Fleurbaey, M. and V. Peragine (2013). Ex ante versus ex post equality of opportunity. Economica 80(317), 118–130.

Fleurbaey, M. and E. Schokkaert (2009). Unfair inequalities in health and healthcare. J. Health Econ 28, 73–90.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics 29(5), 1189–1232.

Hernández, B., A. E. Raftery, S. R. Pennington, and A. C. Parnell (2018, July). Bayesian additive regression trees using bayesian model averaging. Stat. Comput. 28(4), 869–890.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15(3), 651–674.

Hufe, P., R. Kanbur, and A. Peichl (2018). Measuring unfair inequality: Reconciling equality of opportunity and freedom from poverty. Technical Report 11601, IZA Discussion Papers, Bonn.

Hufe, P., A. Peichl, J. Roemer, and M. Ungerer (2017, December). Inequality of income acquisition: the role of childhood circumstances. Soc. Choice Welfare 49(3-4), 499–544.

Krueger, A. (2012). The rise and consequences of inequality in the united states. Speech, Cent. Am. Prog.

Leamer, E. E. (1978). Specification Searches: Ad Hoc Inference with Nonexperimental Data. New York: Wiley.

Lefranc, A., N. Pistolesi, and A. Trannoy (2009, December). Equality of opportunity and luck: Definitions and testable conditions, with an application to income in france. J. Public Econ. 93(11-12), 1189–1207.

Mitnik, P. A., A.-L. Helsø, and V. L. Bryant (2022). Inequality of Opportunity for Income in Denmark and the United States: A Comparison Based on Administrative Data, pp. 317–382. Chicago: University of Chicago Press.

Pareliussen, J. (2017). Income, wealth and equal opportunities in Sweden. Technical Report No. 1394, OECD Economics Department Working Papers.

Peragine, V. and L. Serlenga (2008). Higher education and equality of opportunity in italy. In Research on Economic Inequality, Research on economic inequality, pp. 67–97. Bingley: Emerald (MCB UP).

Pignataro, G. (2012, December). Equality of opportunity: Policy and measurement paradigms. J. Econ. Surv. 26(5), 800–834.

Pistolesi, N. (2009, December). Inequality of opportunity in the land of opportunities, 1968–2001. J. Econ. Inequal. 7(4), 411–433.

Porwal, A. and A. E. Raftery (2022, April). Comparing methods for statistical inference with model uncertainty. Proc. Natl. Acad. Sci. U. S. A. 119(16), e2120737119.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Model selection and accounting for model uncertainty in linear regression models. Journal of the American Statistical Association 92, 179–191.

Raftery, A. E. and Y. Zheng (2003). Discussion:performance of bayesian model averaging. Journal of the American Statistical Association 98(464), 931–938.

Roemer, J. (1998). Equality of Opportunity. Harvard University Press.

Roemer, J. E. (1993). A pragmatic theory of responsibility for the egalitarian planner. Philosophy and Public Affairs 22(2), 146–166.

Roemer, J. E. (1996, February). Theories of Distributive Justice. London, England: Harvard University Press.

Roemer, J. E. and A. Trannoy (2016, December). Equality of opportunity: Theory and measurement. J. Econ. Lit. 54(4), 1288–1332.

Sala-i Martin, X., G. Doppelhofer, and R. I. Miller (2004, September). Determinants of long-term growth: A bayesian averaging of classical estimates (BACE) approach. Am. Econ. Rev. 94(4), 813–835.

Scornet, E., G. Biau, and J.-P. Vert (2015). Consistency of random forests. The Annals of Statistics 43(4), 1716–1741.

Suárez Álvarez, A. and A. Jesús López Menéndez (2018, December). Income inequality and inequality of opportunity in europe: Are they on the rise? In Inequality, Taxation and Intergenerational Transmission, Research on economic inequality, pp. 149–196. Emerald Publishing Limited.

Swedish Fiscal Policy Council (2024). Economic inequality in Sweden: an overview of facts and future challenges. Technical Report 2024/01.

Van De Gaer, D. (1993). Equality of opportunity and investment in human capital. Ph. D. thesis, Catholic University of Leuven.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113(523), 1228–1242.