Hacienda Pública Española/Review of Public Economics by
Instituto de Estudios Fiscales is licensed under CC BY-NC 4.0



Hacienda Pública Española/Review of Public Economics 252-(1/2025): 113-162
© 2025, Instituto de Estudios Fiscales
https://dx.doi.org/10.7866/HPE-RPE.25.1.4

Inequality of Opportunity in Education in Spanish Regions: A Machine Learning Approach*

PABLO BENCOMO-MESA**
Universidad de La Laguna
GUSTAVO A. MARRERO***
Universidad de La Laguna and *EQUALITAS*
GABRIELA SICILIA****
Universidad Autónoma de Barcelona

Received: April, 2024
Accepted: October, 2024

Abstract

Inequality of Opportunity in Achievement (IOpE) measures the importance of factors beyond the student's control in explaining differences in academic performance. Using PISA 2018, we estimate the IOpE for Spanish regions using conditional inference tree (CIT) and forest (CIF). Using CIFs, IOpE is twice as high as those obtained using traditional approaches (on average, 43% compared to 20%). Murcia and Extremadura are among those with the highest IOpE, while Castilla-La Mancha and the Pais Vasco have the lowest IOpE. The circumstances that contribute most to IOpE are the cultural environment at home (number of books) and parental occupation.

Keywords: Inequality of opportunity, Academic achievement, Machine learning, PISA, Spanish regions.

JEL Classification: D63, I24, I28, O52.

1. Introduction

The growing interest in inequality of opportunity (IO) in terms of the level of skills acquired by individuals is based on two facts. First, there is evidence of the positive impact

* Authors acknowledge comments from two anonymous referees. Authors acknowledge financial support from the Ministry of Science, Innovation and Universities of Spain PID2022-137352NB-C44, and Sicilia from PID2023-147616NB-100. Sicilia also acknowledges financial support from Ramon Areces Fundation. Bencomo also thanks the financial support from the Gobierno de Canarias through its FPI program (Tesis2021010086 grant), co-funded by the European Social Fund.

** ORCID ID: 0000-0003-2010-3464.

*** ORCID ID: 0000-0003-4030-0078.

**** ORCID ID: 0000-0002-1078-7762.

of educational outcomes on the future well-being of individuals and on the rate of economic growth and development of countries (Yakovlev and Leguizamon, 2012; Marconi, 2018), where quality education is an important tool for reducing social and economic inequalities (Hofmarcher, 2021; Abdullah *et al*., 2015). Second, international evidence shows that the social and economic components of the family remain a fundamental determinant of students' educational success in both underdeveloped and developed countries (Sirin, 2005; Schütz *et al*., 2008; Jerrim *et al*., 2019). Therefore, one of the most important challenges facing governments and international organisations is to ensure equal opportunities for school-aged people, where everyone has access to quality education and where their academic outcomes depend on their ability and effort rather than on circumstances defined by their social, racial, economic or cultural background (Coleman *et al*., 1966).

In this context, this paper analyses the performance of the Spanish education system in terms of inequality of opportunity in educational achievement (IOpE) from a comparative regional perspective. The study of IOpE at the regional level is particularly relevant in Spain due to the devolution of educational competences to regional governments, which are responsible for the management of educational resources and for a large part of the design and implementation of public policies. The devolution process, which began in the early 2000s, has been long and complex, with very different starting points among the regions, resulting in notable differences in terms of academic outcomes, efficiency and educational equity.

According to the 2018 report of the Programme for International Student Assessment (PISA), the difference between the best (Galicia) and the worst (Canarias) performers in science was almost 40 points, which corresponds to more than one school year. This difference is even more significant if we look at the gap between the percentage of students in Galicia (13.5%) and Canarias (25%) who do not reach the minimum performance level in science. Thus, these marked differences between regions motivate the relevance of comparing IOpE at the Spanish regional level. However, due to the lack of external educational evaluations that allow the comparison of performance between regions, the available evidence in this regard is very limited (Martínez de Lafuente and Vega-Bayo, 2023; Sicilia and Simancas, 2023).

This paper is a contribution to this literature. We use data from the 2018 edition of PISA, in which Spain participated with an expanded sample of students to obtain regionally representative data. These data provide an objective and comparable measure of students' academic performance in three basic skills (mathematics, science and reading) and include extensive information on students' family and school socioeconomic background (Sicilia and Simancas, 2023; Marrero *et al*., 2024), which is a key element for measuring IOpE.

To estimate the effect of circumstances on educational achievement inequality, we use machine learning (ML) techniques. Specifically, we use conditional inference trees (CIT) and conditional inference forests (CIF), which have been shown to be superior to more traditional regression trees and random forests in a number of ways (Hothorn *et al*., 2006). While these methods have recently been applied in the income inequality of opportunity literature to overcome the limitations of traditional regression-based (linear or non-linear) approaches that introduce downward or upward bias (Brunori *et al*., 2019, 2021, 2023; Brunori and Neid-

hofer, 2021; Plassot *et al*., 2022), to our knowledge they have not been applied to measure IOpE. Most previous IOpE studies have used standard parametric or nonparametric methods (Schütz *et al*., 2008; Martins and Veiga, 2010; Ferreria and Gignoux, 2014; Lasso de la Vega *et al*., 2020; Sicilia *et al*., 2022; Marrero *et al*., 2024).[1]

If we apply these results from the income literature to the education setting, estimates using traditional techniques are likely to be highly biased. Moreover, ranking and variable importance analysis could be misleading. By using CIT and CIF, we attempt to overcome the limitations of traditional regression techniques in the education literature. Thus, although this is not a policy-oriented paper, correctly estimating IOpE and the factors that influence it is a necessary first step in understanding the problem of educational inequality in a region and in thinking about policy recommendations.

The research is original in several respects. From an applied perspective, it provides a measure of comparative IOpE at the regional level, including all Spanish regions. This is informative for the design of public policies, given the decentralised nature of educational competences in Spain. From a methodological point of view, the application of ML techniques to the IOpE literature allows us to dimension the effect of the problem of upward and downward bias in our context, where the data structure is different from that of income or health, where ML approaches have already been used to estimate inequality of opportunity. We estimate the entire IOpE distribution using bootstrapping, which provides not only point estimates but also confidence intervals. In conclusion, this paper provides scientific evidence to understand and address one of the most challenging issues in the Spanish education system: the formulation of educational policies and interventions aimed at reducing the existing educational inequalities between regions.

The rest of the paper is organized as follows. Section 2 reviews the background of the IOpE for Spain. Section 3 explains the methodology used to estimate the IOpE. Section 4 describes the database and the variables included in our models. Section 5 includes the analysis and discussion of the results. Section 6 concludes the paper.

2. Inequality of opportunities in educational achievement in Spain

Comparative evidence across education systems on the extent of inequality of opportunity in terms of skills acquisition and learning (achievement) is recent and still scarce (compared, for example, with evidence on inequality of opportunity in terms of income or earnings). Most of it comes from information provided by international large-scale assessments (ILSAs), as they allow comparisons between education systems. The evidence available for the case of European education systems, including Spain in the comparison, is presented below.

Shultz *et al*. (2008) were the first to present comparative evidence for 54 countries on the level of IOpE in primary and secondary education, using data from the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS) in 1995 and 1998, respectively. The authors measure the impact of

socioeconomic status (measured by the number of books at home) on student performance and then examine the relationship between IOpE and certain institutional characteristics or educational policies of education systems (early tracking, attendance at early childhood education, among others). According to this proxy of IOpE, Spain is in the group of countries with the lowest effect of family background on educational achievement in the sample and among European countries.

Martins and Veiga (2010) compare IOpE in 15 European countries using data from the 2003 edition of PISA. The authors find that inequality in students' socioeconomic factors accounts for between 15% and 35% of the total inequality in IOpE, with Spain also being among the countries with the lowest inequality, but also among the countries where socioeconomic inequality is mainly explained by students' background characteristics.

Ferreria and Gignoux (2014) measure IOpE in 57 countries that participated in the 2006 edition of PISA. The authors use a parametric linear regression model and include a wide range of family circumstances. They find that IOpE can be as high as 35% and that it is not correlated with countries' average educational attainment or GDP per capita. The countries with the highest IOpE are Central European countries (Belgium, France, and Germany) and Eastern European countries (Bulgaria and Hungary), while the Nordic countries and Italy have the lowest IOpE. Spain is in the middle.

Lasso de la Vega *et al*. (2020) compare IOpE in 20 European countries using information from PISA 2012 and applying the approach used by Jusot *et al*. (2013), which is based on estimates of a linear regression that includes as explanatory factors variables related to student circumstances, but also other factors related to student effort. As in previous studies, the authors find that the Central European countries show the highest IOpE levels, while the Nordic countries, Ireland and Spain have the lowest.

Sicilia *et al*. (2022) apply a parametric linear regression approach to the 2018 PISA data to estimate IOpE for EU-27 countries and Spanish regions (see also Marrero *et al*., 2024, for European countries). Spain and the Nordic countries have relatively moderate levels of IOpE (around 20%), while the Central European countries –mainly Germany, Belgium, France and the Netherlands– have the highest levels of IOpE (between 38% and 45%). Comparing the Spanish regions, the authors find that the IOpE levels of the Spanish regions –and the variation between them– are lower than those found for the European countries. The IOpE for Spain is 19%, with the Baleares and Castilla-La Mancha at the lower end (16%), while Asturias and Madrid exceed 21% and Murcia has the highest IOpE at 27%. These differences do not seem to be linked to any geographical, economic, or socio-cultural pattern.

Sicilia and Simancas (2023) perform a comprehensive evaluation of the performance of the Spanish education system in terms of efficiency and equity from a regional comparative perspective. To evaluate the educational equity, they compute multiple indicators used in the literature using data from PISA 2018. To measure the influence of the students' socioeconomic background on their performance they use the percentage of the variance in achievement explained by the socioeconomic background (R squared in the regression),

which is close to our measure of IOpE. In line with Sicilia *et al*. (2022), the authors found that Castilla-La Mancha is the most equitable regions, followed by Valencia, both of which are more equitable than the OECD countries on average. In contrast, Asturias, Madrid, and Murcia stand out as the most inequitable regions in the Spanish context.

Finally, Martínez de Lafuente and Vega-Bayo (2023) uses a battery of ML techniques to measure educational mobility and educational inequality in Europe and in Spanish regions. For each region, they use the model that provides better out-of-sample predictions. This strategy generates IOpE estimates that are not fully comparable across regions, as the resulting tuning parameters are likely to be different, which could strongly affect the IOpE estimates and comparability. Nevertheless, they obtain a similar ranking and estimates as in Sicilia *et al*. (2022): the IOpE estimates range from values around 16% (Castilla y León, La Rioja or Valencia) to values around 29% (Asturias, Canarias or Murcia).

With the exception of Sicilia *et al*. (2022), Martínez de Lafuente and Vega-Bayo (2023) and Sicilia and Simancas (2023), there is no comparative work on IOpE in the Spanish regions. And, to the best of our knowledge, except for this second work, we are the first to use ML techniques to estimate IOpE. In addition, we use conditional inference trees and forests, which have been shown to have important advantages over traditional regression trees and random forests (see section 3). We use a parameter tuning approach following Brunori *et al*. (2023). To make IOpE estimates comparable across regions, we select a common set of tuning parameters for all regions. As we will show in the results section, our IOpE estimates are significantly higher than previous estimates, and in some regions they are almost double.

3. Methodology

The most common procedure for estimating IOpE is the one proposed by Ferreira and Guignoux (2014). It is an adaptation of the proposal for estimating IO in income (Bourguignon *et al*., 2007; Ferreira and Guignoux, 2011; Marrero and Rodríguez, 2012). The method is based on estimating a reduced form (usually linear or log-linear) that relates a measure of educational attainment, E, to a set of circumstances $\{C_j\}_{j=1}^{J}$, and an error term (unexplained part), ε: $E = f(C_1, \ldots C_j, \ldots, C_J) + \varepsilon$.

Circumstances are factors beyond the student's control, such as gender or race, place of residence, immigrant status, socio-economic status of parents, cultural environment at home, type of school in which parents enrolled the student, etc. The ordinary least squares (OLS) estimation of this reduced form allows us to orthogonalize the explanatory part of E by the set of circumstances (fitted model), $\hat{E} = \hat{f}(C_1, \ldots C_j, \ldots, C_J)$, of the unexplained part (the residual), $\hat{\varepsilon}$. The ratio between the variance of the explained part and the variance of the total educational achievement, $var(\hat{E})/var(E)$, equal to the coefficient of determination (R^2) of the regression in the case of linearity, measures the percentage of variability in educational performance that is associated with the set of circumstances. This is the measure of IOpE that we use in this paper.

This is the standard measure used in the literature. When measuring IO in terms of income or wealth, other inequality indices are often used, such as Gini, MLD or Theil 0. However, the measures of educational performance that is commonly used come from external assessments (e. g. PISA) and is standardized by the average of the year of the assessment and the competence. Indices such as the Gini or the MLD are affected by this standardization, which would make it difficult to compare the IOpE results across years and competencies. However, the variance is not affected by the standardization (neither the standard deviation nor the coefficient of variation), which is why it is the measure commonly used in the education literature (Ferreira and Guignoux, 2014; Marrero *et al*., 2024).

By using this linear approach, we may be committing at least two types of bias in the fitted part, $\hat{E}$, with opposite signs, which are translated into the estimate of the IOpE (Brunori *et al*., 2019; 2023). The first is a downward bias, which is due to the fact that the circumstances included in the model (those that we can measure and observe), as well as the functional form considered, can only capture part of the total existing IOpE. The second type of bias, which is upward, is due to possible problems of overfitting that we make by introducing irrelevant circumstances into the model (e. g., by including quadratic or cross effects that are not relevant). This leads to an excess of uncertainty in the predictions of E and leads to an upward bias in the estimate of IOpE.

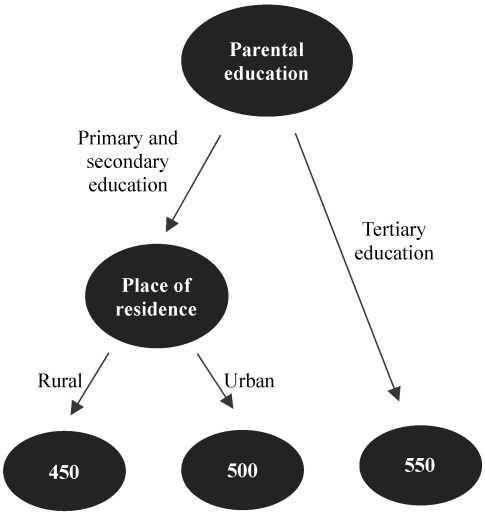
This problem has recently been highlighted in the estimation of IO in income in several papers, such as Brunori *et al*. (2019; 2023), Brunori and Neidhofer (2021), or Plassot *et al*. (2022), among others. Dealing with these biases is not trivial, since conventional methods tend to reduce one bias by increasing the other. For example, the usual way to reduce the downward bias is to introduce interactions or quadratic effects of circumstance into the regression, but this increases the risk of overfitting. Brunori *et al*. (2019) propose to balance the magnitude of these two biases using a regularisation procedure that minimizes the out-of-sample mean squared error (MSE), to obtain IOpE estimates that are not biased in either direction. Supervised machine learning (ML) methods using cross-validation (CV) or similar approaches to assess the out-of-sample MSE is the procedure proposed by these authors. More specifically, they propose to use Conditional Inference Trees (CIT) and Conditional Inference Forests (CIF) (Hothorn *et al*., 2006) because of their flexibility to include nonlinearities and especially because of their advantages over other alternatives (latent variables, lasso, etc.).

3.1. Conditional Inference Trees (CIT)

The goal of the algorithm behind a decision tree (Morgan and Sonquist, 1963; Breiman *et al*., 1984) is to make an out-of-sample prediction of the outcome of a target variable (academic performance, in our case) based on a set of factors or characteristics (our set of circumstances). In order to do this, the space of all possible combinations of circumstances is roughly divided into non-overlapping groups. The name “tree” comes from the way the algorithm generates meaningful sample splits. See Figure 1 for an illustrative example, which will be discussed below.

In addition to predicting a target variable, a decision tree provides a measure of the importance of the explanatory variables and its graphical representation helps to interpret the results. The division of space depends on the procedure used. Appendix 1, part A, explains the approach used in conditional inference trees (CIT).

Figure 1
ILLUSTRATION OF A DECISION TREE



Source: Own elaboration.

To reject or not the null hypothesis, it is necessary to set a significance level (α): a very small α makes us very demanding to reject the null hypothesis of independence; conversely, a very high α makes us not very demanding. Therefore, the higher the α, the more we allow the tree to grow and become "deeper". In general, a tree with a very high α tends to generate problems of overfitting (Hothorn *et al*., 2006), while a tree with a very low α tends to omit relevant information with predictive power. To control the size of the tree (regularisation), we "tune" the α by using CV techniques (Brunori *et al*., 2023), that we refer as α_* (detail is shown in Appendix 1, part B).

For the estimated level of α_*, we estimate the CIT and its end nodes will (univocally) determine a particular group of students, where each group is associated with some categories from a set of circumstances (those that have proven to be significant in each case). As discussed above, a situation of equal opportunity in educational achievement would generate zero partitions. Once the CIT has been estimated, the sample mean of educational achievement is calculated for each end node, which is used as a predictor for every individual within each end node: $\hat{E}^{CIT}$. Finally, the IOpE measure would be: $\frac{var(\hat{E}^{CIT})}{var(E)}$, following the same metric than in the traditional regression approach.

In the example in Figure 1, three groups of students have been created: those who belong to families where the mother has a tertiary education (group 1); and among those whose parents do not have a tertiary education, the division is between living in urban (group 2) or rural (group 3) area. The predicted academic achievement for each group is the average prediction of all the students within the group. Thus, in this example, the students in group 1 are the highest performers and those in Group 3 are the lowest performers. The inequality in educational achievement between these three groups, divided by the total inequality, would be the way to estimate the IOpE in this example.

We present the CIT results for their ease of interpretation and robustness. However, this will not be our preferred method, as there are several estimation problems associated with the CIT (Breiman, 1996, 2001; Friedman *et al*., 2009). Firstly, the results are highly dependent on the sample used and on the presence of missing values: for example, small sample variations can cause the first partition to vary, which already conditions the entire resulting tree. Secondly, they do not work well when highly correlated explanatory factors are included (as is often the case in applications such as ours), since in this situation the tree tends to use the information from a few variables and to omit the rest of the factors (even if they have predictive capacity). In the second case, the trees tend not to use all the valid information available (they use the information inefficiently). Both shortcomings can be addressed using conditional inference forests (Breiman, 2001; Biau and Scornet, 2016), which will be our preferred procedure.

3.2. Coditional Inference Forest (CIF)

A random forest model estimates a large number (N) of trees. Each estimated tree follows the procedure described above, but does not use the same sample or includes the same explanatory factors. A CIF uses the Baggin (bootstrap aggregation) method to generate the datasets used in each of the N estimated trees. Thus, each tree is estimated using a random subsample (without replacement) of approximately 2/3 of the original sample (addressing the first CIT problem commented above), and considers a subsample of $M \leq J$ randomly permuted factors at each node (addressing the second CIT problem commented above). The selection of variables and partitions given the M factors included follows the same procedure as in a CIT, based on the p-values of the hypothesis tests discussed above. Once all the trees in the forest have been estimated, the average of the predicted values obtained for each tree is taken as the result of the prediction for the forest (Hothorn *et al*., 2004; Meinshausen, 2006): $\hat{E}^{CIF}$. Finally, the variance of $\hat{E}^{CIF}$ is calculated, and, as before, its ratio to the total variance is our measure of IOpE: $\frac{var\left(\hat{E}^{CIF}\right)}{var(E)}$.

In the case of the CIF, there are three parameters that are relevant to tune: the α (the same as in the CIT), the number of trees to include (N) and the number of circumstances that are randomly permuted in each case (M). The tuning of these parameters is explained in details in Appendix 1, part C (following Brunori *et al*., 2023). The tuning parameters are referred as α_*, N_* and M_*. As we will see in the results section, the IOpE estimates strongly depend on

the chosen α, and to a lesser extent on M. For this reason, and given that the gains in out-of-sample MSE are very small for certain values of α, it is reasonable to set the same value of these parameters for all regions when making comparisons between regions. We will discuss this aspect in more detail in the results section.

3.3. Circumstances importance

For both CIT and CIF, there are several ways of calculating the importance of each factor (circumstance) in predicting the target variable (academic achievement). In a very intuitive way, the importance of each circumstance is basically calculated by measuring how much predictive power (accuracy) is lost when each circumstance is removed. The greater the loss of predictive power when the variable is removed, the greater the importance of that variable in predicting the target variable (Strobl *et al*., 2008; Hapfelmeier *et al*., 2014). This measure is conditional in the sense that the procedure takes into account the conditional correlation between all variables. Furthermore, the procedure deals with the presence of missing values (Hapfelmeier *et al*., 2014). The results are presented as weights to facilitate interpretation.[2] Each weight associated with each variable ranges between 0 and 100 and adds up to 100.

4. Data and variables

4.1. PISA database

The IOpE is measured using data from the 2018 edition of the PISA assessment, conducted by the Organization for Economic Cooperation and Development (OECD), in which a total of 79 countries (39 from Europe) participated. In this edition, all Spanish regions participated in the assessment with an expanded sample to obtain regionally representative data (MEFP, 2019), which is relevant for our study.

PISA was designed and implemented by the OECD in the late 1990s as a periodic international comparative study of certain student's characteristics and skills. The target population for this assessment is students aged 15-16, who are completing compulsory education in most participating countries. PISA assesses performance in three core competencies: science, reading, and mathematics. It also provides additional information on students' socioeconomic, family, and school backgrounds through questionnaires administered to students, parents, teachers, and school principals. This assessment has been repeated every three years since 2000, and although each edition assesses all of the above competencies, each wave focuses on a specific one.

The level of skill acquisition is measured on a continuous scale with a mean of 500 and a standard deviation of 100 for the participating OECD countries. As explained earlier, for each student we have ten scores for each skill, called plausible values. They are randomly drawn from her score distribution, which is estimated from the responses to each test according to

the Item Response Theory (Rasch, 1960). These plausible values are understood to represent the range of students' abilities and should be fully used to estimate any population statistic (OECD, 2019). Table 1 shows the PISA 2018 sample size and the descriptive values of the science, mathematics and reading test scores for each of the Spanish regions.[3]

Table 1
DESCRIPTIVES OF PISA 2018 TEST SCORES BY REGION

	Science				Mathematics				Reading			
	Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Andalucía	477	90	185	747	473	88	175	726	473	93	171	741
Aragón	497	89	183	747	501	89	198	765	494	93	195	762
Asturias	499	88	212	751	494	88	186	754	498	92	199	764
Baleares	486	85	214	749	487	84	210	742	483	87	204	763
Canarias	472	87	204	772	463	86	183	718	474	87	206	749
Cantabria	501	86	214	757	505	85	196	752	489	88	185	755
Castilla y León	504	88	194	764	506	88	191	742	501	89	193	751
Castilla-La Mancha	488	88	214	751	483	88	202	729	482	91	193	745
Cataluña	491	91	179	764	492	88	212	763	487	95	168	751
Extremadura	478	87	210	734	474	86	182	739	469	90	184	729
Galicia	512	89	187	781	501	87	183	744	496	92	179	746
La Rioja	486	90	198	752	497	91	190	776	466	93	183	760
Madrid	499	89	165	801	501	89	175	780	487	92	172	782
Murcia	488	92	208	744	483	91	160	745	491	95	181	749
Navarra	496	87	216	750	507	86	225	764	476	93	187	742
País Vasco	488	85	183	750	499	85	181	763	476	91	164	757
Valencia	482	85	192	744	478	84	166	744	477	90	186	754
Ceuta	431	83	221	676	426	84	166	664	422	90	202	672
Melilla	452	87	219	688	447	86	222	672	452	88	205	664

Source: Own estimates based on PISA 2018.

4.2. The set of circumstances

Circumstances refer to those factors that are predetermined and beyond the students' control, such as the socio-economic and cultural status of the family, their gender or the school they attend. We consider the same set of circumstances as in Sicilia *et al*. (2022) and Marrero *et al*. (2024), which allows us to compare our results with theirs:

— Gender: a dichotomous variable that takes the value 1 if the gender is female, 0 otherwise.

— Immigrant status of the student: a dichotomous variable that takes the value 1 if the student was born in a country other than Spain, 0 otherwise.

— Immigrant status of parents: a dichotomous variable that takes the value 1 if one of the parents was born in a country other than Spain, 0 otherwise.

— Educational level of the mother (MEdu) and the father (FEdu): Parental education was the most widely used indicator as a unidimensional measure of family socioeconomic status, as it is closely related to other components of socioeconomic status. Following the International Standard Classification of Education (ISCED-11) we construct 4 categories from the students' responses: Primary education or no education: includes ISCED-11 levels 0 (no education) and 1 (primary education); Lower secondary education: includes ISCED-11 level 2 (compulsory secondary education); Upper secondary education: includes ISCED-11 levels 3 (upper secondary education, intermediate vocational training) and 4 (vocational certificate); Tertiary education: includes ISCED-11 levels 5 (higher vocational training) and 6 (university education –including master's degree and doctorate).

— Occupation of the mother (MOcc) and the father (FOcc): It reflects the socioeconomic status of a household (Sirin, 2005). We construct these circumstances from the students' responses, grouping parental occupations into three groups according to the skill level proposed by the International Standard Classification of Occupations (ISCO-08): Skill levels 3 and 4 (high) include ISCO-08 levels 1 (managers), 2 (professionals) and 3 (technicians and associates); Skill level 2 (medium) includes ISCO-08 levels 4 (administrative support workers), 5 (service and sales workers), 6 (skilled workers in agriculture, forestry and fishing), 7 (artisans and related trades) and 8 (plant and machine operators and assemblers); Skill level 1 (low): includes ISCO-08 levels 9 (Elementary Occupations) and 0 (Armed Forces).

— Household Wealth Index (Wealth): Is a composite index constructed by the OECD to capture the level of household wealth based on the possession of material goods in the household such as: a single room for the student, internet access, the number of bedrooms, bathrooms, televisions, cars, mobile phones with internet access, computers, tablets and e-books. Household possessions are often used as a proxy for income because the student often does not know the exact income of the parents (Hanushek *et al*., 2020). The index is standardized, with a mean of zero and a standard deviation of one for OECD countries.

— Number of books at home (Books): The following 4 categories were constructed: 0-25 books, 26-100 books, 100-200 books, and more than 200 books in the home.

— Household Cultural Possessions Index (Cultposs): This is a composite index constructed by the OECD on the basis of students' responses on the possession in the household of the following items: books of classical literature; books of poetry; works of art (e. g. paintings); books of art, music or design; and musical instruments (e. g. guitar, piano). This index reflects the cultural level of the household. The index is standardized, with a mean of zero and a standard deviation of one for OECD countries.

In addition to the individual characteristics and socioeconomic background of the students, we also consider the characteristics of the school. This is because school choice is generally a decision made by families and is beyond the control of students. However, the type of school they attend, as well as their environment, can strongly influence their academic performance (Dills, 2005). Following the literature, we have included two school characteristics:

— School type (Pubschool): A dichotomous variable taking value 1 if the school is subsidized (private management and public financing) or private, and a value of 0 if the institution is public (public management and public financing).

— Peer effect (Peer): To capture the effect of peers on students' skills and learning, the average socioeconomic status of school peers is used (Marrero *et al*., 2024). In this case, we use the mean of the Economic, Social and Cultural Status (ESCS).[4]

Since families' choices of schools for their children are related to the student's circumstances, we have removed this effect from the school variables (Marrero *et al.*, 2024). To do this, we independently regress *school type* and *peer effect* on the set of circumstances and take the residuals of each one (i. e., the part orthogonal to the circumstances). These residuals would capture other factors related to school choice that are independent of socioeconomic factors, such as the importance or value that families place on their children's education.

5. Results

The IOpE estimates are obtained for all three competences (reading, mathematics and science) and we find a great agreement in the IOpE levels and in the ranking of the regions, regardless of the competence area. For the sake of simplicity, we present the results for science, since it is a hybrid domain and produces results that are generally intermediate between those of the other two domains. The main results for the other competencies are available upon request.

The IOpE estimates and the significance of the circumstances are both affected by the type of sample used. Bootstrapping (random sampling with replacement) and the estimation of confidence intervals would reduce this concern. Thus, for each estimated CIT and CIF, we use the bootstrapping procedure to generate the confidence intervals. In our application, there is an additional complexity because we have 10 plausible values for educational achievement (PISA scores) for each individual and subject. Appendix 2 shows the detail to consider these plausible values in the generation of the confidence intervals.

5.1. Alternative scenarios for CIT and CIF

As one purpose of the research is to compare IOpE between regions, we need to use common values of α, N and M for all regions. We show results for three alternative scenarios, which is illustrative to understand the results of our preferred scenario (the tuning parameter scenario).

First, the *baseline scenario*, which consists of setting a small value of α, $\alpha = 0.01$, for CIT and CIF, and a value of M equal to 4 for CIF, which is in line with the "rule of thumb" of dividing the number of factors included by 3 (i.e., 12/3). The value of N is 150 (Appendix 2). In principle, this scenario minimizes potential overfitting problems while introducing downward bias. Second, the *unrestricted or unpruned scenario*, using $\alpha = 0.99$, while keeping the values of M and N as in the baseline. We consider this scenario as one that minimizes the downward bias while potentially incurring in upward bias. Third, a *parameter tuning scenario* that lies between the two previous scenarios and attempts to balance both types of bias (recall from Section 3). For each region, we estimate the tuning values of α_* (for CIT and CIF) and of M_* and N_* (for CIF) following steps 1-8 in section 3.1 for CIT and steps 1-4 in section 3.2 for CIF. We then use the average values of these parameters across regions: for CIT, a common α_* of about 0.05; for CIF, $\alpha_* = 0.80$, $M_* = 5$ and $N_* = 150$. The main results are robust to local variations around these values. Some details of the tuning procedure are given in Appendix 3 (for the CIT), Appendix 4 (for N_* in the CIF) and Appendix 5 (α_* and M_* in the CIF).

We interpret the large differences observed in the resulting α_* between the CIT and CIF approaches as follows. The CIF approach can make better use of the amount of information available without running into serious overfitting problems. By tuning the key parameters in the CIF, we can reduce the downward bias as much as possible, while at the same time being confident of not incurring serious upward bias. Therefore, among the procedures considered, we choose the CIF under the *parameter tuning scenario* as our preferred approach, and the main results in the following sections are based on this choice.

5.2. Estimation of IOpE

We present the main results of our estimation of the IOpE using the CIT and CIF models under the three alternative scenarios considered: *baseline, unrestricted and parameter tuning*. In all cases, the results are the average of those obtained using the ten plausible values in science, and the confidence intervals are obtained using bootstrapping. As a reference point, we take the results of Sicilia *et al*. (2022), who use the same set of circumstances as we do but apply the regression procedure of Ferreira and Guignoux (2014).

For the three scenarios considered, Figure 2 shows the results for all regions. We add the point estimates (marked with a cross) from Sicilia *et al*. (2022). In all graphics, we rank the regions from highest to lowest IOpE according to the average IOpE results in the CIF parameter tuning scenario (our preferred model). Table 2 complements this figure and shows the average IOpE estimates for the different approaches and scenarios considered.

As expected, for both the CIT and the CIF, the IOpE estimate increases significantly with the level of α (the depth of the tree).[5] For example, comparing the two extreme scenarios ($\alpha = 0.01$ and $\alpha = 0.99$), the IOpE estimates almost triple in all regions. This remarkable difference is also observed when comparing the results of our preferred ML estimates (the CIF under the *parameter tuning* scenario) with our reference estimates (Sicilia *et al*., 2022).

Next we compare the results for the CIT and the CIF. Although this is a technical issue, it helps to understand why the CIF (under the *tuning scenario*) is our preferred approach. If we compare the results of the CIT and the CIF, for the same level of α, we see that the mean IOpE for the CIF tends to be higher than the mean for the CIT. The CIF can better optimize the available information and adapt the academic performance data to the existing set of circumstances, reducing the potential downward bias and, more importantly, without incurring significant overfitting problems.[6]

The average IOpE estimates for $\alpha = 0.01$ (*baseline*) are even lower than those obtained in Sicilia *et al*. (2022) in almost all regions, both for CIT and CIF. This is a clear indication of a downward bias in the IOpE estimates for the baseline scenario. In this case, the resulting IOpE estimates can be considered as lower bounds. For illustrative purposes, we include the estimated trees for this scenario (Appendix 6), which we analyse later.

We present estimated trees for $\alpha = 0.01$ because they are of an acceptable size, allowing us to visually inspect and characterize the different resulting types of students (close to be the most representative ones). This inspection can provide valuable information that complements the variable importance analysis provided by the random forest (Section 5.3). For $\alpha = 0.99$, we obtain the opposite result: very high average estimates of the IOpE for CIT and CIF. Therefore, the IOpE under this scenario can be considered as an upper bound.

For the CIT, using the tuning alpha of 0.05, the average values are almost the same as in Sicilia *et al*. (2022). Small differences are observed in La Rioja, Navarra, País Vasco and Castilla-La Mancha. For the CIT, the average estimated IOpE values are around 0.20, ranging from 0.28 in Murcia to 0.17 in Castilla-La Mancha and Valencia. For the CIF, using the tuning alpha of 0.80 (and $M = 5$ and $N = 150$), the estimated IOpE levels are, as expected, much higher: in most cases almost double those obtained for the CIT under the tuning scenario and in Sicilia *et al*. (2022). Moreover, the confidence intervals for the CIF are also much smaller.[7]

Figure 3 shows the maps of IOpE estimates (their means from bootstrapping) using the CIT and CIF under the *parameter tuning* scenario. For comparison purposes, we also show the map using the results in Sicilia *et al*. (2022). To illustrate the differences in IOpE estimates between the different approaches, we use the same scale in the tree maps. We find significant differences in the estimated IOpE across regions, although no clear geographical patterning of IOpE across regions is found. Thus, the differences in IOpE are not just a north-south story; additional factors (for example, related with different regional educational policies) seems to be behind the observed differences.

However, these differences are significantly smaller than those found in other papers comparing European countries (e. g., Marrero *et al*., 2023).[8] Moreover, they appear to be smaller than those observed for GDP per capita, poverty rates or overall income inequality between Spanish regions. From a policy point of view, this has a double meaning: on the one hand, it is positive that the differences between Spanish regions in terms of IOpE are moderate. On the other hand, it is worrying that similar levels of IOpE are then reflected with

different intensity in other macromagnitudes. Understanding the reasons for these different transmission channels is extremely important, but is beyond the scope of this paper.

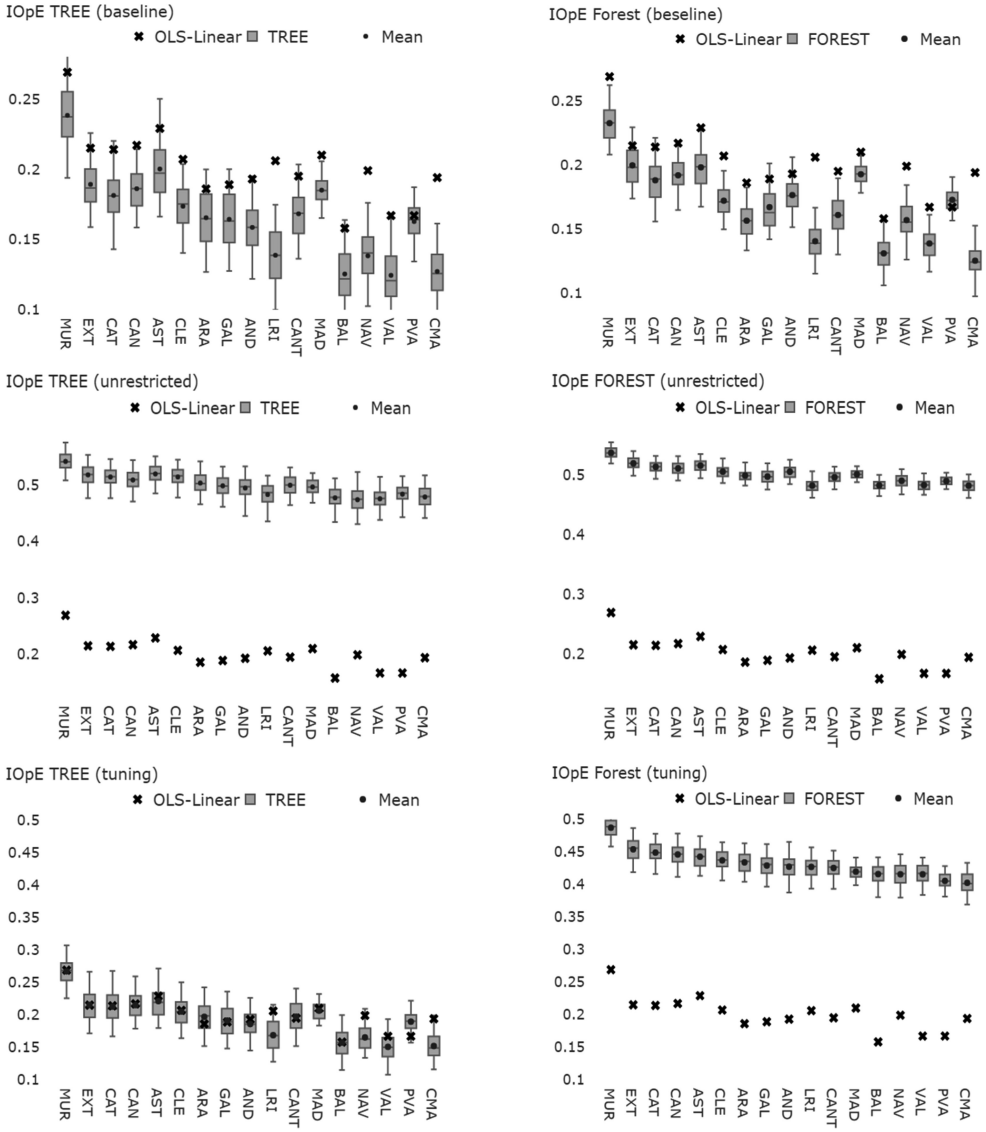
The general trends in the rankings are maintained when comparing CIT with CIF and when comparing the different scenarios, which gives robustness to this relevant result. In general, the regions with the highest IOpE are Murcia and Extremadura, followed by Cataluña, Canarias and Asturias. The regions with the lowest IOpE are Castilla-La Mancha and the País Vasco, followed by Valencia, Navarra, Baleares and Madrid. Comparing the ranking under the CIF *tuning* estimates with that obtained in Sicilia *et al*. (2022), we find some relevant changes: Castilla-La Mancha and Madrid show a significant improvement in the ranking (gaining 6 and 8 positions, respectively); on the contrary, Baleares, Aragón and Galicia show the opposite (losing 4, 7 and 5 positions, respectively).

Table 2
INEQUALITY OF OPPORTUNITY IN EDUCATIONAL ACHIEVEMENT IN SPANISH REGIONS UNDER ALTERNATIVE APPROACHES AND SCENARIOS

Spanish region	CIT Baseline	CIT Unrestricted	CIT Tunning	CIF Baseline	CIF Unrestricted	CIF Tunning	OLS-linear (Sicilia *et al*., 2022)
Andalucía	0.1586	0.4941	0.1859	0.1764	0.5052	0.4272	0.1930
Aragón	0.1655	0.5029	0.1970	0.1564	0.4988	0.4338	0.1860
Asturias	0.2002	0.5190	0.2212	0.1981	0.5155	0.4425	0.2290
Baleares	0.1254	0.4767	0.1572	0.1309	0.4823	0.4158	0.1580
Canarias	0.1861	0.5084	0.2152	0.1919	0.5108	0.4460	0.2170
Cantabria	0.1683	0.4991	0.1972	0.1609	0.4959	0.4252	0.1950
Castilla y León	0.1736	0.5137	0.2052	0.1720	0.5054	0.4369	0.2070
Castilla-La Mancha	0.1272	0.4783	0.1520	0.1251	0.4817	0.4024	0.1940
Cataluña	0.1813	0.5136	0.2130	0.1880	0.5134	0.4489	0.2140
Extremadura	0.1892	0.5174	0.2155	0.1997	0.5192	0.4539	0.2150
Galicia	0.1644	0.4980	0.1912	0.1670	0.4971	0.4290	0.1890
La Rioja	0.1388	0.4827	0.1687	0.1404	0.4819	0.4268	0.2060
Madrid	0.1851	0.4960	0.2066	0.1927	0.5009	0.4195	0.2100
Murcia	0.2383	0.5412	0.2672	0.2326	0.5369	0.4871	0.2690
Navarra	0.1384	0.4735	0.1653	0.1569	0.4900	0.4155	0.1990
País Vasco	0.1628	0.4832	0.1896	0.1727	0.4895	0.4054	0.1670
Valencia	0.1245	0.4751	0.1507	0.1386	0.4827	0.4155	0.1670

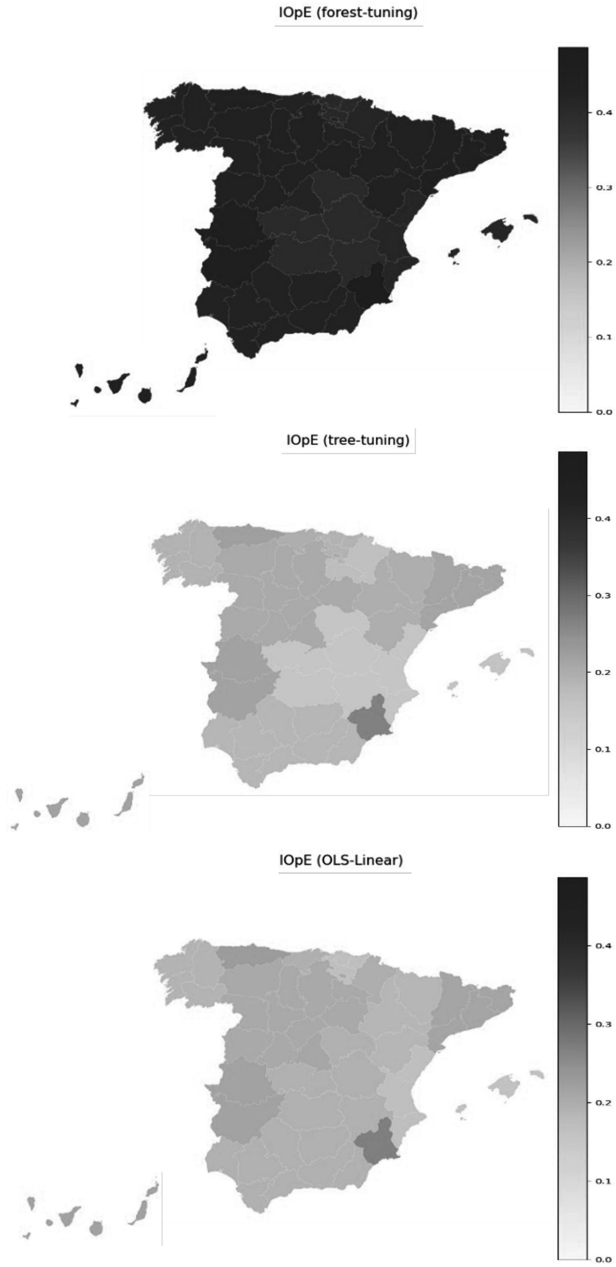
Note: CIT refers to the Conditional Inference Tree approach; CIF to the Conditional Inference Forest. The scenarios are as follows: the baseline scenario sets a $\alpha = 0.01$, for CIT and CIF, and a $M = 4$ and $N = 150$ for CIF. The *unrestricted (or unpruned scenario)* sets an $\alpha = 0.99$, while keeping the values of M and N as in the baseline for the CIF. In the *parameter-tuning* scenario (close to minimize the out-of-sample MSE for all regions), we choose a common value of α, N and M: $\alpha = 0.05$ for the CIT, and an $\alpha = 0.80$ for the CIF; $M = 5$ and $N = 150$ for the CIF. The numbers represent the means of the bootstrapped IOpE distributions for each region under each method and scenario (see Figure 2). The last column refers to point estimates in Sicilia *et al*. (2022) under a traditional linear regression approach.

Figure 2
INEQUALITY OF OPPORTUNITY IN EDUCATIONAL ACHIEVEMENT IN SPANISH REGIONS UNDER ALTERNATIVE APPROACHES AND SCENARIOS



Note: The regions are ordered from the highest to the lowest average value of the IOpE estimated using the CIF under the *parameter-tuning scenario*. The scenarios are as follows: the baseline scenario sets a $\alpha = 0.01$, for CIT and CIF, and a $M = 4$ and $N = 150$. The unrestricted (or unpruned scenario) sets an $\alpha = 0.99$, while keeping the values of M and N as in the baseline for the CIF. In the parameter-tuning scenario (close to minimize the out-of-sample MSE for all regions), we choose a common value of α, N and M: $\alpha = 0.05$ for the CIT, and an $\alpha = 0.80$ for the CIF; $M = 5$ and $N = 150$ for the CIF. The boxes represent the bootstrapped IOpE distribution for each region. Dots are the mean IOpE estimates of the distribution. Crosses are the point estimates in Sicilia *et al*. (2022) under a traditional regression approach.

Figure 3
INEQUALITY OF OPPORTUNITY IN EDUCATIONAL ACHIEVEMENT IN SPANISH REGIONS UNDER ALTERNATIVE APPROACHES AND SCENARIOS



Note: See note in Figure 2. We use the means of the IOpE empirical distribution (the dotted points in Figure 2) in the maps. We use the same scale in the tree maps to see the differences in IOpE estimates between the different approaches.

5.3. The importance of circumstances on educational performance

We measure the importance of each circumstance based on its ability to predict educational performance. We have a measure of the degree of uncertainty in their estimates using bootstrapping. A priori, the most influential factors have the greatest impact on the IOpE. The importance of each circumstance is measured as a percentage of the predictive ability of all circumstances, so the contributions add up to 100 for each region. In terms of which types of circumstances are most important, we look for common patterns across regions. We also analyse the correlation between circumstance importance and IOpE size across regions. Finally, we relate these results to the estimated trees under the baseline scenario.

We first emphasize that, contrary to what we obtain for the level of IOpE, the importance of each circumstance is quite robust to the value of α in CIT and CIF. In general, the groups of most and least important circumstances remain the same regardless of the value of α. The results are very robust to the type of scenario and even to the approach (although some differences are observed when comparing the CIT with the CIF). We therefore focus on the results for our preferred CIF model (*parameter tuning*).

Figure 4 summarizes, for each circumstance, the values of its importance, ordering the regions in the same way as in Figure 2 (from highest to lowest IOpE level according to the CIF *tuning procedure*). For each region, we show the distribution of the mean values of each circumstance over the total (in stacked bars). Figure 5 aggregates results across regions and circumstances to highlight common patterns and key differences between circumstances. We highlight the main findings.

First, a common pattern is that the most relevant factor is the number of books at home (Books), with values close to 30-40% in several cases, followed by the possession of culture and goods at home (HomCulPos and HomPos) and the parents' occupation (MOcc, FOcc), with percentages around 5-20%. For the rest, the percentages are mostly below 10%. However, it is worth noting some differences between the regions with regard to specific circumstances. For example, regarding to gender, the importance of this variable is above 10% (on average) in the País Vasco, Cataluña, Galicia, Aragón, Castilla-La Mancha and Asturias; with regard to immigration, it is above 5% in Murcia, La Rioja and the País Vasco; the peer effect is close to 10% in the Canarias, Madrid and Navarra; similarly, the relevance of the type of school is close to 10% in the Canarias and slightly lower in Valencia and Navarra. The importance of these circumstances is close to zero in the rest of regions. A relevant result is that, controlling for all other circumstances, the contribution of the wealth index at home is below 5% for most regions: for Canarias, Baleares, Aragón, Galicia, Castilla-La Mancha and La Rioja, the contribution is between 5%-8%. A similar result is found for the father's and mother's education, which are not relevant factors once all other circumstances included in the model are controlled for: their contributions are below 5-10% in most regions.

Second, there is no clear overall pattern between the importance of the circumstance and the level of IOpE, at least when we divide the sample into broad groups: high, medium and low levels of IOpE. Thus, we observe regions with a similar distribution of importance levels

across these 3 levels of IOpE. However, within each group there seems to be some correlation. For example, looking at the low IOpE regions, we find that the importance of public school attendance, occupation and parental education decreases with the level of IOpE. On the contrary, IOpE increases with the importance of household wealth, possession of cultural and other goods in the household and possession of books. The importance of the other circumstances remains stable. Understanding the reasons for this result, which goes beyond the purpose of this research, deserves further exploration by looking at specific characteristics and policies within each region.

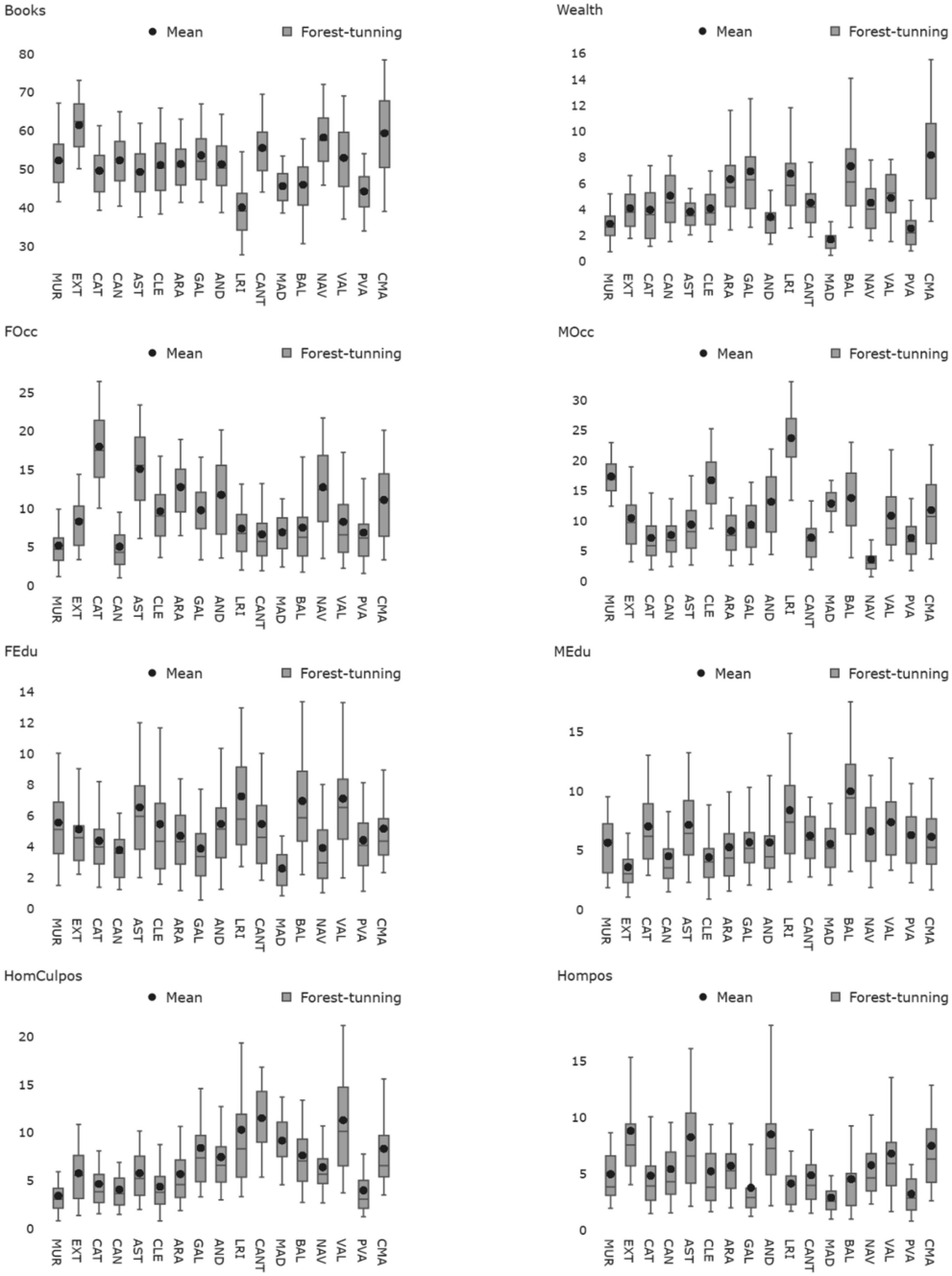
To conclude this section, we link our CIF results to estimated trees under the *baseline* scenario (Annex 4). One shortcoming of the CIF procedure is that it cannot provide an estimated tree. As explained above (Figure 1), a tree is useful because it provides a visual characterisation of the different types of individuals (students in our case) and we can easily associate each type with a particular set of circumstances or characteristics. This set of circumstances may be different for different types and different regions. However, conclusions drawn from a particular tree could be misleading if they are far from the results obtained from a better model, such as our preferred CIF under the *tuning scenario*. According to our results, we do not trust the results for estimated levels of IOpE under the CIT, as they are much smaller. However, the differences are very small when looking at the relevance of each circumstance in explaining the differences in IOpE, as commented above. Thus, our estimated trees can be a good complement to the results described above.

Looking at the trees, although the order of the hierarchy that exists in the tree is not always the same as that of the degree of importance of each circumstance in each region (Figures 4 and 5), we find a very important parallelism between the two. The first finding to note is that the most relevant circumstance (Books) is the initial node in almost all trees. In addition, the other two most relevant circumstances (Hompos and HomCulpos) also appear in many relevant positions in the trees.

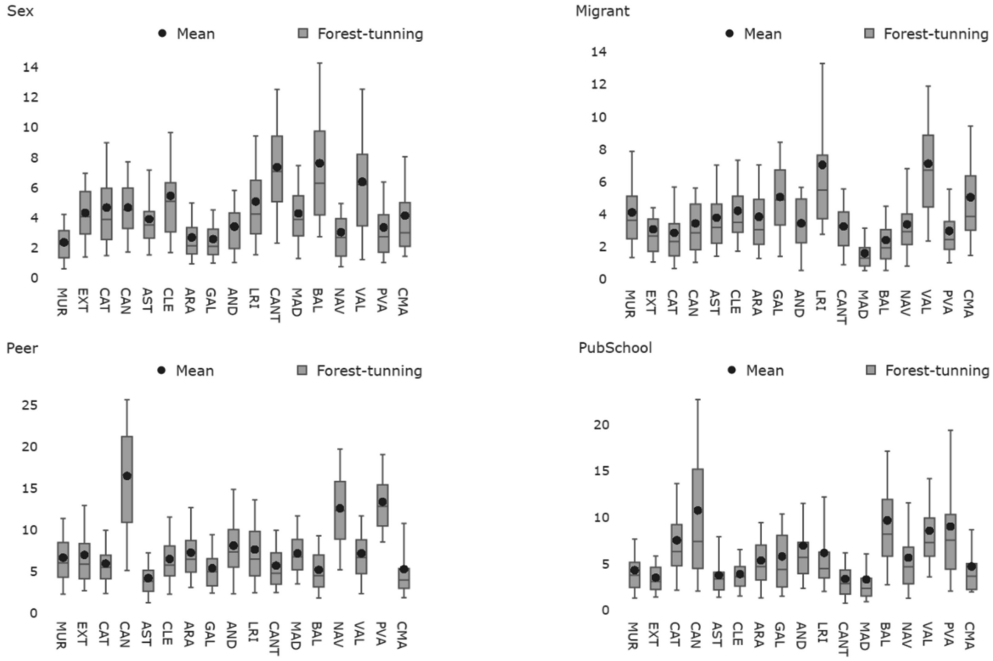
Let us also give some examples associated to regions. In the case of Andalucía, Figure 5 highlights the Books, HomPos, HomCulPos and Mocc circumstances that appear in its tree. In the case of Aragón, Figure 5 highlights that gender is the second most important (a circumstance that is not very important in almost any region), and this fact is also highlighted in the tree of Aragón. In Baleares, the tree is relatively smaller than in the other regions, but the circumstances that appear first in the tree (Book, MOcc and FOcc) are the most relevant in the region. In addition, the circumstances MEdu and Wealth, which have a relevance of more than 5% almost exclusively in Baleares, appear in a lower hierarchical order in the tree.

In the case of Canarias and Navarra, it is worth highlighting the presence of the Peer circumstance in prominent positions in their trees since this circumstance is almost only relevant in these two regions (Figure 4). In the País Vasco, the circumstance Migrant also appears in its trees (and does not appear in practically any other region), which is relevant almost only in this region (and to a lesser extent in Murcia and the Canarias). As a final example, Castilla-La Mancha has a group of circumstances that have a relevance of around or above 10% according to Figure 5 (Book, Hompos, HomCulpos, FOcc) and most of them appear in its tree representation.

Figure 4
IMPORTANCE OF EACH CIRCUMSTANCE ON THE IOPE ESTIMATE USING THE CIF (*PARAMETER-TUNING*)

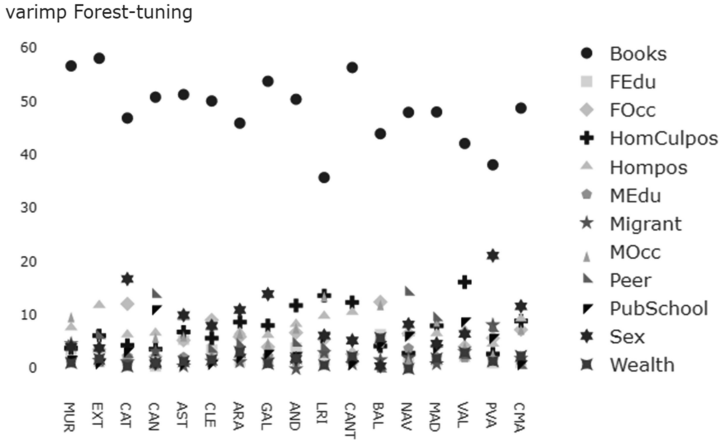


(Continued)



Note: Results are show for the parameter tunning scenario for the CIF (see note in Figure 2). The boxes represent the bootstrapped IOpE distribution for each region. Dots are the mean IOpE estimates of the distribution. Books: books at home; HomCulPos: cultural procession at home; HomePos: possession of good at home; Mocc: mother occupation; Focc: father occupation; PubSchool: public schools asistance; Peer: peer effect of students in the class; Sex: gender of the student; Migrant: immigrant status of the parents; Wealth: home wealth index.

Figure 5
SUMMARY OF THE IMPORTANCE OF EACH CIRCUMSTANCE IN IOPE ACROSS SPANISH REGIONS



Note: See note in Figure 4.

6. Conclusions

Inequality of opportunity in educational achievement (IOpE) refers to the part of inequality in learning that is generated by differences in factors (circumstances) given to students, regardless of their talent and effort. Using the 2018 PISA data, we estimated the IOpE for Spanish regions by applying ML methods, specifically the conditional inference tree (CIT) and the forest (CIF). These methods have recently been used in the income inequality literature to overcome the problems of upward and downward bias of traditional methods, which are based on linear regression. However, to the best of our knowledge, they have not yet been applied to measure IOpE, and the education literature can benefit from this.

The main result is that our IOpE estimates under our preferred ML specification (CIF using *tuning parameters* to minimize out-of-sample MSE) are more than twice the IOpE estimates under linear regression methods in most regions. On average, the IOpE in science for our preferred ML specification has a range of variability across regions from 40.5% to 48.7%, with Murcia, Extremadura, and Cataluña at the top, while Castilla-La Mancha and the Pais Vasco have the lowest levels of IOpE. Using traditional linear regression approaches, estimates of IOpE range from around 16% to 27%. The differences do not seem to be related to any geographical, economic or socio-cultural pattern. In this regard, the differences could be attributed to differences in educational policies (e.g. access to early education, scholarships, the school selection criteria, among others) and educational practices within schools (e.g. grouping criteria, teaching practices, etc.). The analysis of the causal impact of these factors on IOpE falls outside the scope of this work, but it is a future line of research.

The results show that the differences in IOpE between Spanish regions are smaller than those observed between European countries, but also for Spanish regions when using traditional approaches. This result raises the question of whether the substantial differences observed in IOpE between European countries stem from actual disparities or from the methodology employed. Although differences in IOpE across regions are smaller using ML methods, they are not insignificant. Differences in the cultural environment at home (proxy for the number of books at home and other cultural possessions at home) is the aspect that contributes most to IOpE in all regions, followed by parental occupation. Other circumstances, such as gender and migrant status, are relevant in only a few regions. There is no clear pattern between the importance of circumstances and the level of IOpE.

Another notable result is the higher precision of the estimates of the CIF procedure compared to the CIT procedure. These results suggest that the use of CIF not only overcomes the problems of downward and upward bias of linear regression techniques, but also provides more accurate estimates of IOpE. That is, the procedure appears to be able to balance the magnitude of the two biases and thus provide unbiased estimates of the IOpE given the set of circumstances included. This result is extremely important, since the first step in designing effective policies to reduce IOpE is to obtain a good measure of it.

It is important to note that our analysis is focused on providing a more efficient method for measuring inequality of opportunity in education and is therefore primarily descriptive

in nature. Consequently, the results of this research should be viewed as a first step towards studying the IOpE using ML techniques. Therefore, the findings of this research provide solid evidence for the use of novel ML in the field of educational inequality. Although it is not policy-oriented, our work may be relevant in identifying which groups are most vulnerable and at risk of falling behind. If a particular circumstance, such as gender, is found to be insignificant, this means that differences in educational achievement between its groups (e.g. boys and girls) are minimal, and there is no need to be concerned about one group falling behind the other. However, if a circumstance is highly important (e. g. having books at home), schools and policymakers should implement targeted policies to provide additional support to students from families with a poorer home cultural environment, preventing them from falling behind and increasing inequality of opportunity.

Finally, since circumstances cannot be easily influenced by policymaker, it would be worthwhile to study the role of the channels of transmission in the future (e.g. grade repetition, enjoyment of reading, educational or occupational expectations, technological skills, etc.). The channels are factors through which differences in circumstances are ultimately translated into differences in student performance (Marrero *et al*., 2024). In this regard, investigating the mechanisms behind the IOpE using ML methods is a natural extension of this paper.

Appendix 1. Methodological details

Part A. The partition approach of the Conditional Inference Tree

The variable selection procedure in conventional decision trees (CART trees) has been shown to be biased in favour of variables that are continuous or have multiple categories (Breiman, 2001; Strobl *et al*., 2007). To overcome this problem, Hothorn *et al*. (2006) propose an alternative procedure based on so-called recursive binary partitioning, which results in conditional inference trees (CIT) that do not exhibit these biases. Specifically, the partition is chosen according to the strength of the association between the explanatory variable and the target variable, using non-parametric tests (such as chi-squared) to test all possible associations. Briefly, the steps are as follows (see Hothorn *et al*., 2006, for more details):

1. First, we test the null hypothesis, H_o, of overall independence between the target variable E and the entire set of circumstances, $\Phi(E|(C_1,...C_j,...,C_J) = \Phi(E)$, where $\Phi(\cdot)$ is the conditional and unconditional distribution function of E. Given a significance level α, if H_o is not rejected, the procedure ends, and no splitting is made. Note that this is a way of testing for overall IOpE.

2. If the H_o is rejected, the p-values of all the independence tests between E and each of the circumstances C_j, $H_o^j: \Phi\left(E\middle|C_j\right) = \Phi(E)$, are computed, yielding a sequence of p-values $\left\{p_j\right\}_{j=1}^{J}$, one for each circumstance C_j.[9]

3. We choose the circumstance C_{j*} with the smallest p-value, p_{j*}, which represents the one with the highest association conditional on the node where the tree is located.

4. For the chosen circumstance, we determine the partition point, $\bar{c}_{j*}$, which makes the groups in the generated split (i.e., $C_{j*} < \bar{c}_{j*}$ versus $C_{j*} \geq \bar{c}_{j*}$) as different as possible in terms of E.[10]

5. The algorithm is repeated until, for each intermediate node and for each of the circumstances included in the model, we do not reject the H_o of independence. When this happens, the tree stops growing and the resulting node is called the "terminal node". It is the node that allows us to summarize the final output (educational achievement) of a group of individuals who share a set of common circumstances.

Part B. Regularization the size of the tree

We summarise this procedure as follows:

1. Set up a grid of possible α values. In practice, we will use a grid ranging from 0.01 to values below 0.99 (see Appendices 1 and 4 for more details).

2. Divide the total sample (completely randomly) into K equal parts (5 in our application).

3. Estimate the CIT using a subsample containing observations from K-1 partitions (training sample).

4. Predict the target variable E for the K-th omitted partition (the test sample), $\hat{E}^{test}$, compute the prediction errors, $E^{test} - \hat{E}^{test}$, and compute the corresponding (out-of-sample) MSE.

5. Repeat this procedure, each time excluding a different partition.

6. Calculate the mean of the K out-of-sample MSE values obtained.

7. Repeat steps 2-6 for each α.

8. Select the α that minimises the mean of the MSEs in the CV, which we denote by α_*.

Part C. The tuning approach of parameters in the conditional Inference Forest

The CIF, by considering random subsamples in each tree, already allows us to calculate the out-of-sample forecast error with the observations it leaves out (out-of-bag, OOB), which is about 1/3 of the total sample, so it is not necessary to perform a cross-validation exercise. Thus, for each tree estimated in the CIF, we compute the OOB prediction errors, $E^{OOB} - \hat{E}^{OOB}$, and calculate the MSE based on these errors, which are also out-of-sample. The tuning of these parameters is done in two steps, following Brunori *et al*. (2023):

1. A value of N is obtained for which the out-of-sample MSE is practically unchanged, which we denote as N_*. In our case, this value is around 150, i.e., $N_* = 150$.

2. Given N_*, we compute the predictions and associated MSE for a grid of α and M values.

3. We choose the values of M and α that minimise the OOB MSE, which we denote as M_* and α_*.

4. For M_* and α_*, and given N_*, we estimate and predict the CIF.

With all these parameters fixed, we can retrieve the prediction of all educational achievements for all students in the sample, $\hat{E}^{CIF}$, calculate its variance and obtain its ratio to the total variance, $\text{var}(\hat{E}^{CIF})/var(E)$. Note that the fitted values and these variances are calculated for the whole sample (i.e., they are not out-of-sample predictions). The out-of-sample predictions are only used to calculate the MSE and to refine the key parameters of the CIT and CIF.

Appendix 2. Bootstrapping, plausible values and confidence intervals

For each estimated CIT and CIF, we use the bootstrapping procedure to generate the empirical distributions and obtain confidence intervals (i.e., the difference between the 5th and 95th percentiles of the distribution obtained) of the estimated IOpE and the importance of each circumstance. We perform 200 bootstraps in each case.

In our application, there is an additional complexity because we have 10 plausible values for educational achievement (PISA scores) for each individual and subject. Report student performance through plausible values is a standard practice in international surveys such PISA. For each subject, each student answers a limited number of items from which a distribution is computed. Plausible values are randomly extracted from this distribution of results, and they represent the entire range of each student's skill or achievement. Hence, all values must be considered to estimate any population statistic (OECD, 2019). In our setting, for each region and subject, any estimator (e.g., mean, percentiles, variance) must be calculated by computing it separately for each of the ten plausible values and then taking the average. Thus, for each estimator (e.g., for the IOpE), we estimate the IOpE ten times (one for each plausible value) and calculate the average. Taking this aspect into account significantly increases the computational time, so we have only considered this aspect in this last step.[12] We therefore proceed as follows:

1. We start with the first iteration of bootstrapping, $nboot = 1$.

2. We estimate the IOpE (and other estimators, such as the importance of each circumstance) for the ten plausible values.

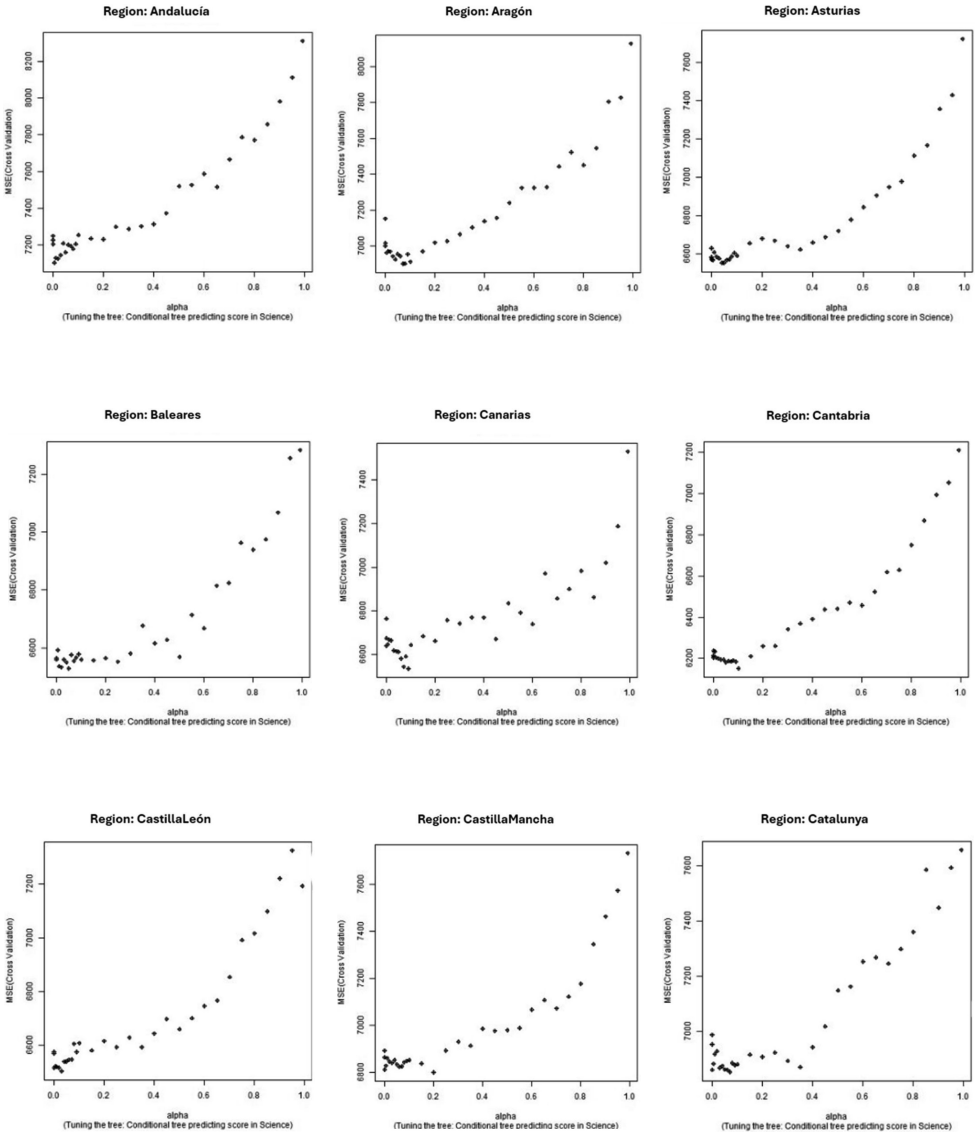
3. For each estimator, we take the average of the ten estimates, the resulting values being the realizations for each $nboot$.

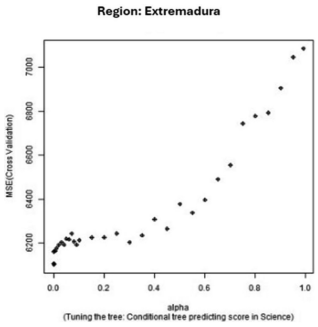
4. We repeat steps 1-3 for each bootstrap ($nboot = 1, ..., 200$).

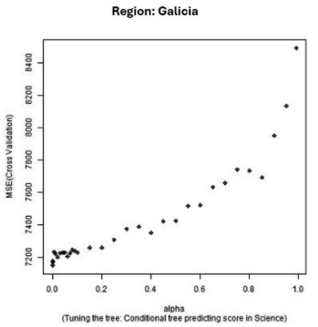
For all 200 realizations, we obtain the empirical distributions of each estimator and calculate the confidence interval as the difference between the 5th and 95th percentiles.

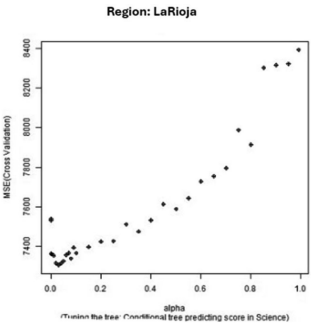
Appendix 3. Tunning the level of α in the CIT for each region

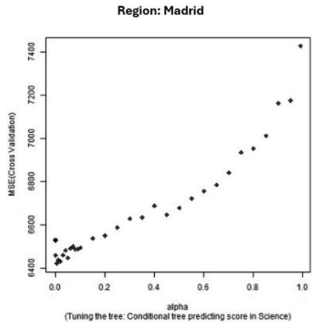
For each Spanish region, we show the relationship between the out-of-sample MSE and the level of regularization α (using Cross Validation averages). See Section 3 for details. For the CIT, we obtain that the level of α_* is less than 0.1 for almost all Spanish regions. We show that the relationship between the out-of-sample MSE and the level of α is clearly increasing for values of α above 0.1 in most regions. At least in our context, the small level of α_* indicates that the CIT procedure easily generates overfitting problems.

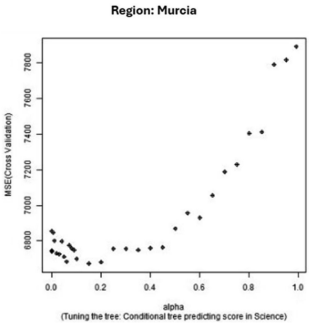


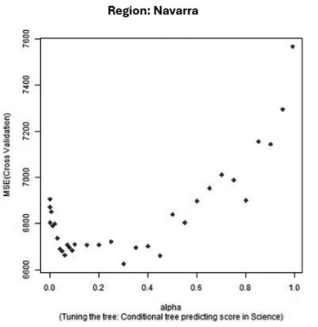

Region: Extremadura
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)

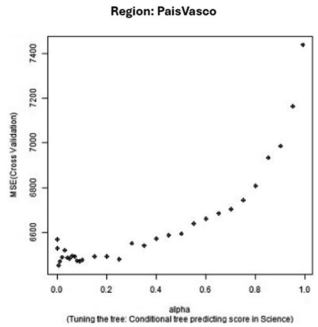

Region: Galicia
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)

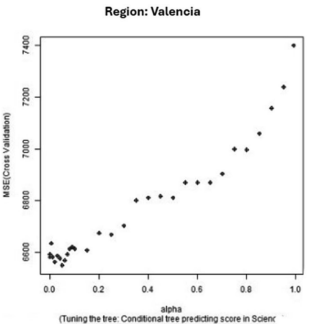

Region: LaRioja
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)


Region: Madrid
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)


Region: Murcia
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)

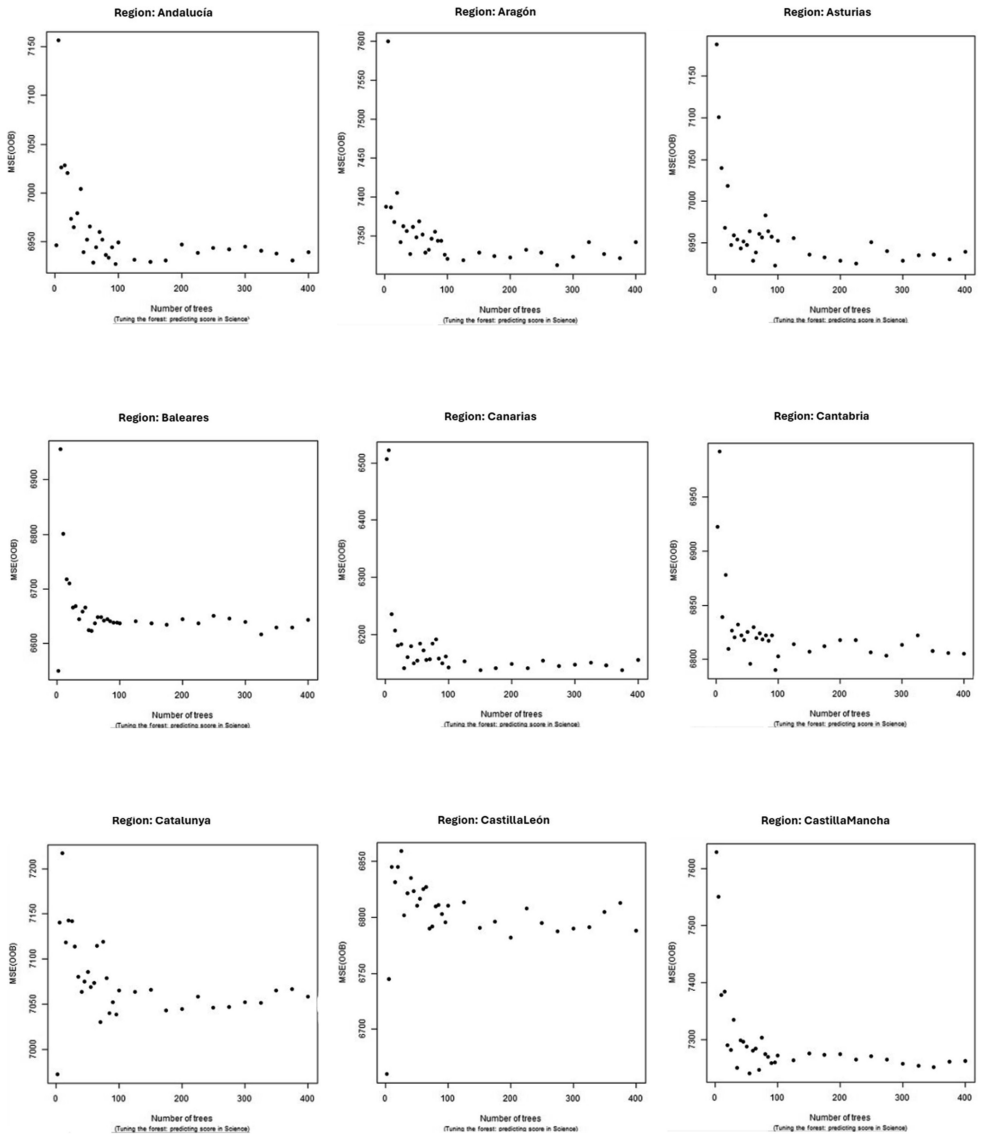

Region: Navarra
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)

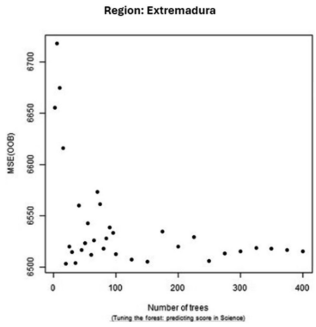

Region: PaisVasco
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Science)

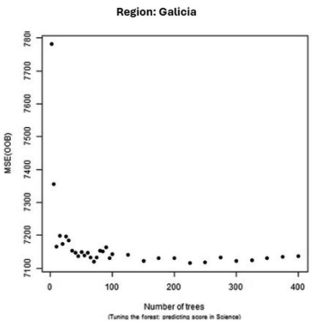

Region: Valencia
MSE(Cross Validation)
alpha
(Tuning the tree: Conditional tree predicting score in Scienc

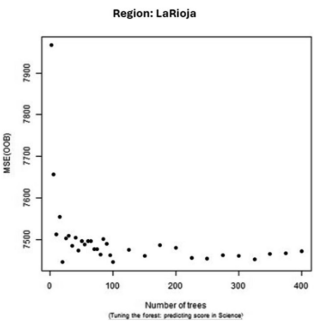
Appendix 4. Tunning the number of trees (*N*) in the CIF for each region

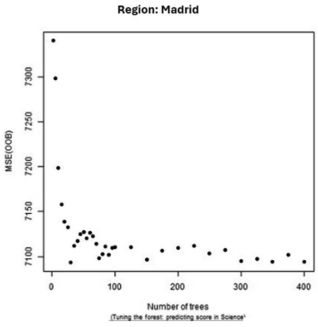
For each Spanish region, we show the relationship between the out-of-bag MSE and the number of trees in the forest. We show this relationship for a given level of α and M (under the baseline scenario), but the shape of the relationship remains practically invariant. See Section 3 for details. For the CIF, the tuning procedure is more complex. We first adjust the number of trees, N. For all regions and given a grid of α and M, we observe that the MSE hardly changes from a value of N between 100 and 200 on. For this reason, we choose $N_* = 150$.

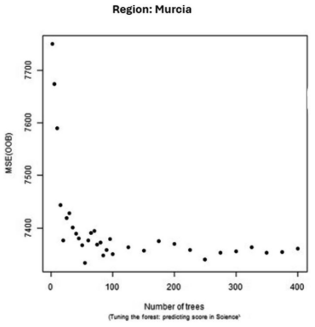


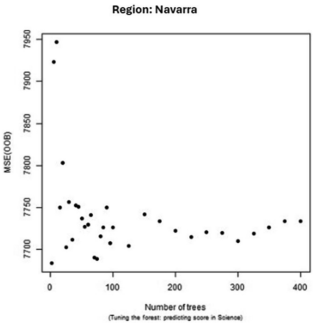

Region: Extremadura
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)

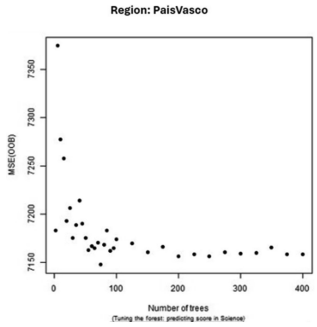

Region: Galicia
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)

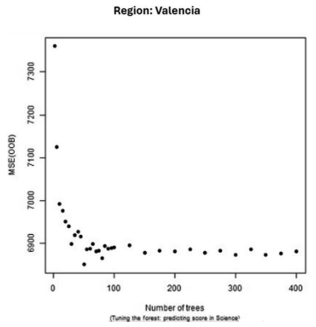

Region: LaRioja
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)


Region: Madrid
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)


Region: Murcia
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)


Region: Navarra
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)

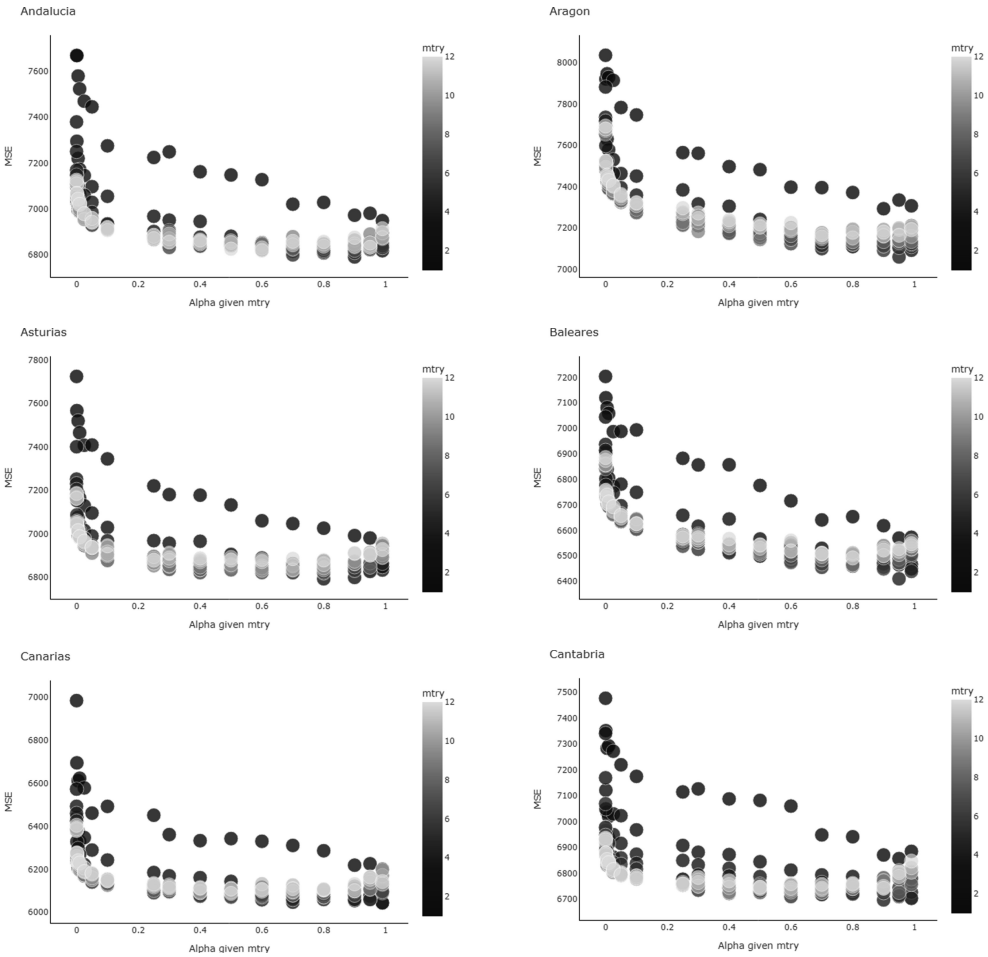

Region: PaisVasco
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)

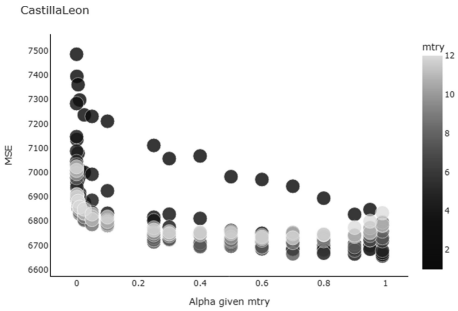

Region: Valencia
MSE(OOB)
Number of trees
(Tuning the forest: predicting score in Science)

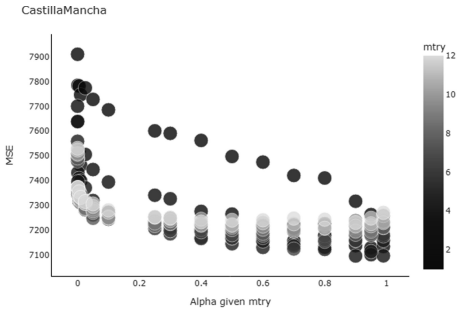
Appendix 5. Tunning the level of α and M, given the value of N in the CIF for each region

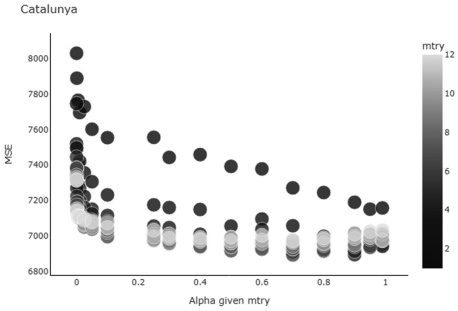
For each Spanish region, we show the relationship between the out-of-bag MSE and the value of α (the key regularization parameter), for different levels of M –"mtry" in the graphics (the number of circumstances that are randomly permuted in each random subsample to estimate the CIF), given the number of trees to include ($N = 150$). See Section 3 for details.

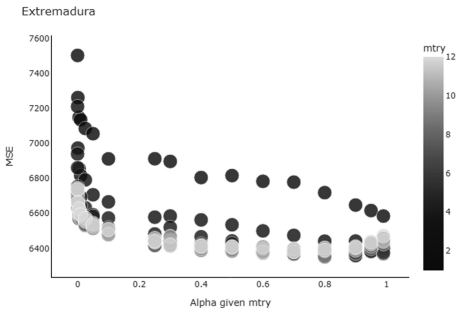
Given the value of N_*, the tuning of the values of α_* and M_* are done simultaneously. We observe that, for a given value of α, the out-of-sample MSE varies very little for different values of M. However, this MSE is significantly reduced for values of α approximately below 0.2 and, in contrast to what happened for the CIT, this function remains relatively flat for values of α above 0.2. At the other extreme, for very high values of α, around 0.8, we observe a small rebound in the MSE in most cases.

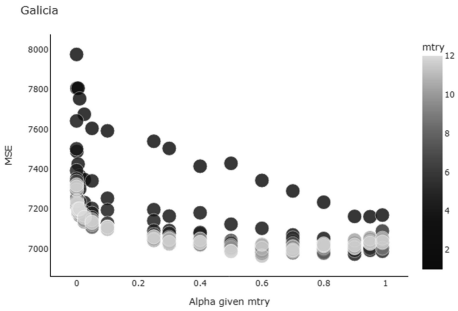


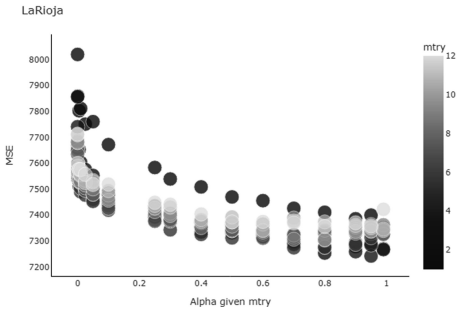

CastillaLeon
MSE
Alpha given mtry
mtry

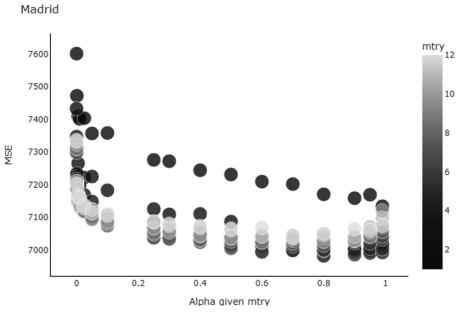

CastillaMancha
MSE
Alpha given mtry
mtry

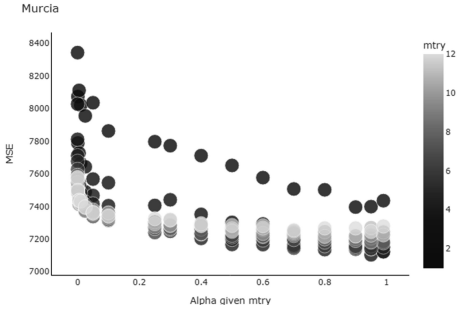

Catalunya
MSE
Alpha given mtry
mtry

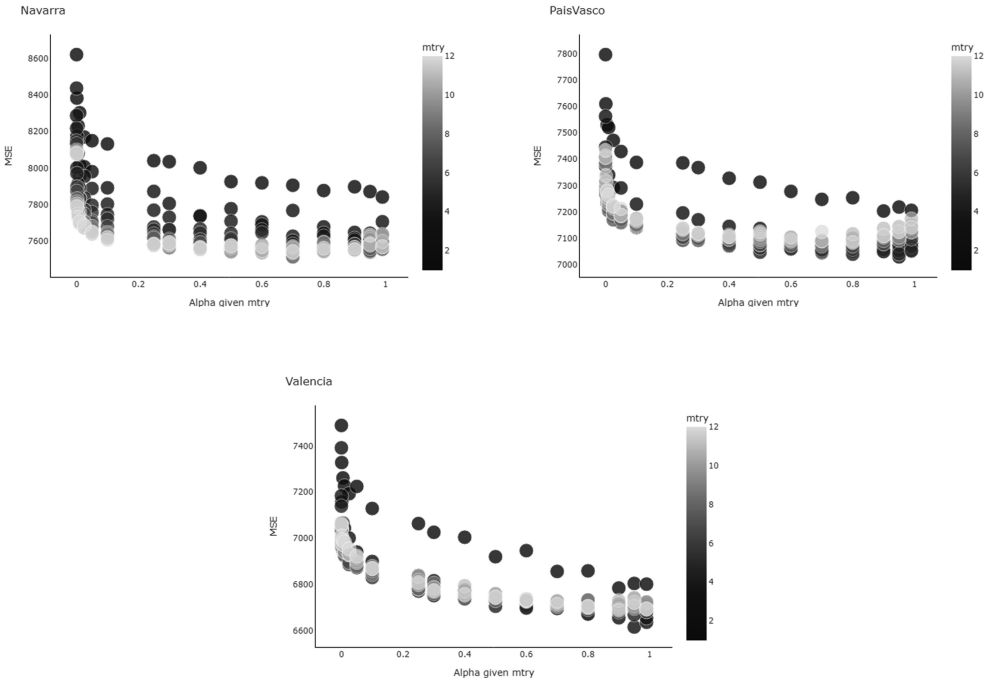

Extremadura
MSE
Alpha given mtry
mtry


Galicia
MSE
Alpha given mtry
mtry


LaRioja
MSE
Alpha given mtry
mtry


Madrid
MSE
Alpha given mtry
mtry

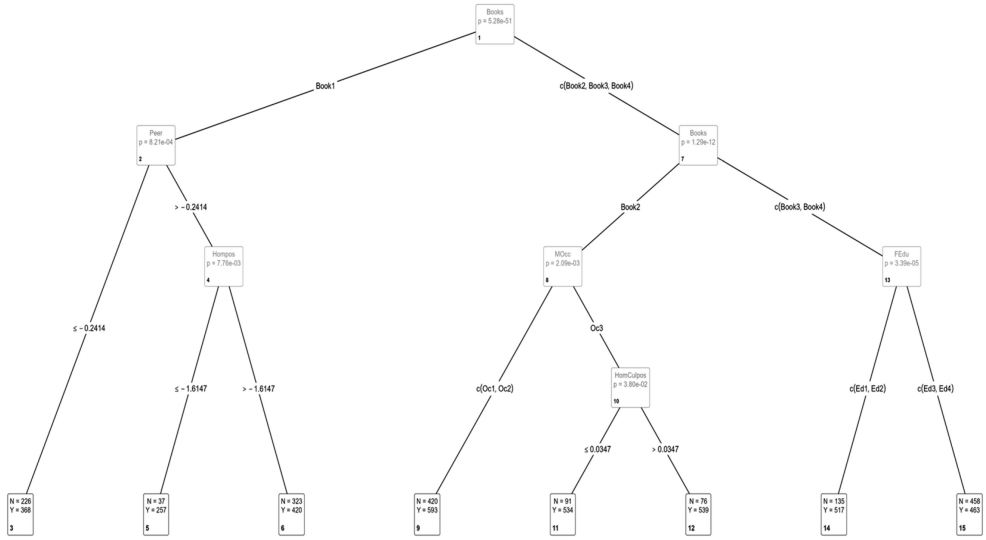

Murcia
MSE
Alpha given mtry
mtry



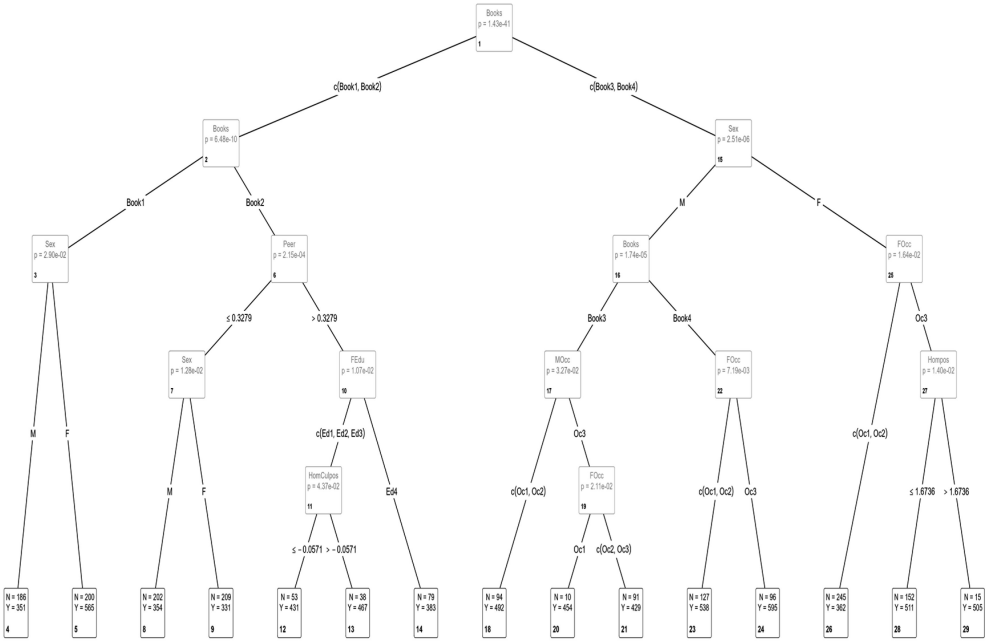
Note: mtry (values for the different curves) refers to "M": the number of circumstances that are randomly permuted in each case random subsample to estimate the CIF.

Appendix 6. Estimation of the CIT under the baseline scenario for each region

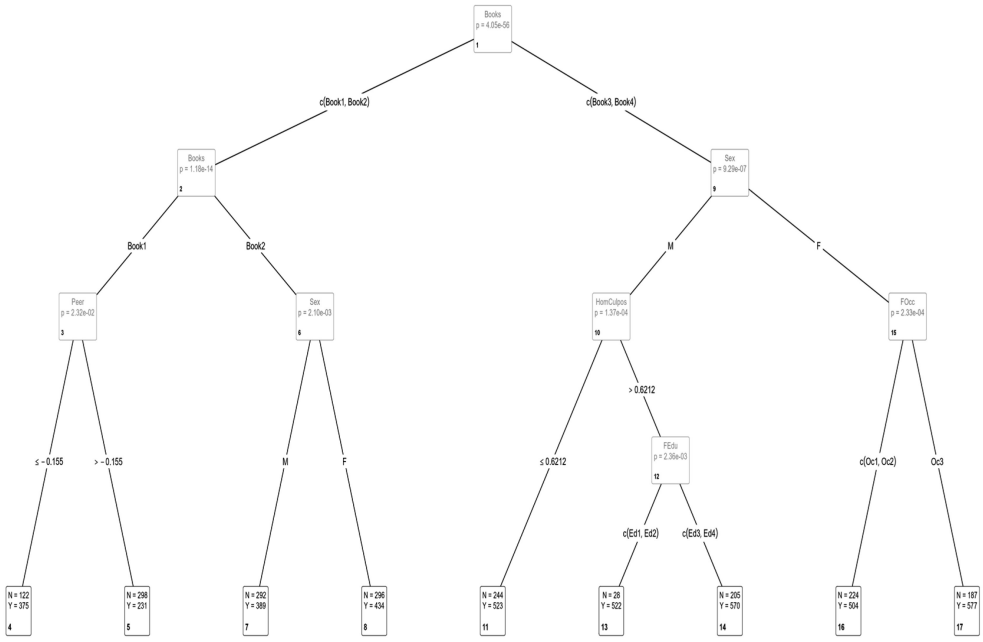
1. Andalucía



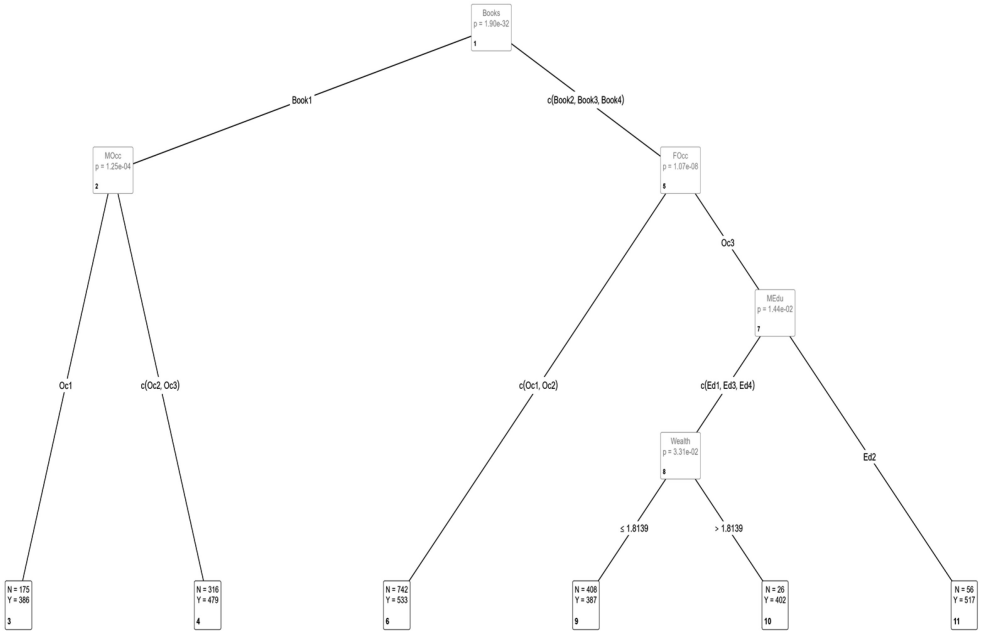
2. Aragón



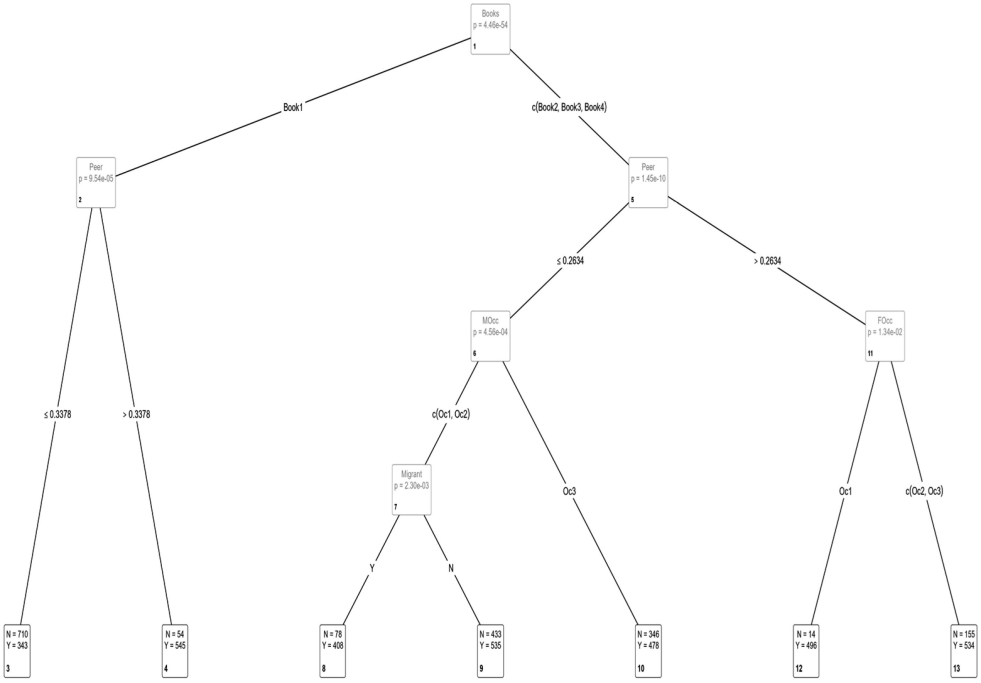
3. Asturias



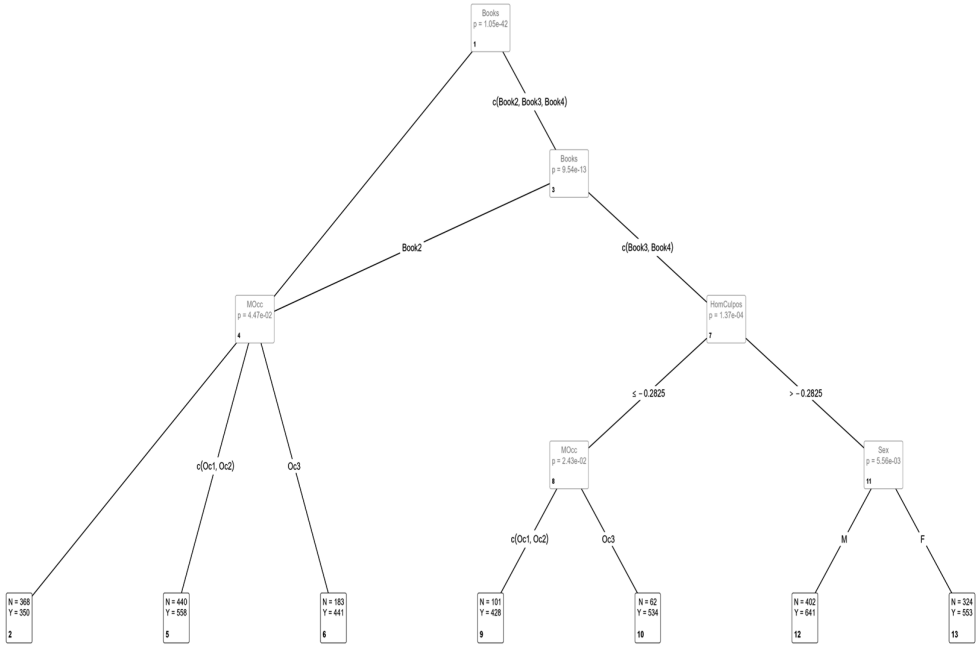
4. Baleares



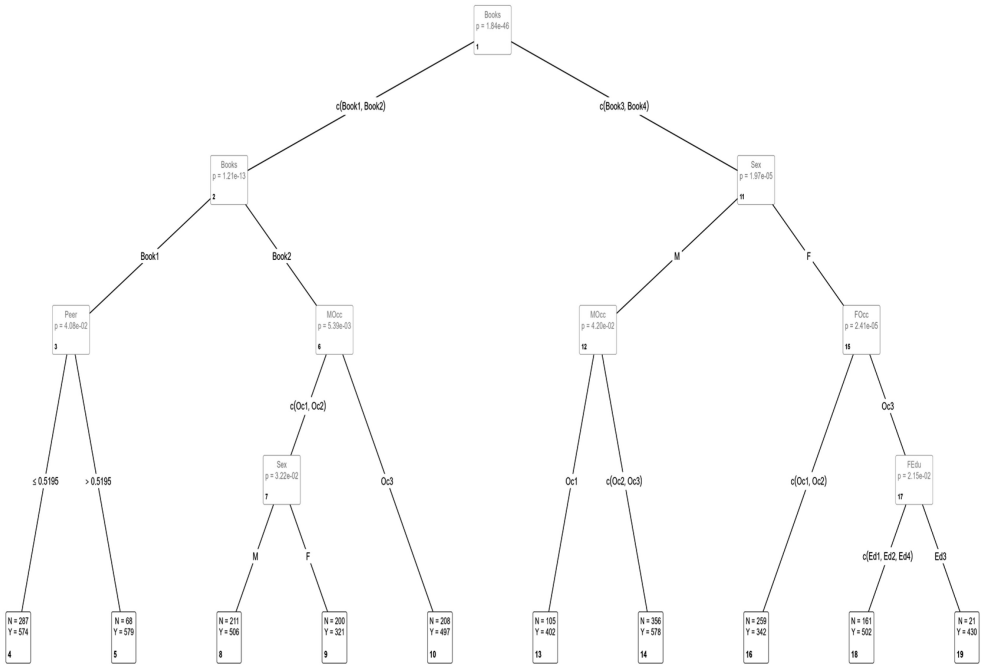
5. Canarias



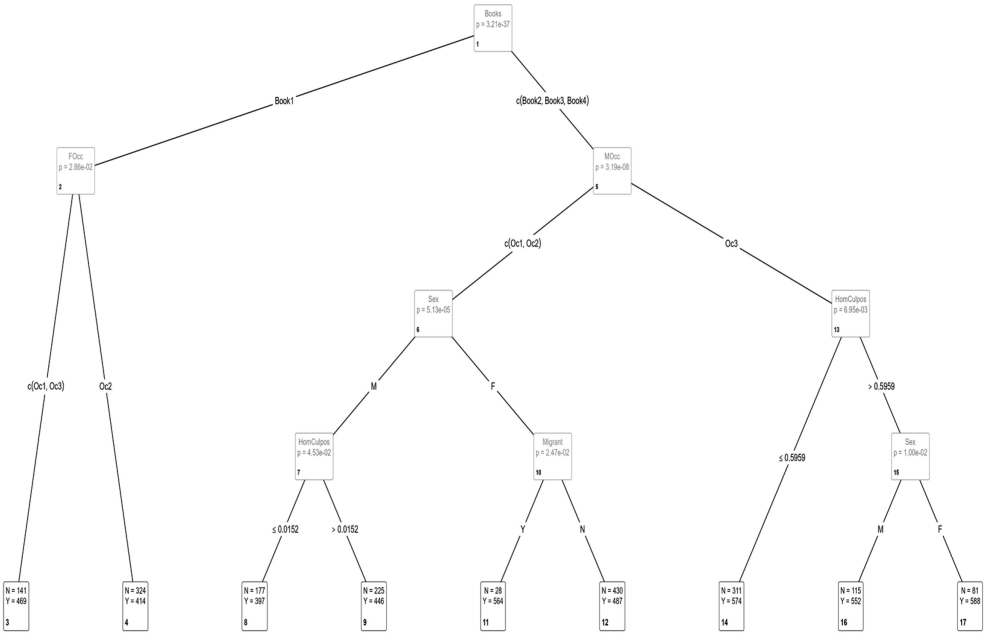
6. Cantabria



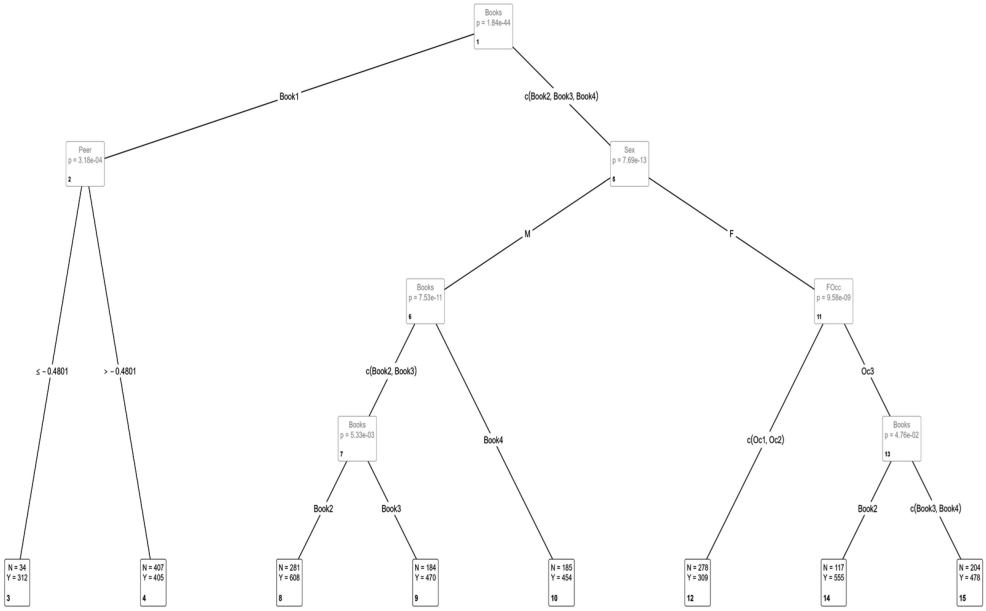
7. Castilla y León



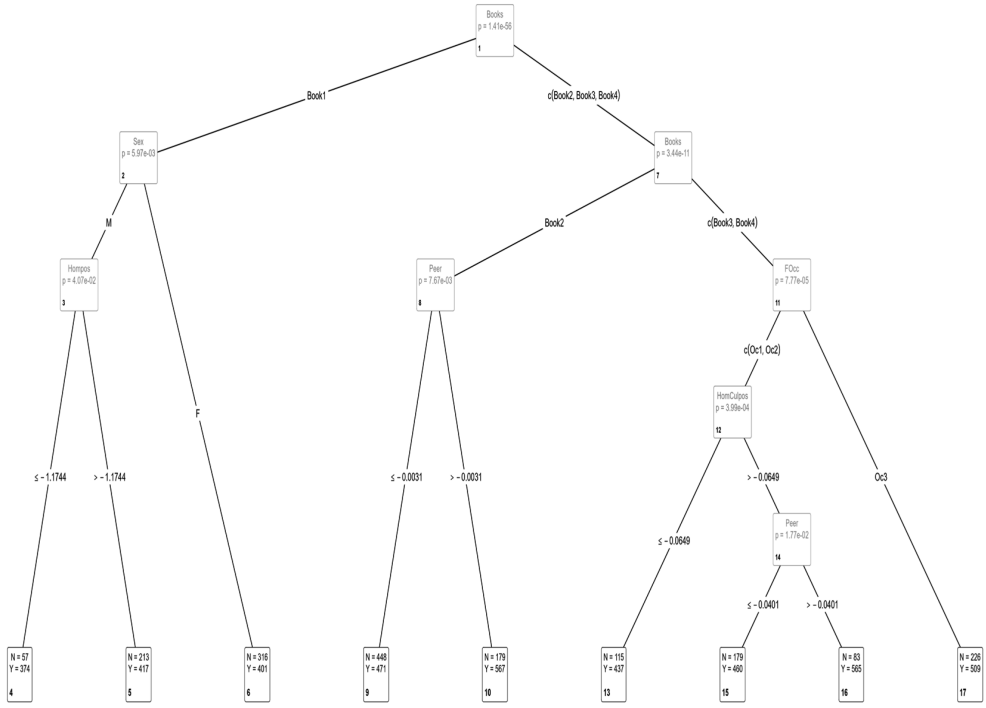
8. Castilla-La Mancha



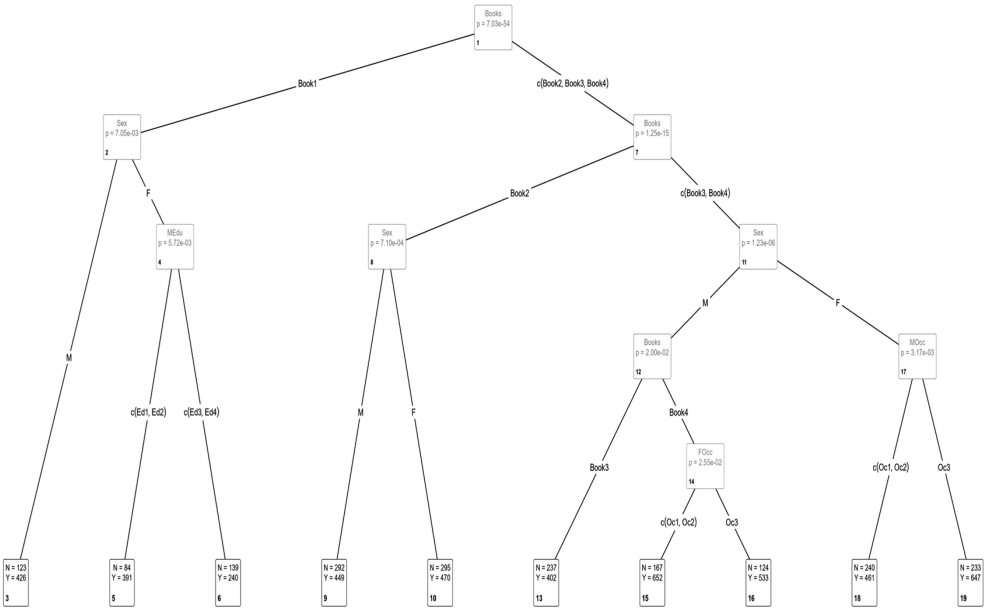
9. Cataluña



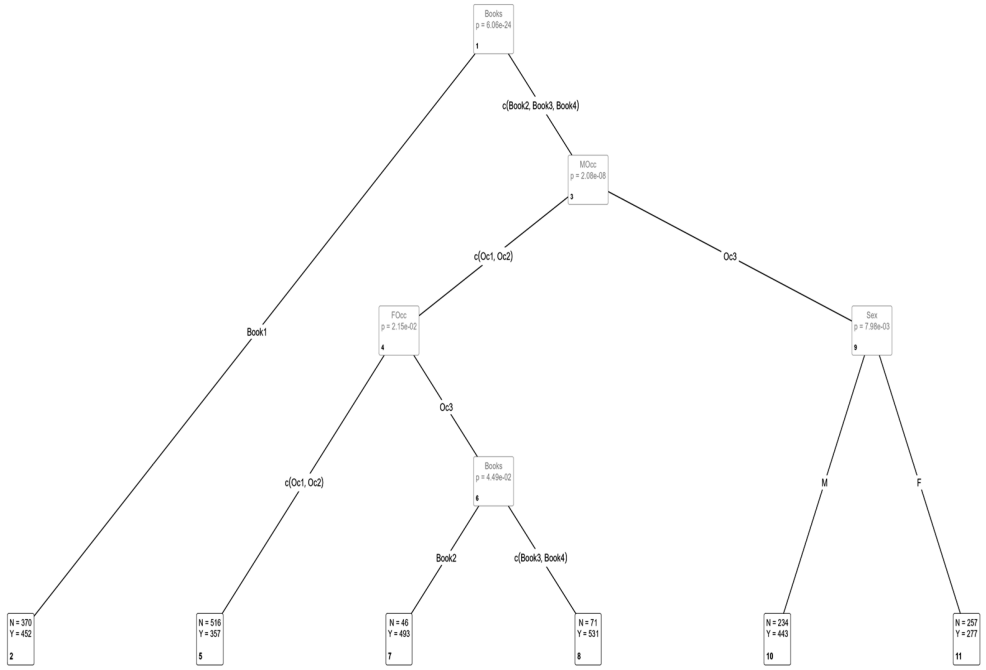
10. Extremadura



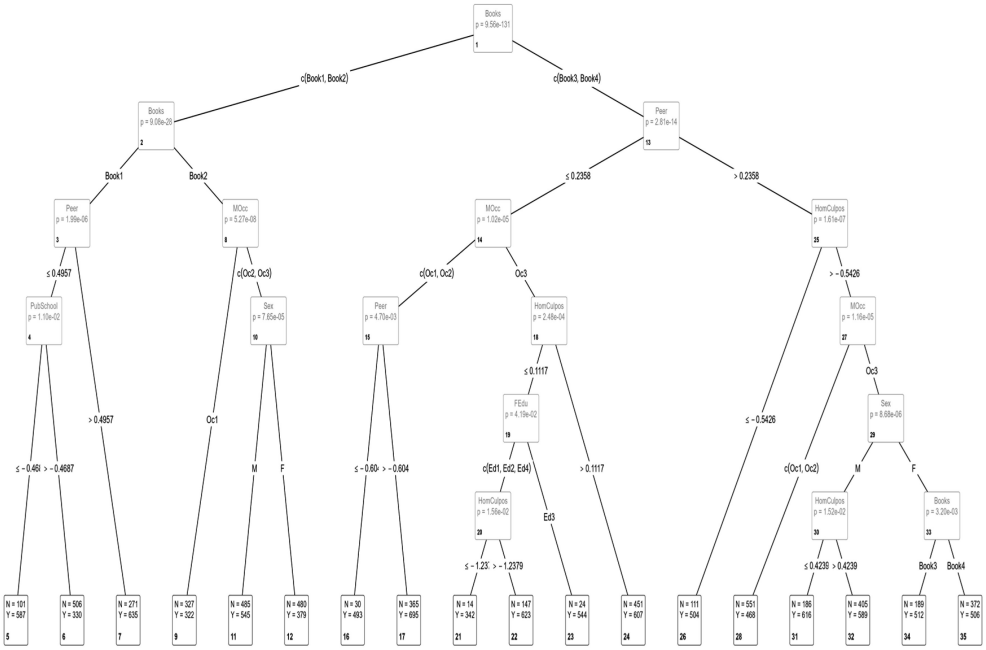
11. Galicia



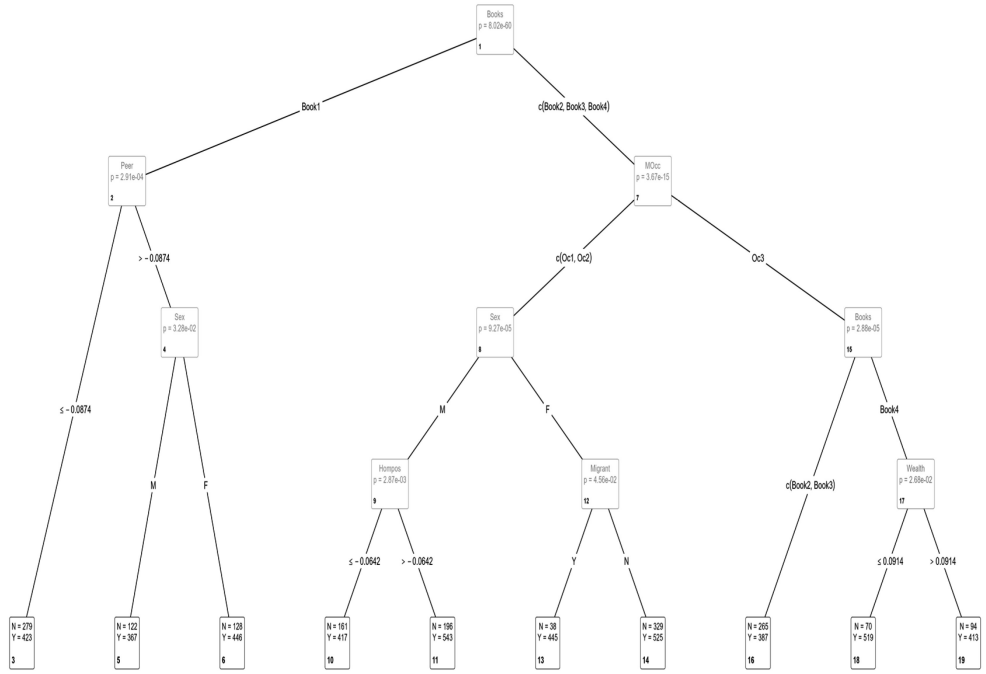
12. La Rioja



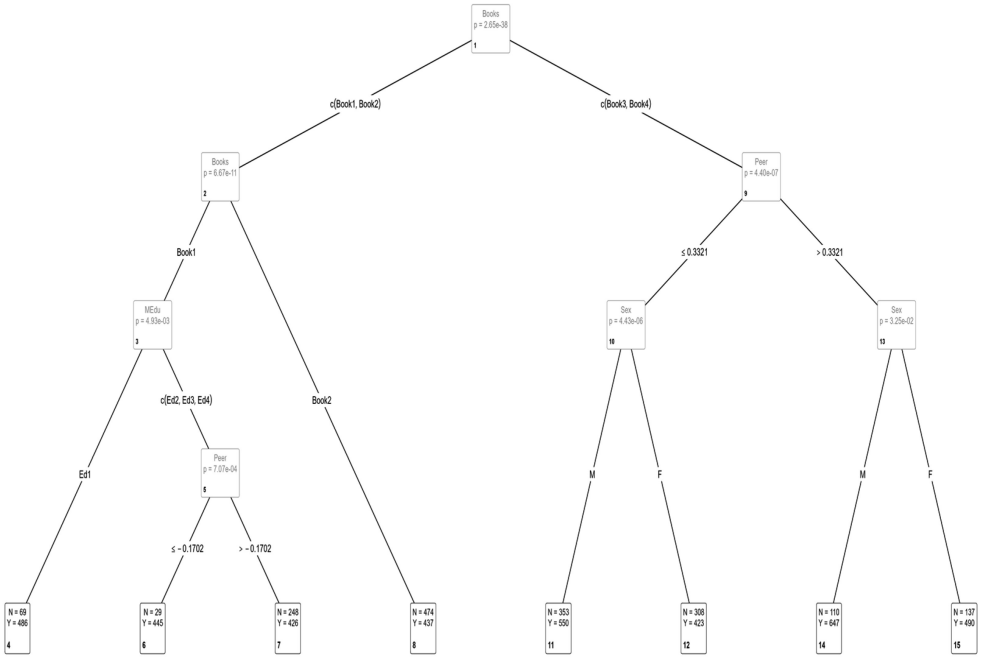
13. Madrid



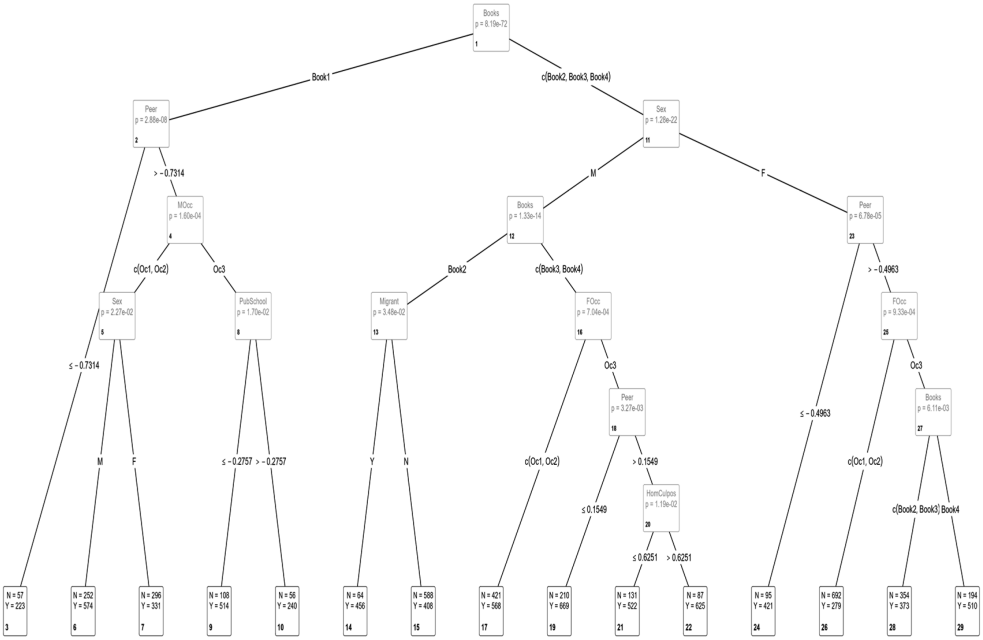
14. Murcia



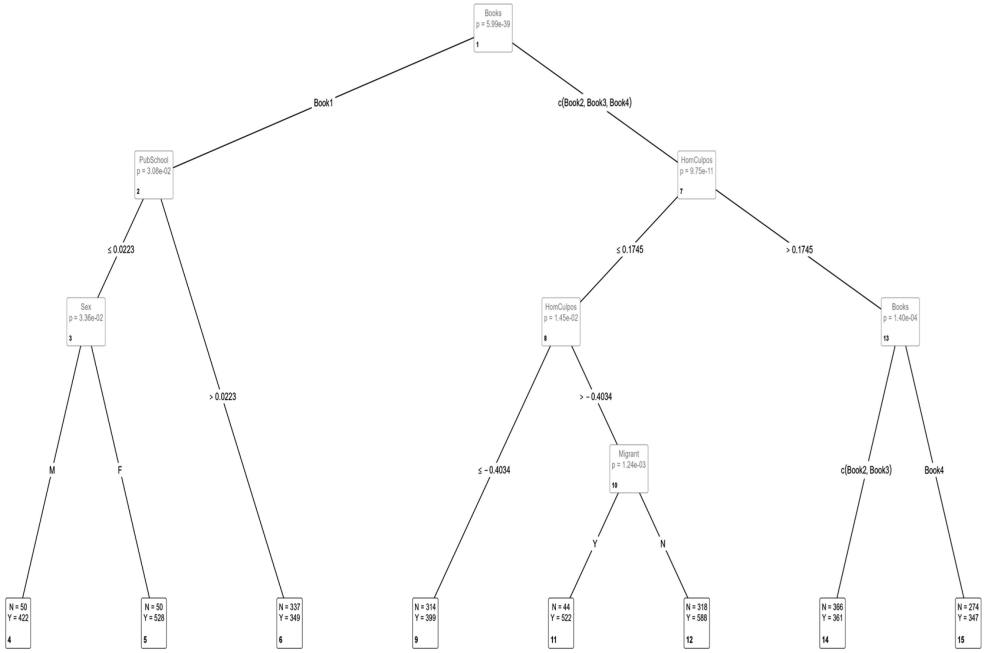
15. Navarra



16. País Vasco



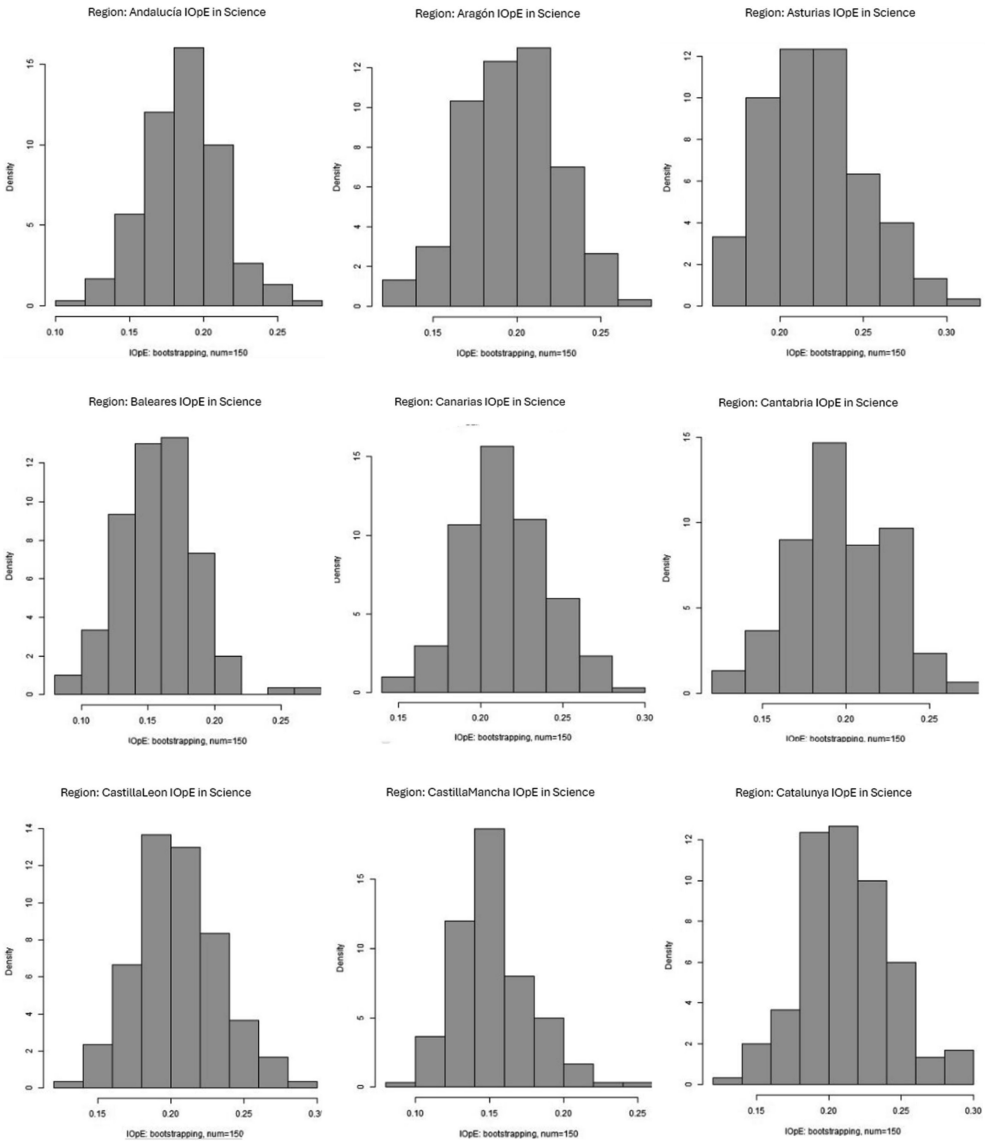
17. Valencia

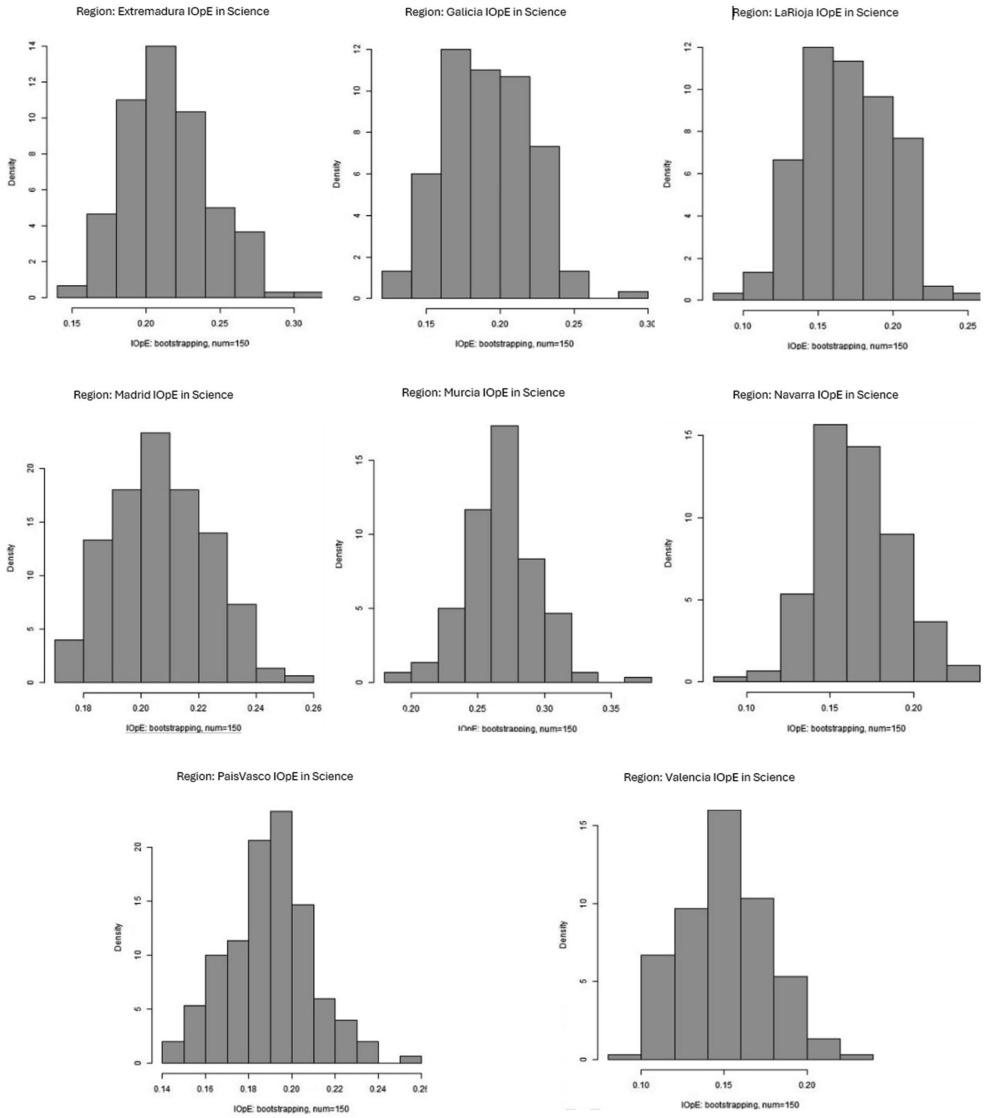

Books
p = 5.99e-39
1
Book1
c(Book2, Book3, Book4)
PubSchool
p = 3.08e-02
2
HomCulpos
p = 9.75e-11
7
≤ 0.0223
> 0.0223
≤ 0.1745
> 0.1745
Sex
p = 3.36e-02
3
HomCulpos
p = 1.45e-02
8
Books
p = 1.40e-04
13
M
F
≤ −0.4034
> −0.4034
c(Book2, Book3)
Book4
Migrant
p = 1.24e-03
10
Y
N
N = 50
Y = 422
4
N = 50
Y = 528
5
N = 337
Y = 349
6
N = 314
Y = 399
9
N = 44
Y = 522
11
N = 318
Y = 588
12
N = 366
Y = 361
14
N = 274
Y = 347
15

Appendix 7. Bootstrapped IOE distribution for Spanish regions

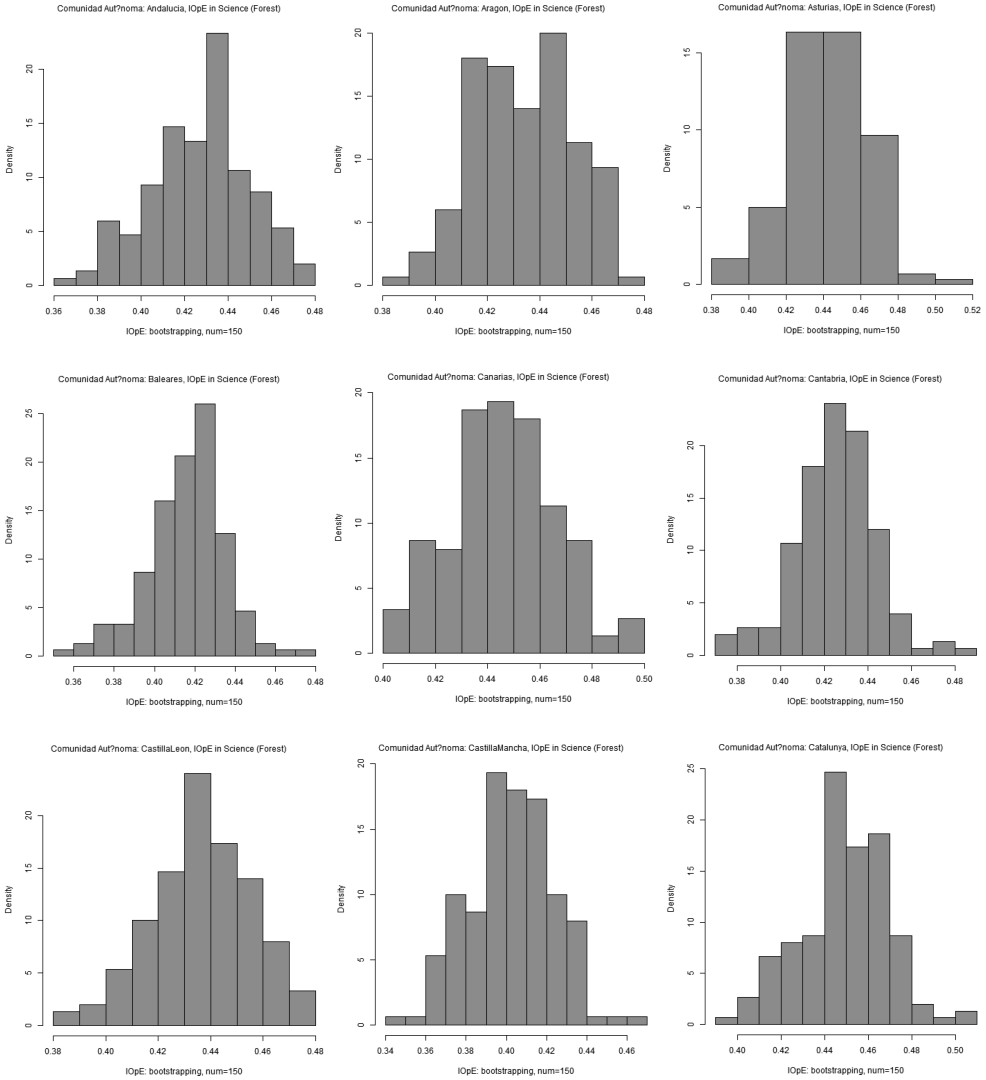
In this Appendix we show the IOpE distribution using bootstrapping for the CIT (A5.1) and CIF (A5.2) models and under the parameter-tuning scenario. The confidence intervals for our IOpE estimates are obtained from these distributions, and the average values are our point estimates.

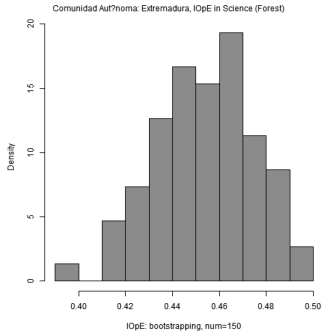
A7.1. Conditional Inference Trees under parameter-tuning scenario

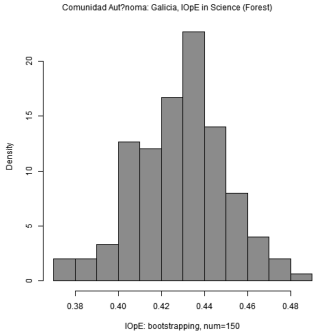


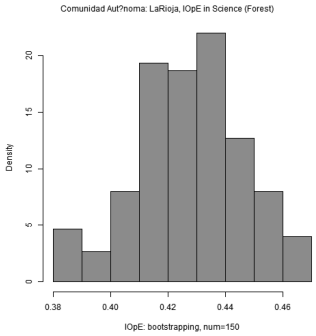

Region: Extremadura IOpE in Science
Region: Galicia IOpE in Science
Region: LaRioja IOpE in Science
Region: Madrid IOpE in Science
Region: Murcia IOpE in Science
Region: Navarra IOpE in Science
Region: PaisVasco IOpE in Science
Region: Valencia IOpE in Science
Density
IOpE: bootstrapping, num=150

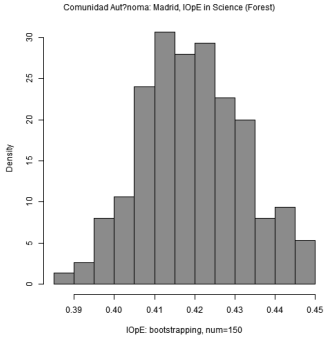
A7.2. Conditional Inference Forest under parameter-tuning scenario

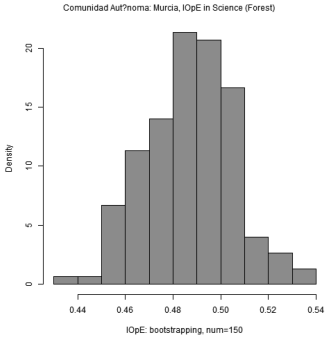


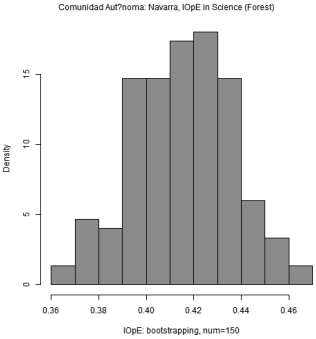

Comunidad Aut?noma: Extremadura, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150

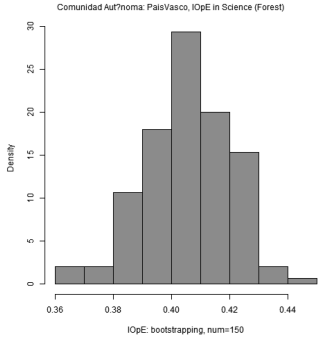

Comunidad Aut?noma: Galicia, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150

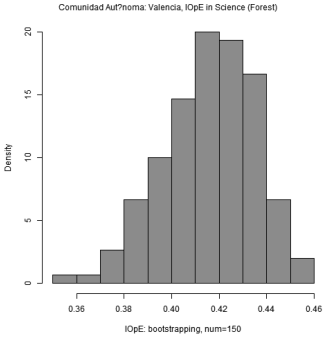

Comunidad Aut?noma: LaRioja, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150


Comunidad Aut?noma: Madrid, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150


Comunidad Aut?noma: Murcia, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150


Comunidad Aut?noma: Navarra, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150


Comunidad Aut?noma: PaisVasco, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150


Comunidad Aut?noma: Valencia, IOpE in Science (Forest)
Density
IOpE: bootstrapping, num=150

Notes

1. There exists a recent report by the COTEC Foundation (Martínez de Lafuente and Vega-Bayo, 2023), which uses ML techniques and estimate educational mobility and inequality of opportunity in educational attainment in European countries and in Spanish regions. They consider a battery of standard regularization techniques (Ridge, Lasso, Elastic Nets) and ML approaches (regression trees, random forest or Boosting). Our preferred approach is the Conditional Inference Forest, which is the one proposed by Brunori *et al*. (2023).

2. This measure depends on the MSE that is lost in each case, a unit that is difficult to interpret. It is usual to normalize each loss by dividing them by the sum of all losses. Thus, the result is a group of weights ranging from 0 to 100 (if we multiply them by 100) and the sum must give 100. In short, each weight measures the relative importance of each circumstance (in predicting accuracy) with respect to the others.

3. Following the recommendations of OECD (2019) for each region and competence, the parameter of interest (mean, standard deviation, etc.) was estimated separately for each of the ten plausible values and then averaged.

4. This index is constructed from four variables: parents' highest level of education; parents' highest level of employment; the number of books in the household; and the household wealth index defined in the main text. The ESCS has a mean of zero and a standard deviation of one for the average OECD student.

5. Since increasing α reduces the downward bias in the IOpE estimates while potentially increasing the upward bias, the IOpE is expected to increase with the level of α.

6. This comparison suggests that in the CIT (and in traditional methods) we leave out relevant information from less important (but potentially predictive) circumstances, whereas the CIF exploits this additional set of information.

7. Appendix 5 shows the IOpE empirical distribution for the *parameter-tuning scenario* for CIT and CIF, which also illustrate these results: the much higher average levels and the much smaller confidence intervals for the CIF than for the CIT.

8. Note that the differences obtained under the CIF tuning scenario are even smaller than under the CIT and in Sicilia *et al*. (2022). This is probably due to the fact that the CIF approach reduces the uncertainty in the estimates, which translates into smaller differences in IOpE between regions.

9. The Bonferroni adjustment is applied to the existing p-values: $1 - (1 - p_j)^J$. This adjustment is a statistical technique used to control for the increase in Type I errors (rejection of the null hypothesis very often) that occurs when conducting multiple comparisons (as in CIT and CIF). We also require that the sample size at the final nodes is at least 1% of the sample size of the database. This makes the results more comparable between Spanish regions, which may have different sample sizes (Table 1). In addition, since there are missing values in some of the explanatory variables in our sample, we ask for three surrogate splits to be set-up (i. e., the parameter "maxsurrogate" is equal to 3). The results are robust to the choice of 2 and 4 in the "maxsurrogate" parameter.

10. This division is obvious and no cut-off point is set if the variable is dichotomous.

11. We do not consider the ten plausible values to tune the parameters in CIT and CIF. We practically obtain the same tuning parameters for each set of plausible values.

References

Abdullah, A., Doucouliagos, H. and Manning, E. (2015), "Does education reduce income inequality? A meta-regression analysis", *Journal of Economic Surveys*, 29(2): 301-316.

Biau, G. and Scornet, E. (2016), "A random forest guided tour", *Test*, 25: 197-227.

Bourguignon, F., Ferreira, F. and Menéndez, M. (2007), "Inequality of opportunity in Brazil", *Review of Income and Wealth*, 53: 585-618.

Breiman, L. (1996), "Bagging predictors", *Machine Learning*, 24: 123-140.

Breiman, L (2001), "Random Forests", *Machine Learning*, 45(1): 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, New York: Chapman and Hall.

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

Brunori, P., Hufe, P. and Mahler, D. (2021), "The roots of inequality: Estimating inequality of opportunity from regression trees and forests", *IZA Discussion Paper*, 14689.

Brunori, P. and Neidhöfer, G. (2021), "The evolution of inequality of opportunity in Germany: A machine learning approach", *Review of Income and Wealth*, 67(4): 900-927.

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

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfield, F. D. and York, R. L. (1966), *Equality of Educational Opportunity. Washington*, DC: U. S., Government Printing Office.

Dills, A. K. (2005), "Does cream-skimming curdle the milk? A study of peer effects", *Economics of Education Review*, 24(1): 19-28.

Ferreira, F. and Gignoux, J. (2014), "The Measurement of Educational Inequality: Achievement and Opportunity", *World Bank Economic Review*, 28(2): 210-246.

Friedman, J., Tibshirani, R. and Hastie, T. (2009), *The elements of statistical learning: Data mining, inference, and prediction*, p. 359, New York: Springer-Verlag New York.

Hanushek, E. A., Kinne, L., Lergetporer, P. and Woessmann, L. (2020), "Culture and Student Achievement: the Inter twined Roles of Patience and Risk-Taking", *NBER Working Papers,* w27484, National Bureau of Economic Research.

Hapfelmeier, A., Hothorn, T., Ulm, K. and Strobl, C. (2014), "A new variable importance measure for random forests with missing data", *Statistics and Computing*, 24, 21-34.

Hofmarcher, T. (2021), "The effect of education on poverty: A European perspective", *Economics of Education Review*, 83: 102124.

Hothorn, T., Lausen, B., Benner, A. and Radespiel-Troeger, M. (2004), "Bagging Survival Trees", *Statistics in Medicine*, 23(1): 77-91.

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

Jerrim, J., Volante, L., Klinger, D. A. and Schnepf, S. V. (2019), *Socioeconomic inequality and student outcomes across education systems. In Socioeconomic Inequality and Student Outcomes*, pp. 3-16, Springer, Singapore.

Jusot, F., Tubeuf, S. and Trannoy, A. (2013), "Circumstances and efforts: how important is their correlation for the measurement of inequality of opportunity in health?", *Health economics*, 22(12): 1470-1495.

Lasso de la Vega, C., Lekuona, A. and Orbe, S. (2020), "Reexamining the inequality of opportunity in education in some European countries", *Applied Economics Letters*, 27(7): 544-548.

Marconi, G. (2018), "Education as a Long-Term Investment: The Decisive Role of Age in the Education-Growth Relationship", *Kyklos*, 71(1): 132-161.

Marrero, G.A. and Rodríguez, J.G. (2012), "Inequality of Opportunity in Europe", *Review of Income and Wealth*, 58(4): 597-621.

Marrero, G.A., Palomino, J.C. and Sicilia, G. (2024), "Educational Attainment in Western Europe: inequality, determinants and channels of transmission", *Journal of Economic Inequality*, 22 (2): 383-41.

Martínez de Lafuente, D. and Vega-Bayo, A. (2023), "Meritocracy and education: Social mobility and inequality of opportunity, a machine learning approach", *Informe COTEC*.

Martins, L. and Veiga, P. (2010), "Do inequalities in parents' education play an important role in PISA students' mathematics achievement test score disparities?", *Economics of Education Review*, 29(6), 1016-1033.

MEFP (2019), *PISA 2018. Programa para la Evaluación Internacional de los Alumnos. Informe Español*, Instituto de Evaluación Educativa, Madrid.

Meinshausen, N. (2006), "Quantile Regression Forests", *Journal of Machine Learning Research*, 7: 983-999.

Morgan, J.N. and Sonquist, J.A. (1963), "Problems in the analysis of survey data, and a proposal", *Journal of the American statistical association*, 58(302): 415-434.

OECD (2019), *PISA 2018 technical report*, OECD publishing.

Plassot, T., Soloaga, I. and Torres, P. (2022), "Inequality of Opportunity in Mexico and its Regions: A Data-Driven Approach", *The Journal of Development Studies*, 58(9): 1857-1873.

Rasch, G. (1960/1980), *Probabilistic models for some intelligence and attainment tests*, Danish Institute for Educational Research, Expanded edition (1980), The University of Chicago Press, Copenhague.

Sicilia, G., Marrero, G. and Palomino, J. (2022), *Desigualdad de oportunidades en el rendimiento educativo en España y Europa*, Fundació La Caixa.

Sicilia, G. and Simancas Rodríguez, R.F. (2023), "Eficiencia y equidad educativa en España: un análisis comparativo a nivel regional", *Hacienda Pública Española/Review of Public Economics*, 245(2): 7-33.

Sirin, S.R. (2005), "Socioeconomic status and academic achievement: A meta-analytic review of research", *Review of Educational Research*, 75(3): 417-453.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. (2007), "Bias in random forest variable importance measures: illustrations, sources and a solution", *BMC Bioinformatics*.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., and Zeileis, A. (2008), "Conditional variable importance for random forests", *BMC bioinformatics*, 9: 1-11.

Schütz, G., Ursprung, H.W. and Wößmann, L. (2008), "Education policy and equality of opportunity", *Kyklos*, 61(2): 279-308.

Yakovlev, P. and Leguizamon, S. (2012), "Ignorance is not bliss: On the role of education in subjective well-being", *The Journal of Socio-Economics*, 41(6): 806-815.

Resumen

La desigualdad de oportunidades en el rendimiento (DOE) mide la importancia de factores fuera del control del estudiante para explicar las diferencias en el rendimiento académico. Utilizando PISA 2018, estimamos la IOpE para las regiones españolas utilizando árbol de inferencia condicional (CIT)

y bosque (CIF). Utilizando CIF, los DOE duplican a los obtenidos utilizando enfoques tradicionales (en promedio, 43% frente a 20%). Murcia y Extremadura se encuentran entre las regiones con mayor IOpE, mientras que Castilla-La Mancha y el País Vasco tienen el IOpE más bajo. Las circunstancias que más contribuyen al IOpE son el entorno cultural en el hogar (número de libros) y la ocupación de los padres.

Palabras clave: desigualdad de oportunidades, rendimiento académico, aprendizaje automático, PISA, regiones españolas.

Clasificación JEL: D63, I24, I28, O52.