Aggregation Trees*

Riccardo Di Francesco†

March 27, 2023

Click here for the most recent version.

Abstract

Uncovering the heterogeneous effects of particular policies or "treatments" is a key concern for researchers and policymakers. A common approach is to report average treatment effects in different subgroups based on observable covariates. However, there is likely to be considerable uncertainty about the appropriate grouping. This paper proposes a nonparametric approach to discovering heterogeneous subgroups in a selection-on-observables framework. The approach constructs a sequence of groupings, one for each level of granularity. Groupings are nested and feature an optimality property. An "honesty" condition allows us to construct valid confidence intervals for the average treatment effect of each group. The use of the proposed methodology is illustrated through an empirical exercise that revisits the effects of maternal smoking on birth weight.

Keywords: Causality, conditional average treatment effects, recursive partitioning, subgroup discovery, subgroup analysis.
JEL Codes: C29, C45, C55

*I especially would like to thank Franco Peracchi for feedback and suggestions. I am also grateful to Hannah Busshoff, Matteo Iacopini, Michael Knaus, Michael Lechner, Tommaso Proietti, seminar participants at University of Rome Tor Vergata and SEW-HSG research seminars, and conference participants at the WEEE 2022 and the 1st Rome Ph.D. in Economics and Finance Conference for comments and discussions. Matias Cattaneo generously provided the data used in the empirical illustration of this paper. The R package for implementing the methodology developed in this paper is available on CRAN at https://CRAN.R-project.org/package=aggTrees. The associated vignette is at https://riccardo-df.github.io/aggTrees/.

†Department of Economics and Finance, University of Rome Tor Vergata, Rome. Electronic correspondence: riccardo.di.francesco@uniroma2.it.

1 Introduction

Uncovering the effects of particular policies or "treatments" is a key concern for researchers and policymakers. Under a selection-on-observables assumption, we can identify and estimate these effects at several levels of granularity, ranging from the coarsest Average Treatment Effect (ATE) to the finest Conditional Average Treatment Effects (CATEs). The ATE characterizes the average impact of the policy, thus measuring its actual effectiveness. On the other hand, the CATEs shed light on the distributional impacts of the policy, which are crucial for decision making if the social welfare criterion that represents the preferences of the policy maker is not "utilitarian" (e.g., Kitagawa & Tetenov, 2021).

However, the CATEs are typically hard to interpret and digest and their estimation may lack precision. Therefore, researchers generally prefer to address treatment effect heterogeneity by reporting the average effect on different subgroups formed by observable covariates such as gender or education. These Group Average Treatment Effects (GATEs) are more interpretable than the CATEs and thus more useful for policy makers. Chernozhukov, Demirer, Duflo, and Fernández-Val (2017) document that, among 189 RCTs published in top economic journals since 2006, 40% report at least one subgroup analysis.

One issue in investigating group heterogeneity is that there could be a large number of ways to form subgroups. Thus, there is likely to be considerable uncertainty about the appropriate grouping. In particular, concerns may arise about which covariates are best suited to explain group heterogeneity and the number of subgroups that should be formed.

Researchers often specify several groups and report only subgroups for which significant heterogeneity is found. However, this strategy, known as p-hacking, results in invalidated inference (e.g., Imbens, 2021). To avoid this risk, journals often require the adoption of a pre-analysis plan that specifies ex ante which subgroups will be investigated. However, this limits the possibility of uncovering unexpected heterogeneity.

This paper proposes an alternative methodology to discover heterogeneous subgroups. The methodology allows researchers to assess whether there is relevant heterogeneity while

avoiding the risk of p-hacking and the drawbacks of pre-analysis plans. This is achieved by estimating the CATEs and aggregating them into a decision tree (Breiman, Friedman, Olshen, & Stone, 1984). This way, we generate a sequence of groupings, one for each level of granularity. The resulting sequence is nested in the sense that subgroups formed at a given level of granularity are never broken at coarser levels. This guarantees consistency of the results across the different granularity levels, which is considered a basic requirement that every classification system should satisfy (e.g., Cotterman & Peracchi, 1992). Moreover, we show that each grouping features an optimality property in that it ensures that the loss in explained heterogeneity resulting from aggregation is minimized.

For a particular grouping, point estimates and standard error for the GATEs are obtained by fitting an appropriate linear model. Under an "honesty" condition (Athey & Imbens, 2016), we can use the estimated standard errors to conduct valid inference about the GATEs as usual, e.g., by constructing conventional confidence intervals and testing particular hypotheses. Honesty is a subsample-splitting technique that requires that different observations are used to form subgroups and estimate the GATEs. In analogy to classical econometrics, this is equivalent to using different subsamples to select and estimate a model. This way, the asymptotic properties of GATE estimates are the same as if the groupings had been exogenously given.

The rest of the paper unfolds as follows. Section 2 discusses the estimands of interest and their identification. Section 3 presents aggregation trees, explaining in detail their implementation. Section 4 compares aggregation and causal trees, juxtaposing the methodological differences and providing some simulation results. Section 5 illustrates an empirical exercise in which the effects of maternal smoking on birth weight are revisited. Section 6 concludes. A description of the variables used in the empirical exercise and proof of formal results are provided in the Appendix.

2 Causal Framework

2.1 Estimands of Interest

We define the estimands of interest using the potential outcomes model (Neyman, 1923; Rubin, 1974). Suppose to have access to a sample of n i.i.d. observations (Y_i, D_i, X_i), where $Y_i \in \mathcal{Y}$ is the outcome targeted by the treatment, $D_i \in \{0, 1\}$ is the binary treatment indicator, and $X_i \in \mathcal{X} \subset \mathbb{R}^p$ is the pre-treatment covariate vector of the i-th unit. We postulate the existence of two potential outcomes $Y_i(1)$ and $Y_i(0)$, denoting the outcome that the i-th unit would experience under each treatment level.

To define the effect of the treatment, we can take the differences in the potential outcomes of each unit and aggregate them at different levels of granularity. The coarsest estimand of interest is the Average Treatment Effect (ATE), $\tau := \mathbb{E}[Y_i(1) - Y_i(0)]$. The ATE characterizes the average impact of the policy, thus measuring its actual effectiveness. However, it does not allow researchers to investigate the distributional aspects of the policy. To this end, we can focus the analysis on the Conditional Average Treatment Effects (CATEs), $\tau(X_i) := \mathbb{E}[Y_i(1) - Y_i(0) | X_i]$. The CATEs provide information at the finest level of granularity that can be achieved with the variables at hand. They allow researchers to understand what drives the effects of a particular policy by relating treatment effect heterogeneity to the observable covariates. However, they are difficult to interpret and communicate to policy makers.

Group Average Treatment Effects (GATEs) represent a way to uncover treatment effect heterogeneity in a digestible and easy-to-communicate form. The GATEs are defined as $\tau_g := \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i \in \mathcal{X}_g\right]$, $g = 1, \ldots, G$, where the sets $\mathcal{X}_1, \ldots, \mathcal{X}_G$ form a partition of $\mathcal{X}$. The definition of GATEs requires choosing a value of G, which controls the level of granularity to deploy, and a partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$ of $\mathcal{X}$. If subgroups are formed according to the levels of a single discrete variable $Z_i \subset X_i$, each GATE simplifies to $\tau_g = \mathbb{E}[Y_i(1) - Y_i(0) | Z_i = g]$.

2.2 Identification

All the estimands discussed in the previous section are defined in terms of potential outcomes. However, each unit is either treated or not treated. Thus, we observe only one potential outcome per unit, and further assumptions are needed for identification.

To identify ATE, GATEs, and CATEs, we need the following standard assumptions (e.g., Imbens & Rubin, 2015):

Assumption 2.1. *(SUTVA):* $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

Assumption 2.2. *(Exogeneity of the covariates):* $X_i(1) = X_i(0) = X_i$*, where potential covariates are defined analogously to potential outcomes.*

Assumption 2.3. *(Unconfoundedness):* $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i | X_i$

Assumption 2.4. *(Common support):* $0 < p(X_i) < 1$*, where* $p(X_i) \equiv \mathbb{P}(D_i = 1 | X_i)$ *is the conditional treatment probability (or propensity score).*

SUTVA assumes the absence of interference between units, thus ruling out spillover effects. The exogeneity of the covariates stipulates that the covariates are not affected by treatment assignment, implying that controlling for X_i does not hinder causal estimation. The unconfoundedness assumption, also known as "selection-on-observables," requires that X_i contains all the "confounders," i.e., all covariates jointly affecting treatment assignment and the outcome.[1] The common support assumption states that each unit must have a non-zero probability of belonging to the treatment and control groups.

Under Assumptions 2.1–2.4, the CATEs are identified from observable data:

[1] X_i can also include additional "heterogeneity variables" not necessary for identification but for which treatment effect heterogeneity is of interest. The sets of confounders and heterogeneity variables can overlap in any way or be disjoint.

$$
\begin{aligned}
&\mathbb{E}\left[Y_i | X_i, D_i = 1\right] - \mathbb{E}\left[Y_i | X_i, D_i = 0\right] \\
&\quad = \mathbb{E}\left[Y_i\left(1\right) | X_i, D_i = 1\right] - \mathbb{E}\left[Y_i\left(0\right) | X_i, D_i = 0\right] \quad (\textit{by SUTVA}) \\
&\quad = \mathbb{E}\left[Y_i\left(1\right) | X_i\right] - \mathbb{E}\left[Y_i\left(0\right) | X_i\right] \quad (\textit{by Unconfoundedness}) \qquad (2.1) \\
&\quad = \mathbb{E}\left[Y_i\left(1\right) - Y_i\left(0\right) | X_i\right] \\
&\quad = \tau\left(X_i\right)
\end{aligned}
$$

where the quantities in the first line are easily estimated using any supervised learning method. The ATE and the GATEs can be rewritten as expectations of the CATEs, thus being identified under the same assumptions.

3 Aggregation Trees

GATEs estimation requires forming subgroups according to one or more observable covariates. However, there may be a large number of ways to form subgroups and there is likely to be considerable uncertainty about the appropriate grouping. In the absence of subject-matter knowledge, a data-driven methodology for discovering heterogeneous subgroups is particularly attractive.

The methodology proposed in this paper consists of three steps. First, an estimation step constructs an estimate $\hat{\tau}\left(\cdot\right)$ of $\tau\left(\cdot\right)$. Second, a tree-growing step approximates the estimated $\hat{\tau}\left(X_i\right)$ by a binary hierarchical tree to construct the set of admissible groupings, that is, groupings that maximize systematic between-group heterogeneity. Third, a tree-pruning step uses a cost-complexity criterion to generate a sequence of nested subtrees, one for each level of granularity. Each subtree provides an optimal grouping of the observations, where optimality means that, for each granularity level, groupings minimize the loss in explained heterogeneity resulting from aggregation.

The next two subsections discuss how the tree-growing and the tree-pruning steps may be interpreted at the population level, followed by some remarks on estimation and inference.

3.1 Tree-Growing Step

To define the GATEs we need to select a partition of the covariate space $\mathcal{X}$. Then, we can identify each GATE as the expectation of $\tau(\cdot)$ within each group.

This is equivalent to constructing a multivariate step-function $f \in \mathcal{F}$ that well approximates $\tau(\cdot)$ by partitioning $\mathcal{X}$ into strata and then treats $\tau(\cdot)$ as constant within each stratum. Under the mean squared error (MSE) criterion the best piecewise constant approximation to $\tau(\cdot)$ is the solution to the following problem:[2]

$$\min_{\{(c_g, \mathcal{X}_g)\}_{g=1}^{G}} \mathbb{E}\left[(\tau(X) - f(X))^2\right] \quad \text{s.t.} \quad f(X) = \sum_{g=1}^{G} c_g \mathbb{1}\left(X \in \mathcal{X}_g\right) \tag{3.1}$$

where $\mathbb{1}(\cdot)$ is an indicator of the truth of its argument, $c_1, \ldots, c_G$ are constants, and the sets $\mathcal{X}_1, \ldots, \mathcal{X}_G$ form a partition of the covariate space $\mathcal{X}$. We can show that, for any partition of $\mathcal{X}$, the optimal constants are $c_g^* = \mathbb{E}\left[\tau(X) \,|\, X \in \mathcal{X}_g\right]$. Thus, solving problem (3.1) identifies the GATEs regardless of how groups are formed (see Appendix C).

The question is then how to find the optimal partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$. Following the approach of Breiman et al. (1984), partitions of $\mathcal{X}$ can be constructed by recursively stratifying the covariate space using axis-aligned splits. Let x_j be some particular value of the j-th covariate. Starting with a region of the covariate space $\mathcal{R}_m \subseteq \mathcal{X}$, consider a candidate splitting variable j and splitting point s. Define the corresponding subregions as:[3]

$$\mathcal{R}_{m+1}(j, s) = \left\{X | x_j \leq s\right\}, \quad \mathcal{R}_{m+2}(j, s) = \left\{X | x_j > s\right\}$$

The split occurs on a given pair (j, s), and the population is stratified accordingly. The process is then repeated in the resulting subregions, thus obtaining finer and finer partitions of $\mathcal{X}$. The whole procedure can be described by the shape of a decision tree: the *"root"* (i.e., the node with no *"parent"*) corresponds to $\mathcal{X}$, the m-th internal node represents subregion

[2] We are implicitly assuming that $\tau(\cdot)$ is not a step-function.

[3] In the case of categorical splitting variables, s corresponds to a subset of possible levels of j, and the inequality signs are replaced by $\in$ and $\notin$. Continuous variables need to be discretized.

$\mathcal{R}_m$ and has two *"children"* nodes representing subregions $\mathcal{R}_{m+1}$ and $\mathcal{R}_{m+2}$, and associated with the *"leaves"* (i.e., the collection of terminal nodes) is a partition of the covariate space.

Ideally, we would like to explore the space of all possible trees and pick the one whose associated partition minimizes (3.1). However, it is generally infeasible to enumerate all the distinct binary decision trees. Consider the situation where the random vector X is composed of p binary covariates, and let $\mathcal{D}$ be the *"depth"* of a given tree (that is, the number of nodes connecting the root to the furthest leaf). We can show that $L_{\mathcal{D}} = \prod_{d=1}^{\mathcal{D}} (p - (d-1))^{2^{d-1}}$ is a lower bound for the number of distinct binary decision trees grown by recursively partitioning $\mathcal{X}$ and having a depth equal to or lower than $\mathcal{D}$ (see Appendix D). $L_{\mathcal{D}}$ quickly diverges as p grows. For example, fixing $\mathcal{D} = 3$ and letting $p = 10$ yields a lower bound of 3,317,760, while letting $p = 20$ leads to a bound of 757,926,720. Things only worsen with categorical covariates taking more than two values or continuous covariates discretized using a large number of bins.

To cope with this issue, Breiman et al. (1984) suggest a "greedy" approach that partitions each region $\mathcal{R}_m \subseteq \mathcal{X}$ by choosing the split that minimizes the MSE within the resulting subregions $\mathcal{R}_{m+1}$ and $\mathcal{R}_{m+2}$. This process is then iterated until some particular *"stopping criterion"* is met, for instance the maximum depth $\mathcal{D}_{max}$ of the tree. This approach is greedy in that it ignores that a suboptimal split could yield better results at later steps and is generally considered to be a reasonable way of circumventing the exhaustive search of the space of all possible trees.

At each step, the optimal greedy split is placed to explain as much heterogeneity as possible within the two resulting subregions. In particular, the splitting variable j and the splitting point s are chosen to solve the following problem:

$$\min_{j,s} \mathbb{V}\left(\tau\left(X\right) | X \in \mathcal{R}_{m+1}\left(j,s\right)\right) + \mathbb{V}\left(\tau\left(X\right) | X \in \mathcal{R}_{m+2}\left(j,s\right)\right) \tag{3.2}$$

The greedy approach partitions each subregion $\mathcal{R}_m \subseteq \mathcal{X}$ in a way that maximizes systematic heterogeneity between the resulting subgroups, thus constructing a set of admissible group-

ings described by a tree $\mathcal{T}_0$. Notice that the first greedy split yields the optimal non-greedy partition (associated with the leaves of a 1-depth tree), while subsequent splits are likely to be only greedy-optimal.

3.2 Tree-Pruning Step

Choosing the size of the tree is important. On the one hand, too deep trees might capture unimportant details of $\tau(\cdot)$. On the other hand, too shallow trees might miss important structure. This trade-off between the tree size and the accuracy of the approximation can be formalized by the following *"cost-complexity"* criterion:

$$C_\alpha(\mathcal{T}) = \sum_{\ell=1}^{|\mathcal{T}|} Q_\ell(\mathcal{T}) + \alpha|\mathcal{T}| \tag{3.3}$$

with $\ell = 1, \ldots, |\mathcal{T}|$ an index for the terminal nodes of a tree $\mathcal{T}$, and $Q_\ell(\mathcal{T}) = \mathbb{V}(\tau(X)|X \in \mathcal{R}_\ell)$. The first term on the right-hand side of (3.3) measures the variability of $\tau(\cdot)$ in region $\mathcal{R}_\ell$ around its piecewise constant approximation given by $\mathbb{E}[\tau(X)|X \in \mathcal{R}_\ell]$. This quantity can be arbitrarily reduced by growing deeper trees. This justifies the regularization term $\alpha|\mathcal{T}|$ that penalizes the model according to the cost-complexity parameter $\alpha \in [0, \infty)$ and the number of terminal nodes $|\mathcal{T}|$ of $\mathcal{T}$.

The parameter α governs the balance between the accuracy and the interpretability of the model. Define a subtree $\mathcal{T} \subset \mathcal{T}_0$ as any tree that can be obtained by collapsing any number of internal nodes of $\mathcal{T}_0$ and let $\mathcal{T}_\alpha \subseteq \mathcal{T}_0$ be the smallest subtree for which (3.3) is minimized. For each α, a unique $\mathcal{T}_\alpha$ exists, which can be identified by *"weakest link pruning:"* starting from $\mathcal{T}_0$, we iteratively collapse the internal node that gives the slightest increase in the accuracy of the approximation. This "weakest" node is defined in terms of the impurity $Q_\ell(\mathcal{T})$ of its children.

Following this procedure, we can generate a sequence of nested subtrees $\mathcal{T}_{\alpha_0}, \mathcal{T}_{\alpha_1} \ldots, \mathcal{T}_{\alpha_{max}}$, where $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_{max} < \infty$ are threshold values such that all α in a given interval lead to the same subtree and $\mathcal{T}_{\alpha_{max}}$ corresponds to the tree's root. Associated with each

subtree in the sequence is a partition of the covariate space. Therefore, the tree-pruning step generates a sequence of groupings, one for each threshold value $\alpha_0 < \alpha_1 < \cdots < \alpha_{max}$. Because the sequence is nested, subgroups formed at a given level of granularity are never broken at coarser levels. This guarantees consistency of the results across the different granularity levels, which is considered a basic requirement that every classification system should satisfy (e.g., Cotterman & Peracchi, 1992). Moreover, because each $\mathcal{T}_{\alpha_k}$ is obtained by collapsing the weakest node of $\mathcal{T}_{\alpha_{k-1}}$, the tree-pruning step constructs optimal groupings by aggregating the two subgroups for which the loss in explained heterogeneity resulting from aggregation is minimized.

3.3 Estimation and Inference

Using the sample analogs of (3.2)–(3.3) is infeasible, as $\tau(\cdot)$ is never observed. This paper proposes to use an estimate $\hat{\tau}(\cdot)$ of $\tau(\cdot)$ in the tree-growing and tree-pruning steps to construct the sequence of optimal groupings. Then, for a particular granularity level, we can estimate the GATEs in several ways. In randomized experiments, taking the difference between the mean outcomes of treated and control units in each group is an unbiased estimator of the GATEs. Equivalently, we can obtain the same point estimates in addition to their standard errors by estimating via OLS the following linear model:

$$Y_i = \sum_{l=1}^{|\mathcal{T}_\alpha|} L_{i,l}\,\gamma_l + \sum_{l=1}^{|\mathcal{T}_\alpha|} L_{i,l}\,D_i\,\beta_l + \epsilon_i \tag{3.4}$$

with $|\mathcal{T}_\alpha|$ the number of leaves of a particular tree $\mathcal{T}_\alpha$, and $L_{i,l}$ a dummy variable equal to one if the i-th unit falls in the l-th leaf of $\mathcal{T}_\alpha$. Exploiting the random assignment to treatment, we can show that each β_l identifies the GATE in the l-th leaf.

In observational studies, estimating model (3.4) would yield biased GATE estimates due to the selection into treatment. To get unbiased estimates, we can use the orthogonal estimator of Semenova and Chernozhukov (2021) to estimate the best linear predictor of $\tau(\cdot)$ given a set of dummies denoting leaf membership. The key idea is to construct a random variable

Γ_i, generally called score, such that $\tau(X_i) = \mathbb{E}[\Gamma_i | X_i]$, and project it onto $L_{i,1}, \ldots, L_{i,|\mathcal{T}_\alpha|}$. For instance, consider the following doubly-robust score (Robins & Rotnitzky, 1995):

$$\Gamma_i^* = \mu(1, X_i) - \mu(0, X_i) + \frac{D_i [Y_i - \mu(1, X_i)]}{p(X_i)} - \frac{(1 - D_i)[Y_i - \mu(0, X_i)]}{1 - p(X_i)}$$

where $\mu(D_i, X_i) = \mathbb{E}[Y_i | D_i, Z_i]$ is the conditional mean of Y_i and $p(X_i) = \mathbb{P}(D_i = 1 | X_i)$ is the propensity score. Because $\tau(X_i) = \mathbb{E}\left[\Gamma_i^* | X_i\right]$, this score is a natural candidate. We recognize that it depends on unknown functions $\eta(X_i) := \{\mu(1, X_i), \mu(0, X_i), p(X_i)\}$ and make this explicit by writing $\Gamma_i^* = \Gamma_i^*(\eta)$. We refer to η as nuisance functions, as they are not of direct interest but necessary to construct a plug-in estimate $\Gamma_i^*(\hat{\eta})$ of $\Gamma_i^*(\eta)$ that we aim to regress on $L_{i,1}, \ldots, L_{i,|\mathcal{T}_\alpha|}$. Semenova and Chernozhukov (2021) show that $\Gamma_i^*(\eta)$ is a Neyman-orthogonal score (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, & Robins, 2018), that is, its plug-in estimate $\Gamma_i^*(\hat{\eta})$ is insensitive to bias in the estimation of $\hat{\eta}$. They then suggest the following two-stage procedure. First, construct an estimate $\hat{\eta}$ of the nuisance functions η using K-fold cross-fitting: split the sample into K folds of similar sizes and, for each $k = 1, \ldots, K$, estimate $\hat{\eta}_k$ using all but the k-th folds. Second, construct $\widehat{\Gamma}_i^* := \Gamma_i^*(\hat{\eta}_k)$, where the observation i belongs to the k-th fold, and estimate via OLS the following linear model:

$$\widehat{\Gamma}_i^* = \sum_{l=1}^{|\mathcal{T}_\alpha|} L_{i,l} \beta_l + \epsilon_i \tag{3.5}$$

As before, each β_l identifies the GATE in the l-th leaf. Semenova and Chernozhukov (2021) show that thanks to the Neyman-orthogonality of Γ_i^*, the OLS estimator $\hat{\beta}_l$ of β_l is root-n consistent and asymptotically normal, provided that the product of the convergence rates of the estimators of the nuisance functions $\mu(\cdot, \cdot)$ and $p(\cdot)$ is faster than $n^{1/2}$. This allows using machine learning estimators such as random forests and LASSO to estimate the nuisance functions, as they are shown to achieve an $n^{1/4}$ convergence rate and faster under particular conditions.

However, GATE estimates may show some bias if we use the same data to construct

the tree and to estimate models (3.4)–(3.5), leading to invalid inference. One way out is to grow "honest" aggregation trees (Athey & Imbens, 2016). Honesty is a subsample-splitting technique that requires that different observations are used to form the subgroups and estimate the GATEs. For this purpose, we split the observed sample into a training sample $\mathcal{S}^{tr}$ and an honest sample $\mathcal{S}^{hon}$ of arbitrary sizes. We use $\mathcal{S}^{tr}$ to estimate $\tau(\cdot)$ and construct the tree $\mathcal{T}_0$ and, for a particular grouping $\mathcal{T}_\alpha$, we use $\mathcal{S}^{hon}$ to estimate (3.4)–(3.5). This way, the asymptotic properties of the estimators discussed above are the same as if the groupings had been exogenously given. Honesty generally comes at the expense of a larger mean squared error, as fewer observations are used to estimate $\hat{\tau}(\cdot)$, construct the tree, and compute GATE estimates.

4 Comparison with Causal Trees

This section compares aggregation trees with the causal trees proposed by Athey and Imbens (2016). In the next subsection, I highlight the methodological differences between the two approaches, which differ in the splitting strategy used to construct trees and in the output they provide. Then, I compare them in a simulation exercise to study their performance under different scenarios.

4.1 Methodological Differences

Aggregation and causal trees have the same aim: discovering heterogeneous subgroups by approximating $\tau(\cdot)$ with a multivariate step-function constructed via recursive partitioning. However, they differ in two main ways: how trees are constructed and what output they provide.

Trees are typically constructed by greedily minimizing an assumed loss function based on the MSE criterion (see Section 3.1). Let $\mathcal{T}$ be a tree constructed using a training sample $\mathcal{S}^{tr}$, and let $\mathcal{S}^{te}$ be an independent test sample. Then, when heterogeneous treatment effects

are the object of the analysis, one wants to build a tree that minimizes:

$$EMSE\left(\mathcal{T}\right) = \mathbb{E}\left[MSE\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right)\right]$$

where the expectation is taken over the joint distribution of the training and test samples and:[4]

$$\begin{aligned} MSE\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right) &= \frac{1}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \left\{ \left[\tau_i - \tilde{\tau}\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right)\right]^2 - \tau_i^2 \right\} \\ &= \frac{1}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \tilde{\tau}^2\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right) - \frac{2}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \tau_i \, \tilde{\tau}\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right) \end{aligned}$$

with $\tau_i \equiv \tau\left(X_i\right)$ and $\tilde{\tau}\left(x, \mathcal{S}, \mathcal{T}\right)$ an estimate of $\tau\left(\cdot\right)$ within the leaf $\ell\left(x, \mathcal{T}\right)$ of $\mathcal{T}$ where x falls obtained using observations in the sample $\mathcal{S}$. In practice, trees are constructed by greedily minimizing an in-sample version $MSE\left(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \mathcal{T}\right)$.[5]

The key challenge in a causal inference framework is that we do not observe τ_i. Thus, $MSE\left(\cdot, \cdot, \cdot\right)$ is an infeasible criterion and needs to be estimated. Causal and aggregation trees differ in how they estimate this criterion. Athey and Imbens (2016) propose the following estimator:

$$\widehat{MSE}_{CT}\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right) = \frac{1}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \tilde{\tau}_{CT}^2\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right) - \frac{2}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \tilde{\tau}_{CT}\left(X_i, \mathcal{S}^{te}, \mathcal{T}\right) \tilde{\tau}_{CT}\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right)$$

with:

$$\tilde{\tau}_{CT}\left(x, \mathcal{S}, \mathcal{T}\right) = \hat{\mu}\left(1, x, \mathcal{S}, \mathcal{T}\right) - \hat{\mu}\left(0, x, \mathcal{S}, \mathcal{T}\right)$$

$$\hat{\mu}\left(d, x, \mathcal{S}, \mathcal{T}\right) = \frac{1}{|i \in \mathcal{S} : X_i \in \ell\left(x, \mathcal{T}\right), D_i = d|} \sum_{i \in \mathcal{S} : X_i \in \ell(x, \mathcal{T}), D_i = d} Y_i, \quad d = 0, 1$$

[4] We are departing from the standard criterion $\mathbb{E}[\{\tau_i - \tilde{\tau}(X_i, \mathcal{S}^{tr}, \mathcal{T})\}^2]$ by subtracting $\mathbb{E}[\tau_i^2]$. Because this term does not depend on an estimator, the tree that minimizes the standard criterion also minimizes $EMSE\left(\cdot\right)$.

[5] This is what Athey and Imbens (2016) denote as the "adaptive" case, where the same sample is used to both construct and estimate the tree. Athey and Imbens (2016) also consider an alternative "honest" criterion $MSE\left(\mathcal{S}^{te}, \mathcal{S}^{hon}, \mathcal{T}\right)$ that uses different samples for construction of the tree ($\mathcal{S}^{tr}$) and treatment effect estimation ($\mathcal{S}^{hon}$). For simplicity, here I compare aggregation and causal trees focusing on the adaptive case.

In classical randomized experiments, $\widehat{MSE}_{CT}\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right)$ is an approximately unbiased estimator of $MSE\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right)$, as $\mathbb{E}\left[\tau_i | i \in \mathcal{S}^{te} : i \in \ell\left(x, \mathcal{T}\right)\right] = \mathbb{E}\left[\tilde{\tau}_{CT}\left(x, \mathcal{S}^{te}, \mathcal{T}\right)\right]$, with the expectations taken over the distribution of the test samples. We can construct causal trees by greedily minimizing the following in-sample counterpart:

$$\widehat{MSE}_{CT}\left(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \mathcal{T}\right) = -\frac{1}{|\mathcal{S}^{tr}|} \sum_{i \in \mathcal{S}^{tr}} \tilde{\tau}^2_{CT}\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right)$$

On the other hand, I propose a "plug-in" estimator of $MSE\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right)$, where I plug-in an estimate $\hat{\tau}\left(\cdot\right)$ of $\tau\left(\cdot\right)$ constructed from the training sample:

$$\widehat{MSE}_{AT}\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right) = \frac{1}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \tilde{\tau}^2_{AT}\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right) - \frac{2}{|\mathcal{S}^{te}|} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}_i \, \tilde{\tau}_{AT}\left(X_i, \mathcal{S}^{tr}, \mathcal{T}\right)$$

with:

$$\tilde{\tau}_{AT}\left(x, \mathcal{S}, \mathcal{T}\right) = \frac{1}{|i \in \mathcal{S} : X_i \in \ell\left(x, \mathcal{T}\right)|} \sum_{i \in \mathcal{S} : X_i \in \ell(x, \mathcal{T})} \hat{\tau}_i$$

If we use a consistent estimator of τ_i, then $\widehat{MSE}_{AT}\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right)$ is an approximately unbiased estimator of $MSE\left(\mathcal{S}^{te}, \mathcal{S}^{tr}, \mathcal{T}\right)$, even if the assignment to treatment is random only conditional on X_i. As before, we can construct aggregation trees by greedily minimizing the in-sample version $\widehat{MSE}_{AT}\left(\mathcal{S}^{tr}, \mathcal{S}^{tr}, \mathcal{T}\right)$. This is equivalent to choosing the splits to minimize the conditional variance of $\hat{\tau}_i$ in the resulting nodes (see equation 3.2).

The splitting strategy of the aggregation trees should result in a lower sampling variance, especially in the presence of covariates that affect the mean outcomes but do not affect the treatment effects. To see this, consider the following example. Write the potential outcomes as $Y_i\left(d\right) = \phi\left(X_i\right) + \frac{1}{2}\left(2d - 1\right)\tau\left(X_i\right) + \epsilon_i$, with $\phi\left(\cdot\right)$ a model for the mean effect. Moreover, assume that $\phi\left(X_i\right) = \frac{1}{2}X_{i1} + X_{i2}$ and $\tau\left(X_i\right) = \frac{1}{2}X_{i1}$. While exploring all the possible values of a particular covariate as splitting points, we move one observation at a time from one region of the covariate space to its complement (the child nodes). As each observation belongs to either the treated or the control group, this would change the sample average of the observed

outcomes of only one group and, through this, $\tilde{\tau}_{CT}(\cdot,\cdot,\cdot)$. Because X_{i2} has a strong impact on the mean outcomes, moving a single observation from one child node to the other according to the values of this covariate will likely produce a large change in the sample average of the observed outcomes of one group. Thus, we expect $\tilde{\tau}_{CT}(\cdot,\cdot,\cdot)$ to vary greatly with the splitting point, although X_{i2} does not enter the model for the treatment effects. This may also lead the estimator to find spurious splits on X_{i2}. In contrast, if the CATEs are precisely estimated, $\tilde{\tau}_{AT}(\cdot,\cdot,\cdot)$ should not be affected by moving a single observation from one child node to the other according to the values of X_{i2}. Thus, we expect $\tilde{\tau}_{AT}(\cdot,\cdot,\cdot)$ to vary less with the splitting point, lowering the sampling variance of the estimation.

After a deep tree has been constructed, the standard practice is to prune it according to an assumed cost-complexity criterion. Aggregation and causal trees rely on the same criterion, which is composed of two terms (see equation 3.3). The first term corresponds to the criterion $MSE\left(\mathcal{S}^{tr},\mathcal{S}^{tr},\mathcal{T}\right)$ used for constructing the trees and measures the in-sample goodness-of-fit of the tree. The second term is a regularization component that penalizes the model's complexity, defined as the number of splits of the tree. Regularization is needed to prevent overfitting: the in-sample $MSE\left(\mathcal{S}^{tr},\mathcal{S}^{tr},\mathcal{T}\right)$ always decreases with additional splits, even in the cases where its out-of-sample counterpart $MSE\left(\mathcal{S}^{te},\mathcal{S}^{tr},\mathcal{T}\right)$ actually increases.

The cost-complexity criterion is additive in these two terms and is characterized by a non-negative cost-complexity parameter that controls the relative weight of the two components. Athey and Imbens (2016) suggest using cross-validation to choose the optimal cost-complexity parameter, thus selecting the "best" tree and providing the researcher with a single partition of the covariate space. In contrast, aggregation trees refrain from selecting a single partition and explore different granularity levels by varying the cost-complexity parameter, thus generating a consistent sequence of optimal groupings.

4.2 Simulation Results

I choose the DGPs to replicate the simulation study in Athey and Imbens (2016). Potential outcomes are generated as follows:

$$Y_i(d) = \phi(X_i) + \frac{1}{2}(2d-1)\tau(X_i) + \epsilon_i, \quad d = 0, 1$$

with $\phi(\cdot)$ a model for the mean effect and $\epsilon_i \sim \mathcal{N}(0, 0.01)$. Observed outcomes are generated according to Assumption 2.1. The treatment is always randomly assigned as in a Bernoulli experiment, with the marginal probability of treatment equal to 0.5. Covariates are generated as $X_i \sim \mathcal{N}(0, 1)$ and are independent of one another and of ϵ_i. Under this setting, CATEs are identified from observable data (see equation 2.1).

I consider three different designs that differ in the number of covariates p and the models for the mean effect $\phi(\cdot)$ and the treatment effect $\tau(\cdot)$:

$$\begin{aligned}
&\textit{Design 1.}\ p = 2, && \phi(X_i) = \frac{1}{2}X_{i1} + X_{i2}, && \tau(X_i) = \frac{1}{2}X_{i1} \\
&\textit{Design 2.}\ p = 10, && \phi(X_i) = \frac{1}{2}\sum_{j=1}^{2} X_{ij} + \sum_{j=3}^{6} X_{ij}, && \tau(X_i) = \sum_{j=1}^{2} X_{ij}\mathbb{1}\left(X_{ij} > 0\right) \\
&\textit{Design 3.}\ p = 20, && \phi(X_i) = \frac{1}{2}\sum_{j=1}^{4} X_{ij} + \sum_{j=5}^{8} X_{ij}, && \tau(X_i) = \sum_{j=1}^{4} X_{ij}\mathbb{1}\left(X_{ij} > 0\right)
\end{aligned}$$

Notice that each design contains covariates that enter both the models for the mean effect and the treatment effect, covariates that enter the model for the mean effect but not the model for the treatment effect, and "noise" covariates that do not affect outcomes (*Design 1* does not feature any noise covariate). For each design, I consider four sample sizes, $n \in \{500, 1000, 2000, 4000\}$. Thus, I consider overall twelve different scenarios.

After drawing a sample of size n, I first construct a causal tree and then two aggregation trees. I spit the sample into a training sample $\mathcal{S}^{tr}$, used to construct the trees, and an honest sample $\mathcal{S}^{hon}$, used to estimate the GATEs. Thus, all trees are honest. The two samples are of equal sizes.

To build the aggregation trees, I estimate $\hat{\tau}(\cdot)$ using the X-learner (Künzel, Sekhon, Bickel, & Yu, 2019) and the causal forest (Athey, Tibshirani, & Wager, 2019) estimators. Then, I use the estimates to perform the tree-growing and tree-pruning steps. All these steps use only observations in $\mathcal{S}^{tr}$. To ensure a fair comparison, I select the subtrees with the same number of leaves of the cross-validated causal tree.

To estimate the GATEs, the causal trees pick the difference in mean outcomes between the treated and control units in each leaf, as in Athey and Imbens (2016). On the other hand, the aggregation trees construct and average doubly-robust scores (see Section 3.3). Honest regression forests and 5-fold cross-fitting are used to estimate the propensity score and the conditional mean function of the outcome. The GATEs are always estimated using observations in $\mathcal{S}^{hon}$.

I rely on an external validation sample of size $10,000$ to assess the quality of the approximation. This large number of observations helps to minimize the sampling variance. Three performance measures are computed: the squared bias, the variance, and the mean squared error for the prediction of each observation in the validation sample:

$$
\begin{aligned}
Bias^2(x) &= \left[\frac{1}{R}\sum_{r=1}^{R}\hat{\tau}_r(x) - \tau(x)\right]^2 \\
Var(x) &= \frac{1}{R}\sum_{r=1}^{R}\left[\hat{\tau}_r(x) - \frac{1}{R}\sum_{r=1}^{R}\hat{\tau}_r(x)\right]^2 \\
MSE(x) &= Bias^2(x) + Var(x)
\end{aligned}
$$

with x a generic point in the validation sample, R the number of replications, and $\hat{\tau}_r(\cdot)$ the CATE function estimated by the tree at the r-th replication. I summarize these performance measures by averaging over the validation sample.

Additionally, I estimate model (3.4) to get standard errors for the GATEs estimated by the causal trees and model (3.5) to get standard errors for the GATEs estimated by the aggregation trees. Both models are estimated using observations in $\mathcal{S}^{hon}$. I use the estimated standard errors to construct conventional 95% confidence intervals.

Table 4.1 displays the results obtained with $R = 1,000$ replications (see Appendix B for results with adaptive trees). Broadly speaking, aggregation trees perform well relative to causal trees, showing a better prediction accuracy uniformly over the considered scenarios. The superior performance of aggregation trees is particularly noticeable in *Design 1*, where the MSE of causal trees is between 36% and 79% larger than the MSE of aggregation trees. The advantage of aggregation trees is smaller in *Design 2* (between 6-19%) and *Design 3* (between 1-8%). In all designs the advantage increases with the sample size. AT_{XL} and AT_{CF} show similar perfomances.

	Design 1				*Design 2*				*Design 3*			
	500	1,000	2,000	4,000	500	1,000	2,000	4,000	500	1,000	2,000	4,000
Panel 1: $\overline{MSE}$												
AT_{XL}	0.141	0.084	0.050	0.031	0.701	0.611	0.483	0.303	1.423	1.356	1.223	0.985
AT_{CF}	0.141	0.085	0.051	0.031	0.690	0.607	0.480	0.300	1.416	1.338	1.215	0.964
CT	0.193	0.127	0.082	0.056	0.746	0.655	0.533	0.358	1.449	1.372	1.257	1.039
Panel 2: $\overline{Bias^2}$												
AT_{XL}	0.074	0.038	0.022	0.012	0.619	0.500	0.322	0.176	1.340	1.252	1.026	0.682
AT_{CF}	0.079	0.041	0.024	0.013	0.594	0.483	0.305	0.159	1.308	1.206	0.990	0.627
CT	0.093	0.047	0.024	0.013	0.630	0.533	0.362	0.178	1.338	1.274	1.100	0.750
Panel 3: $\overline{Var}$												
AT_{XL}	0.067	0.047	0.029	0.019	0.081	0.111	0.161	0.127	0.084	0.104	0.197	0.304
AT_{CF}	0.062	0.044	0.027	0.018	0.096	0.124	0.175	0.141	0.108	0.132	0.226	0.337
CT	0.100	0.080	0.058	0.043	0.116	0.122	0.172	0.181	0.111	0.098	0.157	0.289
Panel 4: Coverage for 95% CI												
AT_{XL}	0.98	0.95	0.92	0.84	0.95	0.96	0.96	0.96	0.95	0.93	0.93	0.94
AT_{CF}	0.98	0.96	0.92	0.84	0.95	0.95	0.96	0.95	0.94	0.94	0.92	0.92
CT	0.90	0.91	0.93	0.92	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94
Panel 5: $\overline{\lvert\mathcal{T}\rvert}$												
	2.56	3.57	4.5	6.1	1.18	1.43	2.07	3.04	1.11	1.29	1.79	3.26

Table 4.1: Comparison with causal trees. The first three panels report the average over the validation sample of MSE ($\overline{MSE}$), $Bias^2$ ($\overline{Bias^2}$) and Var ($\overline{Var}$). The fourth panel reports coverage rates for 95% confidence intervals. The last panel reports the average number of leaves in the different designs. All trees are honest.

The second and the third panels of Table 4.1 provide additional insights into the superior performance of aggregation trees by investigating the two components of the mean squared error. All the considered estimators have some bias, which gets larger in more complex

designs. The bias of aggregation trees is smaller than the bias of causal trees, with few exceptions where the two estimators tie. However, causal trees never feature the lowest bias in any scenario. Aggregation trees also feature a lower sampling variance than causal trees, resulting from the splitting strategy discussed in the previous subsection. This effect is particularly pronounced in *Design 1*, where the variance of causal trees is between 49% and 132% larger than the variance of aggregation trees. However, in *Design 3* causal trees outperform aggregation trees in terms of sampling variance in larger samples, although by a smaller margin (5-25%).

Finally, the fourth panel of Table 4.1 displays coverage rates for 95% confidence intervals. Both aggregation and causal trees feature rates close to the nominal rate. However, causal trees never achieve the nominal rate, and aggregation trees are generally conservative.

Comparing these results with Table B.1 allows us to evaluate the costs and benefits of honesty. Using different data for constructing the trees and treatment effect estimation greatly benefits inference. Coverage rates of adaptive trees are generally below the nominal rate, particularly those of causal trees that can be as low as 66%. However, honesty comes at the expense of a larger mean squared error. The cost of honesty in terms of MSE is between 9% and 34% for causal trees, with two exceptions in *Design 2* and *Design 3* where, for the smallest sample size, honest trees have a lower MSE than adaptive trees. For aggregation trees, the price to be paid is higher, ranging between 4% and 63%. This happens because the estimation of the CATEs and the GATEs hinges on flexible machine learning estimators, which are generally more sensitive to using fewer observations in training the models than the GATE estimator of causal trees.

5 Empirical Example

As documented in Almond, Chay, and Lee (2005), infants born at low birth weight (LBW) can impose substantial costs on society, with estimated expected costs of delivery

and initial care exceeding 100,000$ (at prices of year 2000) for babies weighing 1,000 grams at birth.[6] Moreover, LBW is associated with a higher risk of death within one year of birth. For these reasons, birth weight is considered the primary measure of a baby's health and is often the direct target of health policies. Thus, understanding what causes LBW is crucial.

The effect of maternal smoking on LBW has received considerable attention in the literature, for it is regarded as one of the most significant and modifiable risk factors. Several studies confirm that smoking during pregnancy is associated with lower average birth weight, with estimated ATEs ranging between -600 and -100 grams (e.g., Almond et al., 2005; Abrevaya, 2006). As for effect heterogeneity, it is now well understood that the effects are increasingly negative with the mother's age (Abrevaya, Hsu, & Lieli, 2015; Lee, Okui, & Whang, 2017; Zimmert & Lechner, 2019; Fan, Hsu, Lieli, & Zhang, 2022). Finally, treatment heterogeneity has also been investigated. Cattaneo (2010) considers different smoking intensities as different treatments and shows that higher smoking intensities lead to more negative effects. Heiler and Knaus (2021) find that heterogeneous effects can be partly explained by different smoking behavior of ethnic and age groups.

5.1 Data

I analyze the same data set as Almond et al. (2005), also used by Cattaneo (2010) and Heiler and Knaus (2021).[7] The clean data contain 435,124 observations measured in Pennsylvania between 1989 and 1991. The outcome of interest is the infant's weight at birth in grams. The treatment indicator is equal to one if the mother smoked during pregnancy and zero otherwise. The pre-treatment covariate vector contains 39 confounders and heterogeneity variables, providing information on the mother's and father's background characteristics (age, ethnicity, whether the mother was married or foreign-born), mother's behavior possibly associated with smoking (whether she drank alcohol during pregnancy, how many drinks per week), maternal medical risk factors not affected by smoking during pregnancy, and birth

[6] An infant is considered born at low birth weight if she weighs less than 2,500 grams at birth.

[7] I thank Matias Cattaneo for sharing the full data.

characteristics (e.g., the month of birth, whether the infant is first born, number and quality of prenatal care visits). See Table A.1 in Appendix A for a description of all the variables used in the analysis.

To avoid common support issues, I drop children whose mothers and fathers were particularly young or old at birth, or who attended more than thirty prenatal care visits. Moreover, I drop children whose mothers used to consume more than ten alcoholic drinks per week during pregnancy. Overall, 596 observations are dropped from the original data set. Table A.2 in Appendix A reports the summary statistics for the treated and control groups in the final sample. The table shows sample averages and standard deviations for each variable, together with two measures of difference in the distribution across treatment arms: the normalized difference, measuring the difference between the locations of the distributions, and the logarithm of the ratio of standard deviations, measuring the difference in the dispersion of the distributions. Overall, the sample appears to be sufficiently balanced, with only five relatively unbalanced covariates: `meduc`, `unmarried` and `feduc` show strong differences in locations, while `alcohol` and `n_drink` show strong differences in the dispersion. These results are robust to the inclusion of the dropped observations.

Recall that the identification of ATE, GATEs and CATEs requires Assumptions 2.1–2.4 to hold. Assumption 2.1 (SUTVA) may raise some concerns, as passive smoking might generate spillover effects. However, only around 19% of the mothers report smoking, thus reducing the risk of interference. Assumption 2.2 (exogeneity of the covariates) is considered reasonable in this setting, as all covariates, including maternal medical risk factors, are believed not to be affected by the treatment (Almond et al., 2005). Assumption 2.3 (unconfoundedness) requires to control for all variables that influence both the infant's weight at birth and the decision of the mother to smoke. Parents' economic factors, such as their income, might affect the decision to smoke. However, they are unlikely to causally influence the infant's birth weight, so their absence in the data set is not a concern for identification. Finally, Figure A.1 in Appendix A shows kernel density estimates of the propensity scores. Overall,

there is substantial overlap in the propensity scores for the treated and control groups. However, the number of control units with a large score is quite low. Therefore, to increase the credibility of the analysis, I trim the sample by dropping all the observations that have an estimated propensity score larger than 0.6. This reduces the sample size by $8,354$ units. Most of the summary statistics shown in Table A.2 stay unchanged. The trimming reduces the proportion of mothers consuming alcohol and the average number of weekly drinks during pregnancy in the treatment group, resulting in a more balanced sample.

5.2 Constructing the Sequence of Groupings

To construct the sequence of optimal groupings, we first need to estimate the CATEs. For this purpose, I use the causal forest estimator of Athey et al. (2019). To achieve valid inference about the GATEs, I split the sample into a training sample and an honest sample of equal sizes. The forest is built using only the training sample. Figure A.2 displays the estimated CATEs sorted according to their size together with 95% confidence intervals. Almost all the predicted effects are negative and statistically different from zero at the 5% significance level.

I then construct the set of admissible groupings by approximating the estimated CATEs via a decision tree. Again, I use only observations in the training sample. Given the discrete nature of the variables at hand, there is no need for discretization. Once the tree has been constructed, I use observations in the honest sample to estimate the GATEs by constructing and averaging doubly-robust scores (Robins & Rotnitzky, 1995).

Figure 5.1 displays the results. To understand how the tree is constructed, we start from the root, which provides the estimated ATE (-204 grams). We split observations into children born from adult mothers (the root's left child) and children born from young mothers (the root's right child). Among all the possible groupings composed of two groups, this is the one that maximizes heterogeneity in the treatment effects. The first group represents 77% of units and features an estimated GATE of around -217 grams. The second group represents

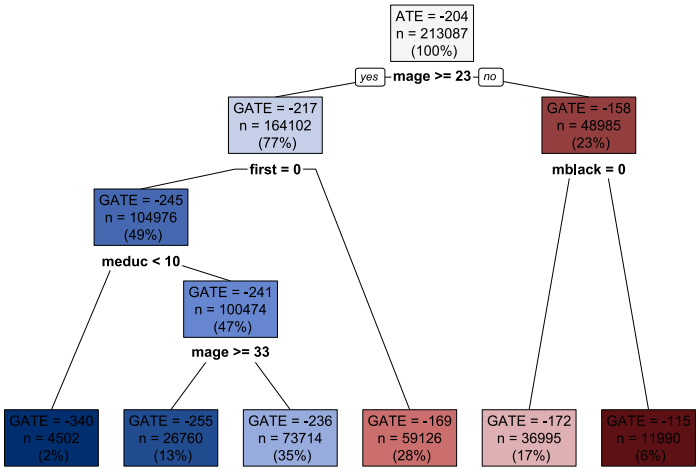

Figure 5.1: Aggregation tree, constructed in the training sample. Each node displays the GATE and the number and percentage of units belonging to each subgroup. The GATEs are estimated by averaging doubly-robust scores constructed using the honest sample. Blue and red shades denote groups with GATEs stronger (i.e., more negative) and lighter (i.e., more positive) than the ATE.

23% of units and features an estimated GATE of around −158 grams.

We then further divide these groups into smaller subgroups. Children born from adult mothers are divided according to whether they are first born, while children born from young mothers are divided according to whether their mother is black. We thus obtain four groups represented by as many nodes.

We repeat this process until some stopping criterion is met. Finally, we construct the sequence of optimal groupings by progressively aggregating the two subgroups for which the loss in explained heterogeneity resulting from aggregation is smallest. Figure 5.2 provides a visualization of this process. Reading the figure from left to right and from top to bottom, each panel corresponds to a grouping in the sequence, each obtained by progressively collapsing the node for which the loss in explained heterogeneity is minimized. The figure makes it immediate to appreciate that the sequence of groupings is consistent. Because groupings are nested, we never break previous groups when moving to coarser levels, thus obtaining consistency of the results across the different granularity levels.

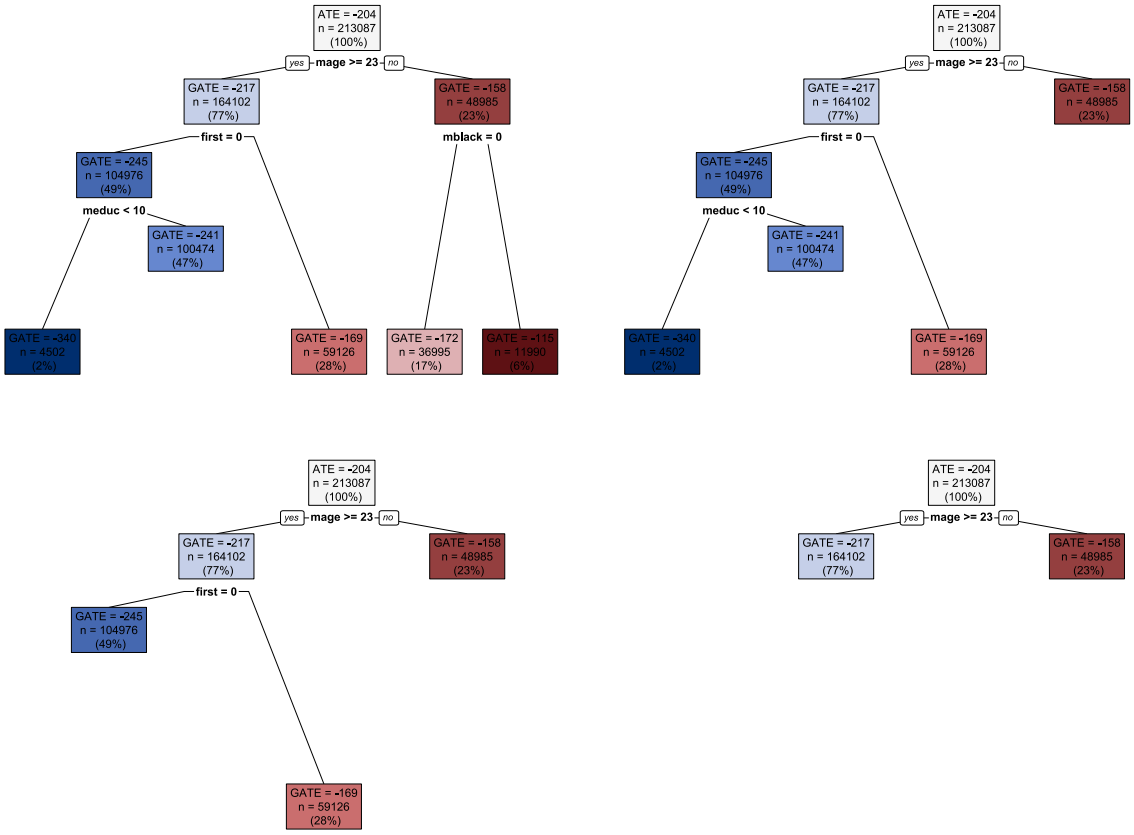

Figure 5.2: Sequence of optimal groupings, obtained by progressively collapsing the node for which the loss in explained heterogeneity is minimized.

5.3 Discussion

Researchers interested in uncovering the effects of a particular policy generally face two crucial tasks: assessing systematic treatment effect heterogeneity and, if evidence of heterogeneity is found, understanding the mechanisms behind.

One way to tackle the first task is to look at the distribution of the estimated CATEs. For instance, we could look at Figure A.2 and conclude that there is systematic heterogeneity in the treatment effects. However, because of estimation noise high variation in the predictions does not necessarily imply that effects are heterogeneous. A better approach consists of investigating whether different subgroups feature different average effects. For this purpose, I choose the optimal grouping composed of four groups (displayed on the top right panel

of Figure 5.2) and estimate model (3.5) to get standard errors for the GATEs. I use only observations in the honest sample to estimate the model. Table 5.1 reports point estimates and 95% confidence intervals. We find strong differences in treatment effects, with the estimated GATEs ranging from -339 grams for the most affected group (*Leaf 1*) to -157 grams for the least affected group (*Leaf 4*). However, *Leaf 3* and *Leaf 4* feature similar GATEs. Table 5.1 also displays the differences in the GATEs across all pairs of groups, together with p-values to test the null hypotheses that each difference equals zero. To account for multiple hypotheses testing, I adjust the p-values using the procedure of Holm (1979). The difference in the GATEs of *Leaf 3* and *Leaf 4* is about 10 grams. We fail to reject the null hypothesis that this difference is zero at any conventional confidence level. On the other hand, the differences between all other pairs of groups are statistically significant. Overall, these results provide evidence of systematic heterogeneity in treatment effects.

To understand what factors drive this heterogeneity, we look at how treatment effects relate to the observable covariates. One possibility is to ask which variables have been

	Leaf 1	*Leaf 2*	*Leaf 3*	*Leaf 4*
GATEs	-339.522 [-390.478, -288.566]	-240.812 [-257.217, -224.407]	-168.656 [-191.806, -145.506]	-157.884 [-168.903, -146.865]
Leaf 1	-	-	-	-
	(-)	(-)	(-)	(-)
Leaf 2	98.709	-	-	-
	(0.001)	(-)	(-)	(-)
Leaf 3	170.866	72.156	-	-
	(0.001)	(0.001)	(-)	(-)
Leaf 4	181.638	82.929	10.772	-
	(0.001)	(0.001)	(0.410)	(-)

Table 5.1: Point estimates and 95% confidence intervals for the GATEs. Leaves are sorted in increasing order of the GATEs. Additionally, differences in the GATEs across all pairs of leaves are displayed. p-values to test the null hypothesis that a single difference is zero are adjusted using Holm's procedure and reported in parenthesis under each point estimate.

used by the tree-growing process to construct groups and measure their relative importance. However, we should not conclude that covariates not used for splitting are not related to heterogeneity. If two covariates are highly correlated, trees generally split on only one of them.

A better alternative is to ask how the average characteristics of the units vary across subgroups (e.g., Chernozhukov et al., 2017). Table 5.2 shows how the average value of some selected covariates changes across *Leaves 1–Leaves 4* (see Table A.3 for the remaining covariates). We find that the least affected group is composed of children born to younger mothers and fathers, suggesting that the effects are more negative at higher ages (this is in line with results from previous literature, e.g., Abrevaya et al., 2015; Zimmert & Lechner, 2019). On the other hand, no first-born infants belong to the most affected group, which also features the lowest average parental educational attainment and number of prenatal care visits. Additionally, such group shows the lowest proportion of mothers that suffered

	Leaf 1		*Leaf 2*		*Leaf 3*		*Leaf 4*	
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
Parent's characteristics								
`mage`	29.331	(0.073)	29.872	(0.013)	28.255	(0.016)	19.596	(0.009)
`meduc`	7.174	(0.040)	13.363	(0.006)	13.822	(0.009)	11.381	(0.008)
`fage`	31.645	(0.094)	32.002	(0.017)	30.472	(0.022)	22.740	(0.019)
`feduc`	8.175	(0.053)	13.371	(0.008)	13.707	(0.010)	11.375	(0.012)
Birth characteristics								
`first`	0.000	(-)	0.000	(-)	1.000	(-)	0.664	(0.002)
`plural`	0.018	(0.002)	0.016	(0.001)	0.019	(0.001)	0.011	(0.001)
`n_prenatal`	7.811	(0.062)	11.222	(0.010)	11.785	(0.012)	10.040	(0.018)
Maternal medical risk factors								
`diabete`	0.016	(0.002)	0.021	(0.001)	0.022	(0.001)	0.009	(0.001)
`anemia`	0.012	(0.002)	0.008	(0.001)	0.005	(0.001)	0.016	(0.001)
`hyper`	0.014	(0.002)	0.017	(0.001)	0.044	(0.001)	0.031	(0.001)

Table 5.2: Average characteristics of units in each leaf, obtained by regressing each covariate on a set of dummies denoting leaf membership. Standard errors are estimated via the Eicker-Huber-White estimator. Leaves are sorted in increasing order of the GATEs.

from pregnancy-associated hypertension. Finally, the proportion of twins or higher births and the proportion of mothers suffering from diabetes or anemia do not vary much across groups. Overall, these results suggest that several maternal medical risk factors and birth characteristics are unrelated to treatment effect heterogeneity, which is mainly driven by parent's characteristics.

6 Conclusion

The methodology proposed in this paper provides a nonparametric data-driven approach to discovering heterogeneous subgroups in a selection-on-observables framework. The approach constructs a sequence of groupings, one for each level of granularity. The sequence is nested in the sense that subgroups formed at a given level of granularity are never broken at coarser levels. We show that each grouping is "optimal" in the sense that the loss in explained heterogeneity resulting from aggregation is minimized.

For a particular grouping, point estimates and standard error for the GATEs are obtained by fitting an appropriate linear model. Under an honesty condition, we can use the estimated standard errors to construct asymptotically valid symmetric confidence intervals for the GATEs.

An empirical exercise revisiting the effect of maternal smoking on children's birth weight illustrates the practical implementation of the proposed methodology. Results suggest that treatment effect heterogeneity is mainly driven by the mother's and father's characteristics and is unrelated to several maternal risk factors and birth characteristics.

Although the aggregation trees illustrated in the empirical section of this paper have been grown using all the available units, we can, in principle, construct trees on separate subpopulations of interest defined by the values of a categorical covariate (e.g., gender). This way, the researcher can exploit the available prior knowledge to address specific research questions.

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics*, *21*(4), 489–519.

Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, *33*(4), 485–505.

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, *120*(3), 1031–1083.

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth; Brooks.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, *155*(2), 138–154.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (2017). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint arXiv:1712.04802*.

Cockx, B., Lechner, M., & Bollens, J. (2023). Priority to unemployed immigrants? a causal machine learning evaluation of training in belgium. *Labour Economics*, *80*.

Cotterman, R., & Peracchi, F. (1992). Classification and aggregation: An application to industrial classification in cps data. *Journal of Applied Econometrics*, *7*(1), 31–51.

Fan, Q., Hsu, Y.-C., Lieli, R. P., & Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, *40*(1), 313–327.

Heiler, P., & Knaus, M. C. (2021). Effect or treatment heterogeneity? policy evaluation with aggregated and disaggregated treatments. *arXiv preprint arXiv:2110.01427*.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, *35*(3), 157–74.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.

Kitagawa, T., & Tetenov, A. (2021). Equality-minded treatment choice. *Journal of Business & Economic Statistics*, *39*(2), 561–574.

Knaus, M. C. (2022). Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, *25*(3), 602–627.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Lee, S., Okui, R., & Whang, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, *32*(7), 1207–1225.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczyc*, *10*, 1–51.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Semenova, V., & Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289.

Zimmert, M., & Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.

Appendix A Data

Label	Description
OUTCOME.	
bweight	Infant birth weight (in grams)
TREATMENT.	
smoke	=1 if mother smoked during pregnancy
COVARIATES.	
Mother's characteristics.	
mage	Mother's age
meduc	Mother's educational attainment
mwhite	=1 if mother is white
mblack	=1 if mother is black
mhispan	=1 if mother is hispanic
foreign_born	=1 if mother is foreign born
unmarried	=1 if mother is unmarried
alcohol	=1 if mother drank alcohol during pregnancy
n_drink	Number of drinks per week during pregnancy
Father's characteristics.	
fage	Father's age
feduc	Father's educational attainment
fwhite	=1 if father is white
fblack	=1 if father is black
fhispan	=1 if father is hispanic
Birth characteristics.	
birthmonth1	=1 if birth in January
birthmonth2	=1 if birth in February
birthmonth3	=1 if birth in March
birthmonth4	=1 if birth in April
birthmonth5	=1 if birth in May
birthmonth6	=1 if birth in June
birthmonth7	=1 if birth in July
birthmonth8	=1 if birth in August
birthmonth9	=1 if birth in September
birthmonth10	=1 if birth in October
birthmonth11	=1 if birth in November
birthmonth12	=1 if birth in December
first	=1 if the infant is first born
plural	=1 if twins or greater birth
n_prenatal	Number of prenatal care visits
prenatal0	=1 if no prenatal visit
prenatal1	=1 if first prenatal visit in first trimester of pregnancy
prenatal2	=1 if first prenatal visit in second trimester of pregnancy
prenatal3	=1 if first prenatal visit in third trimester of pregnancy
adequacy1	=1 if adequacy of care is adequate (Kessner Index)
adequacy2	=1 if adequacy of care is intermediate (Kessner Index)
adequacy3	=1 if adequacy of care is inadequate (Kessner Index)
Maternal medical risk factors.	
diabete	=1 if mother is diabetic
anemia	=1 if mother is anemic
hyper	=1 if mother had pregnancy-associated hypertension

Table A.1: Description of variables in the data set.

	Treated		Controls		Overlap measures	
	($n_t = 81,388$)		($n_c = 353,140$)			
	Mean	(S.D.)	Mean	(S.D.)	$\hat{\Delta}_j$	$\hat{\Gamma}_j$
mage	25.503	(5.372)	27.340	(5.553)	-0.336	-0.033
meduc	11.783	(1.883)	13.088	(2.430)	-0.600	-0.255
mwhite	0.850	(0.357)	0.865	(0.342)	-0.043	0.044
mblack	0.147	(0.354)	0.116	(0.321)	0.090	0.099
mhispan	0.020	(0.140)	0.031	(0.173)	-0.068	-0.209
foreign_born	0.019	(0.138)	0.056	(0.229)	-0.190	-0.504
unmarried	0.444	(0.497)	0.196	(0.397)	0.552	0.225
alcohol	0.045	(0.207)	0.007	(0.080)	0.243	0.943
n_drink	0.123	(0.729)	0.013	(0.212)	0.205	1.237
fage	28.451	(6.556)	29.640	(6.264)	-0.185	0.046
feduc	11.686	(2.628)	13.102	(2.800)	-0.522	-0.064
fwhite	0.831	(0.375)	0.857	(0.350)	-0.072	0.068
fblack	0.162	(0.369)	0.123	(0.329)	0.112	0.115
fhispan	0.028	(0.166)	0.033	(0.178)	-0.025	-0.068
birthmonth1	0.081	(0.273)	0.078	(0.268)	0.011	0.017
birthmonth2	0.074	(0.262)	0.075	(0.264)	-0.003	-0.004
birthmonth3	0.082	(0.274)	0.086	(0.280)	-0.015	-0.023
birthmonth4	0.076	(0.265)	0.083	(0.277)	-0.027	-0.043
birthmonth5	0.081	(0.273)	0.087	(0.282)	-0.022	-0.032
birthmonth6	0.083	(0.277)	0.086	(0.280)	-0.009	-0.013
birthmonth7	0.092	(0.289)	0.089	(0.284)	0.011	0.016
birthmonth8	0.094	(0.291)	0.089	(0.285)	0.017	0.024
birthmonth9	0.090	(0.286)	0.087	(0.282)	0.011	0.016
birthmonth10	0.086	(0.281)	0.084	(0.277)	0.009	0.013
birthmonth11	0.078	(0.269)	0.077	(0.267)	0.003	0.005
birthmonth12	0.082	(0.275)	0.079	(0.270)	0.012	0.019
first	0.367	(0.482)	0.438	(0.496)	-0.146	-0.029
plural	0.015	(0.120)	0.016	(0.127)	-0.015	-0.060
n_prenatal	10.210	(3.989)	11.125	(3.395)	-0.247	0.161
prenatal0	0.025	(0.156)	0.007	(0.086)	0.141	0.603
prenatal1	0.718	(0.450)	0.838	(0.368)	-0.292	0.200
prenatal2	0.204	(0.403)	0.124	(0.330)	0.216	0.200
prenatal3	0.047	(0.212)	0.026	(0.159)	0.114	0.289
adequacy1	0.631	(0.483)	0.762	(0.426)	-0.287	0.125
adequacy2	0.258	(0.437)	0.184	(0.388)	0.178	0.121
adequacy3	0.105	(0.306)	0.049	(0.216)	0.210	0.347
diabete	0.018	(0.132)	0.018	(0.135)	-0.006	-0.022
anemia	0.014	(0.119)	0.008	(0.092)	0.057	0.266
hyper	0.019	(0.138)	0.029	(0.168)	-0.065	-0.202

Table A.2: Balance between treatment and control groups. The last two columns report the estimated normalized differences ($\hat{\Delta}_j$) and logarithms of the ratio of standard deviations ($\hat{\Gamma}_j$).

	Leaf 1		*Leaf 2*		*Leaf 3*		*Leaf 4*	
	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)	Mean	(S.D.)
mwhite	0.879	(0.005)	0.890	(0.001)	0.913	(0.001)	0.745	(0.002)
mblack	0.072	(0.004)	0.095	(0.001)	0.064	(0.001)	0.246	(0.002)
mhispan	0.095	(0.004)	0.020	(0.001)	0.013	(0.001)	0.059	(0.001)
foreign_born	0.149	(0.005)	0.046	(0.001)	0.053	(0.001)	0.041	(0.001)
unmarried	0.155	(0.005)	0.124	(0.001)	0.115	(0.001)	0.581	(0.002)
alcohol	0.007	(0.001)	0.011	(0.001)	0.011	(0.001)	0.007	(0.001)
n_drink	0.019	(0.004)	0.023	(0.001)	0.022	(0.001)	0.017	(0.001)
fwhite	0.871	(0.005)	0.884	(0.001)	0.908	(0.001)	0.724	(0.002)
fblack	0.072	(0.004)	0.102	(0.001)	0.069	(0.001)	0.263	(0.002)
fhispan	0.097	(0.004)	0.022	(0.001)	0.015	(0.001)	0.066	(0.001)
birthmonth1	0.094	(0.004)	0.077	(0.001)	0.078	(0.001)	0.080	(0.001)
birthmonth2	0.082	(0.004)	0.075	(0.001)	0.076	(0.001)	0.076	(0.001)
birthmonth3	0.086	(0.004)	0.087	(0.001)	0.083	(0.001)	0.084	(0.001)
birthmonth4	0.075	(0.004)	0.082	(0.001)	0.083	(0.001)	0.078	(0.001)
birthmonth5	0.081	(0.004)	0.088	(0.001)	0.086	(0.001)	0.082	(0.001)
birthmonth6	0.086	(0.004)	0.087	(0.001)	0.085	(0.001)	0.085	(0.001)
birthmonth7	0.079	(0.004)	0.090	(0.001)	0.089	(0.001)	0.087	(0.001)
birthmonth8	0.091	(0.004)	0.089	(0.001)	0.091	(0.001)	0.092	(0.001)
birthmonth9	0.086	(0.004)	0.087	(0.001)	0.089	(0.001)	0.089	(0.001)
birthmonth10	0.088	(0.004)	0.084	(0.001)	0.083	(0.001)	0.084	(0.001)
birthmonth11	0.070	(0.004)	0.076	(0.001)	0.079	(0.001)	0.079	(0.001)
birthmonth12	0.082	(0.004)	0.078	(0.001)	0.078	(0.001)	0.084	(0.001)
prenatal0	0.020	(0.002)	0.007	(0.001)	0.002	(0.001)	0.020	(0.001)
prenatal1	0.467	(0.007)	0.866	(0.001)	0.918	(0.001)	0.660	(0.002)
prenatal2	0.357	(0.007)	0.104	(0.001)	0.066	(0.001)	0.257	(0.002)
prenatal3	0.149	(0.005)	0.018	(0.001)	0.011	(0.001)	0.057	(0.001)
adequacy1	0.340	(0.007)	0.787	(0.001)	0.854	(0.001)	0.569	(0.002)
adequacy2	0.378	(0.007)	0.170	(0.001)	0.124	(0.001)	0.313	(0.002)
adequacy3	0.272	(0.007)	0.038	(0.001)	0.018	(0.001)	0.111	(0.001)

Table A.3: Average characteristics of units in each leaf, obtained by regressing each covariate on a set of dummies denoting leaf membership. Standard errors are estimated via the Eicker-Huber-White estimator. Leaves are sorted in increasing order of the GATEs.

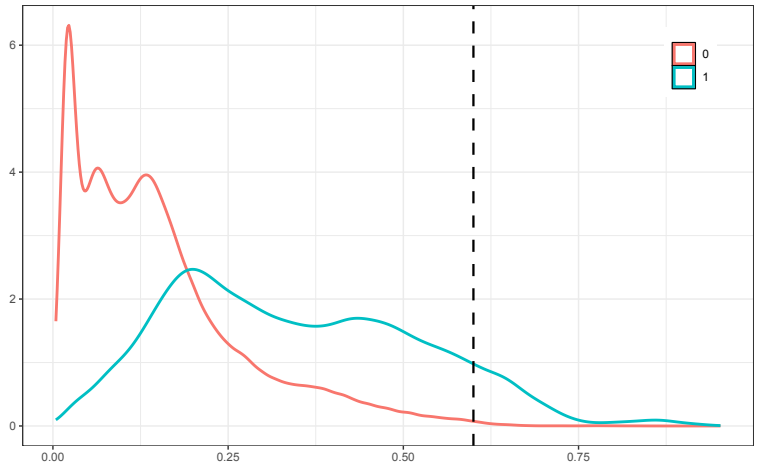

Figure A.1: Kernel density estimates of the propensity score. Propensity scores are estimated by an honest regression forest. The dashed line shows the sample trimming.

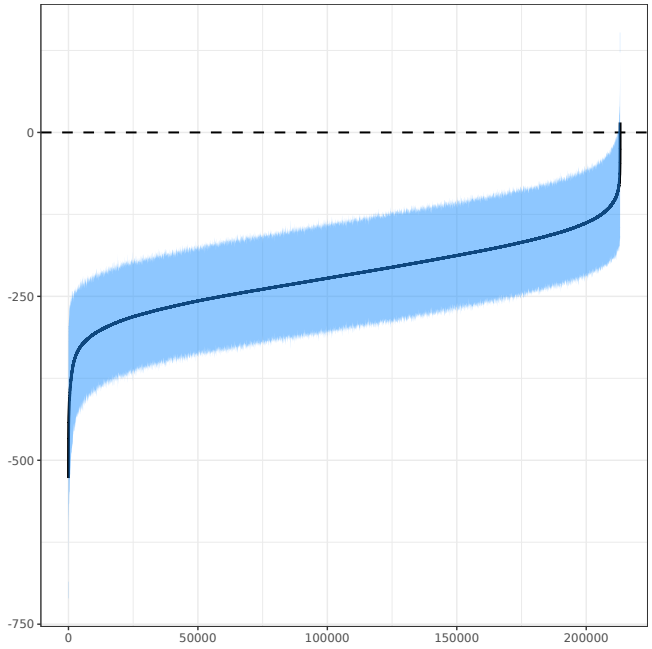

Figure A.2: Sorted CATEs and 95% confidence intervals. Predictions on the honest sample are shown. Standard errors are smoothed by a Nadaraya-Watson regression.

Appendix B Further Simulation Results

	Design 1				*Design 2*				*Design 3*			
	500	1,000	2,000	4,000	500	1,000	2,000	4,000	500	1,000	2,000	4,000
Panel 1: $\overline{MSE}$												
AT_{XL}	0.095	0.056	0.035	0.020	0.612	0.480	0.298	0.210	1.364	1.235	0.976	0.725
AT_{CF}	0.094	0.057	0.035	0.020	0.611	0.478	0.295	0.207	1.358	1.228	0.952	0.688
CT	0.169	0.107	0.072	0.047	0.756	0.599	0.399	0.280	1.461	1.360	1.118	0.868
Panel 2: $\overline{Bias^2}$												
AT_{XL}	0.046	0.026	0.015	0.008	0.458	0.299	0.168	0.137	1.240	1.002	0.654	0.466
AT_{CF}	0.048	0.028	0.016	0.008	0.445	0.282	0.150	0.121	1.209	0.968	0.599	0.413
CT	0.017	0.009	0.004	0.003	0.410	0.232	0.096	0.074	1.192	0.932	0.543	0.327
Panel 3: $\overline{Var}$												
AT_{XL}	0.049	0.030	0.020	0.012	0.154	0.181	0.131	0.073	0.124	0.232	0.321	0.259
AT_{CF}	0.046	0.028	0.019	0.011	0.166	0.196	0.145	0.086	0.149	0.261	0.353	0.275
CT	0.152	0.099	0.068	0.044	0.346	0.367	0.303	0.206	0.269	0.428	0.575	0.540
Panel 4: Coverage for 95% **CI**												
AT_{XL}	0.97	0.91	0.84	0.71	0.93	0.94	0.96	0.96	0.91	0.89	0.92	0.93
AT_{CF}	0.96	0.92	0.84	0.71	0.92	0.93	0.95	0.96	0.89	0.87	0.89	0.92
CT	0.72	0.74	0.73	0.74	0.75	0.73	0.77	0.79	0.78	0.66	0.68	0.70
Panel 5: $\overline{\lvert\mathcal{T}\rvert}$												
	2.88	3.89	5.09	7.08	1.47	2.06	3.08	3.93	1.25	1.82	3.26	5.63

Table B.1: Comparison with causal trees. The first three panels report the average over the validation sample of MSE ($\overline{MSE}$), $Bias^2$ ($\overline{Bias^2}$) and Var ($\overline{Var}$). The fourth panel reports coverage rates for 95% confidence intervals. The last panel reports the average number of leaves in the different designs. All trees are adaptive.

Appendix C Best Greedy Approximation

Fix a partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$ of $\mathcal{X}$. Then, the approximation problem (3.1) is equivalent to:

$$\min_{c_1,\ldots,c_G} \mathbb{E}\left[\left(\tau(X) - \sum_{g=1}^{G} c_g \mathbb{1}\left(X \in \mathcal{X}_g\right)\right)^2\right] \tag{C.1}$$

Write the first-order conditions:

$$\begin{aligned}
&\mathbb{E}\left[\left(\tau(X) - \sum_{t=1}^{G} c_t \mathbb{1}\left(X \in \mathcal{X}_t\right)\right) \mathbb{1}\left(X \in \mathcal{X}_g\right)\right] = 0, \quad g = 1, \ldots, G \\
\Longrightarrow\ &\mathbb{E}\left[\tau(X) \mid X \in \mathcal{X}_g\right] - c_g = 0, \quad g = 1, \ldots, G \\
\Longrightarrow\ &c_g^* = \mathbb{E}\left[\tau(X) \mid X \in \mathcal{X}_g\right] \equiv \tau_g, \quad g = 1, \ldots, G
\end{aligned}$$

Therefore, for any partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$ of the covariate space, the best MSE approximation is achieved by setting $c_g = c_g^*,\ g = 1, \ldots, G$.

In order to construct the optimal greedy partition $\mathcal{X}_1, \ldots, \mathcal{X}_G$, the algorithm starts with a region of the covariate space $\mathcal{R}_m \subseteq \mathcal{X}$ and iteratively stratifies the population minimizing the MSE within the resulting subregions:

$$\min_{j,s} \mathbb{E}\left[\left(\tau(X) - \tau_{m+1}\right)^2 | X \in \mathcal{R}_{m+1}(j,s)\right] + \mathbb{E}\left[\left(\tau(X) - \tau_{m+2}\right)^2 | X \in \mathcal{R}_{m+2}(j,s)\right] \tag{C.2}$$

which is equivalent to problem (3.2).

Appendix D Bounding the Number of Trees

Theorem D.1. *Define the "depth" $\mathcal{D}$ of a binary decision tree as the number of nodes connecting the root to the furthest leaf. Let X be a p-vector of binary covariates. Then, the number of distinct decision trees constructed by recursively partitioning $\mathcal{X}$ and having a depth equal to or lower than $\mathcal{D}$ is bounded from below by $L_{\mathcal{D}} = \prod_{d=1}^{\mathcal{D}} (p - (d-1))^{2^{d-1}}$.*

Proof. The proof is a matter of careful counting and relies on the fundamental theorem of counting. Define a *symmetric $\mathcal{D}$-depth tree* as any binary decision tree such that the number of nodes connecting the root to each leaf equals $\mathcal{D}$. The root is considered a 0-depth tree.

Start from the whole covariate space $\mathcal{X}$, i.e., from the unique 0-depth tree. Since all the p covariates are binary, there is a unique candidate splitting point s for each. Therefore, there exist p distinct candidate pairs (j, s) for the first split. It follows that it is possible to build p distinct symmetric 1-depth trees.

Now, fix a symmetric 1-depth tree, assuming without loss of generality that the split occurred on the first covariate. A symmetric 2-depth tree is then obtained by splitting both leaves of the nested symmetric 1-depth tree. As a split already occurred on the first covariate, there exist $p - 1$ distinct candidate pairs (j, s) for splitting each terminal node. Therefore, from a given symmetric 1-depth tree it is possible to build $(p-1)^2$ distinct symmetric 2-depth trees. By the fundamental theorem of counting, the number of distinct symmetric 2-depth trees equals $p\,(p-1)^2$.

By a similar argument, it is easy to count the number of distinct symmetric 3-depth trees that can be constructed from any symmetric 2-depth tree, which equals $(p-2)^4$. Again, from the fundamental theorem of counting it follows that the number of distinct symmetric 3-depth trees equals $p\,(p-1)^2\,(p-2)^4$.

Iterating the argument, we can write a closed-form expression of the number of symmetric $\mathcal{D}$-depth trees that can be constructed using p binary covariates:

$$L_{\mathcal{D}} = \prod_{d=1}^{\mathcal{D}} (p - (d-1))^{2^{d-1}} \tag{D.1}$$

Notice that any binary decision tree with a depth equal to or lower than $\mathcal{D}$ can be regarded as a subtree of a given symmetric $\mathcal{D}$-depth tree, that is, it can be obtained by collapsing a certain number of internal nodes of the latter. Therefore, the set of symmetric $\mathcal{D}$-depth trees is a subset of all the possible distinct binary decision trees that can be constructed by recursively partitioning $\mathcal{X}$ whose depth is at most $\mathcal{D}$. It follows that $L_{\mathcal{D}}$ is a lower bound for the number of such trees. □

Remarks. Equation (D.1) has a nice interpretation. Notice that a symmetric $\mathcal{D}$-depth tree is composed of $2^{\mathcal{D}}$ terminal nodes. Therefore, the formula reflects the fact that starting from any symmetric $(d-1)$-depth tree, 2^{d-1} leaves must be split to form a symmetric d-depth tree, and that $p-(d-1)$ candidate pairs (j,s) exist for each of these splits.

Notice also that we cannot grow symmetric trees with depth $\mathcal{D} > p$: in such cases, $L_{\mathcal{D}} = 0$. Moreover, $L_{p-1} = L_p$: starting from any symmetric $(p-1)$-depth tree, each leaf can be split choosing one and only one candidate pair (j,s), hence only one symmetric p-depth tree can be constructed for each of the distinct symmetric $(p-1)$-depth trees.

In the case of p categorical covariates with k categories each, Theorem D.1 holds if we substitute $p\,(k-1)$ for p.