Who Should the Taxman Visit? Evidence from Door-to-Door Tax Enforcement in Indonesia - Latest Version (link)[*]

Paulo Antonacci[†] Muhammad Khudadad Chattha [‡]

November 3, 2025

Abstract: In this paper we study the effects and optimal design of a large-scale tax-nudging campaign in Indonesia, where officials conducted door-to-door visits to more than 30,000 property owners.To estimate the program the program's average and heterogeneous impact on property-tax compliance we propose a modified causal forest estimator. We find that visits increased payment rates by an average of 7.8 p.p., with effects persisting for at least four subsequent years and being particularly pronounced among lower-value properties. We then use individualized treatment-effect estimates—rather than conventional compliance risk scores—to design targeting rules that answer "Who should the taxman visit?" The optimal targeting depends on a prioritization rule induced by policy objective:A participation-maximizing rule, which seeks to induce the largest number of taxpayers to comply, prioritizes lower-value properties and delivers meaningful gains in both participation and revenue. A revenue-maximizing rule directs visits toward higher-value properties, generating substantial fiscal returns with little change in the number of payers. By contrast, a standard risk-based rule that targets likely noncompliers is effectively neutral—allocating effort broadly across the value distribution. However it yields no discernible improvements in either revenue or participation. These results suggest that risk-based prioritization is ill-suited to nudging contexts and make explicit the equity–efficiency trade-offs inherent in algorithmic targeting for tax enforcement.

Keywords: Tax Compliance, Behavioral Nudges, Causal Forest, Policy Learning

JEL Codes: H21, H26, H71, H83, C99, D04, D91

[*] We are grateful for valuable comments from Peter Arcidiacono, Kate Bundorf, Michael Pollmann, Cem Özdemir and Naranggi Pramudya Soko. In memoriam, we thank Raka Rizky Fadilla for his exceptional research assistance, work ethic and kindness. We thank Carolyn Turk, Oleksii Balabushko, Alma Kanani, Rama Krishnan Venkateswaran and Daniel Ortega for their overall guidance. We are especially thankful to Nuryanto, Head of the Finance Agency of Kota Gorontalo, and Suprianto Kadir, Head of Revenue Field at the Finance Agency of Kota Gorontalo, for their support in implementing this project. All errors are our own. For correspondence, please contact: paulo.antonacci@duke.edu, mchattha@worldbank.org

[†] Duke University

[‡] World Bank

1 Introduction

The question “Who should the taxman visit?” goes to the heart of tax compliance strategy: given limited resources, authorities must decide where to direct enforcement efforts and how to encourage voluntary compliance. Tax agencies worldwide, along with international organizations like the OECD, IMF, and World Bank, have studied this issue extensively. Their recommendation is that auditing and other interventions should be risk-based and strategic – in other words, the taxman should “visit” those taxpayers most likely to be non-compliant or impactful for revenue (Khwaja et al., 2011). Recent work adds nudges and behavioral insights that can raise participation without heavy-handed audits. Yet when treatment effects are heterogeneous, three natural objectives—*targeting deliquency*, *maximizing revenue*, and *maximizing participation*—will generally induce different targeting priorities. Limited resources then make these objectives mutually constraining, transforming targeting into an explicit equity–optimality trade-off.

Improving compliance is especially challenging in developing economies, where tax administrations often operate with sparse data and rely on discretionary rather than data-driven decisions. In such contexts, individual bureaucrats may decide whom to audit, which households to visit, or which firms to inspect based on limited information or subjective judgment. As governments increasingly adopt e-government systems and digital recordkeeping, new opportunities arise for evidence-based and algorithmic decision-making in enforcement (Battaglini et al., 2025; Haseeb & Vyborny, 2022; Knebelmann et al., 2024; Bachas et al., 2025). Yet these developments also raise important questions about algorithmic fairness—that is, how data-driven targeting may affect equity and accountability in public administration. If designed carelessly, algorithmic tools might inadvertently reinforce social or spatial inequities, for instance by systematically focusing enforcement on poorer or more visible taxpayers (Black et al., 2022). Most existing applications optimize outcome prediction (e.g., risk or fraud detection) rather than the prioritization problem relevant for nudges. We instead study policy learning—selecting whom to visit to maximize expected treatment effects, which are unobserved -only estimated - at decision time(Manski, 2004; Tetenov, 2012; Yadlowsky et al., 2025). This reframes algorithmic assistance from predicting outcomes to allocating scarce enforcement toward units with the highest expected causal returns.

In this paper we investigate the door-to-door property-tax enforcement campaign in the city of Gorontalo, Indonesia (Map 1). Property tax is the city’s main locally generated revenue, and rising delinquency and payment delays prompted the initiative. Between November 2020 and December 2021, tax officers equipped with GPS-enabled tablets attempted visits to all 53,134 registered properties. When contact was made, the system logged geolocation and respondent identity in real time; those properties are coded as treated. Each address was visited at most once. In total, 31,730 properties were recorded as visited (59.9% coverage). Activity peaked in January 2021 with 8,776 visits and declined in March, before current-year bills became payable, as administrative priorities shifted.

We estimate the causal effect of door-to-door visits on property-tax compliance and develop an extension of the causal-forest/R-learner framework that accounts for pre-treatment confounding when estimating conditional average treatment effects (CATEs). The estimator employs Neyman-orthogonal scores and sample-splitting with cross-fitting to enable valid inference on heterogeneity (Nie & Wager, 2021; Wager & Athey, 2018; Chernozhukov, Lee, et al., 2025). By contrasting visited properties with comparable unvisited properties, we measure changes in payment timing and persistence across years. We link CATEs to taxpayer and property characteristics to characterize responsiveness to visits. We then study targeting under capacity constraints (Yadlowsky et al., 2025; Athey & Wager, 2021), comparing assignment rules tied to objectives in public financial management and tax-audit manuals and quantifying their performance.

The results from the intervention evaluation provide clear evidence that the door-to-door intervention significantly improved property tax compliance in the City of Gorontalo, Indonesia. After the full rollout of field visits in 2021, we calculate the ATE of 7.8% p.p., or 12.3% increase relative to an average baseline compliance of 63.4%. These findings accord with a large experimental literature on tax nudges Pomeranz and Vila-Belda (2019) and Antinyan and Asatryan (2024). Among the subset of studies on property taxation,[1] the average effect of reminders and related nudges is modest but consistently positive, typically in the range of 1–7 percentage points, with most estimates clustered around 3–4 percentage points. Delivery channel matters: face-to-face contact and phone calls outperform letters and emails (Ortega & Sanguinetti, 2013; Ortega & Scartascini, 2020; Hallsworth et al., 2017). The most comparable setting to ours is Weigel (2020), who study door-to-door visits and report large relative gains off a near-zero baseline (approximately 0.1% compliance). Our results, achieved against a far higher starting compliance rate, underscore the potency of in-person outreach even where baseline compliance is already substantial.

These effects remain large and statistically significant in every post-rollout year, tapering

[1] Brockmeyer et al. (2021) in Mexico; Castro and Scartascini (2015), Cruces et al. (2025), and Schächtele et al. (2023) in Argentina; Del Carpio (2013) in Peru; Pfeifer and Pacheco (2020) and Schächtele et al. (2022) in Brazil; Chirico et al. (2019) in the United States; John and Blume (2018) in the United Kingdom; Okunogbe (2021) in Liberia; Weigel (2020) in the Democratic Republic of Congo; Collin et al. (2025) in Tanzania; Manwaring and Regan (2023) in Uganda.

gradually to about 6.1 percentage points by year four. Intertemporal correlations of $\hat{\tau}$ across years indicate persistence. The pattern is consistent with behavioural change observed in the literature on income-tax audits. While property taxes leave no room for underreporting — assessments are externally determined—visits can recover arrears and shift beliefs about future enforcement: enforcement in one period affects compliance in later periods. A large literature in high-income settings documents sizable and long-lived audit effects: programs often more than pay for themselves (Boning et al., 2024); spillovers raise reported liabilities by roughly 55% of the audit adjustment in the following year (Kleven et al., 2011), with persistence up to five years in the U.K. (Advani et al., 2023) and six years in Norway even absent penalties (Hebous et al., 2023). We find comparably large but gradually attenuating effects over a five-year horizon in a developing-country property-tax context.

We also find that these gains were heterogeneous: smaller and lower-value properties, as well as taxpayers with moderate prior compliance, exhibited the largest behavioral responses, while wealthier or chronically non-compliant owners responded little. These heterogeneous impacts, stable across 2021–2024, indicate that enforcement effectiveness depends strongly on taxpayer characteristics. Taken together, our program evaluation results show that the door to door program can can meaningfully shift compliance while providing a fertile design optimal, data-driven targeting rules that allocate enforcement resources where they yield the highest returns.

We use heterogeneity in treatment effects and the properties of our estimator to investigate the impact of different treatment prioritization rules (Yadlowsky et al., 2025). Optimizing allocation with CATEs estimates reveals a trade-off between revenue and participation. When compared to the Average Treatment Effect: A participation-maximizing rule meaningfully increases both tax participation and revenue. A revenue-maximizing yielding substantial revenue gains but little change in the share of payers. By contrast, a conventional risk-based rule that targets likely non-compliers is close to neutral, delivering no discernible gains in either outcome.

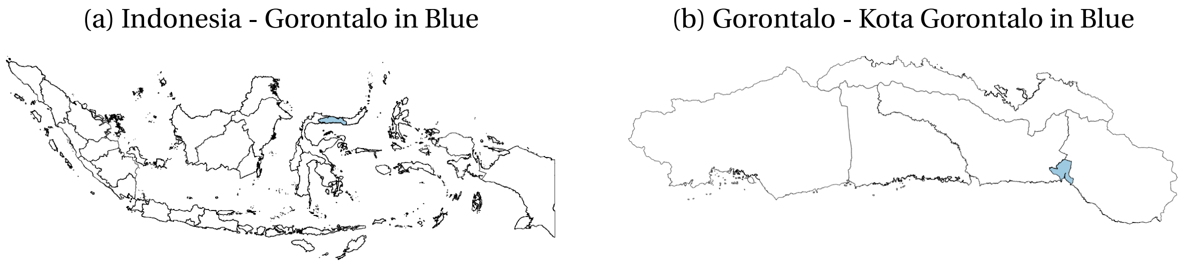
These choices have clear distributional consequences. To place policies on a common footing, we summarize performance with a single scalar targeting metric. A revenue-maximizing rule concentrates enforcement on high-value properties and nearly coincides with the most progressive benchmark—rank-ordering visits by property value. By contrast, a participation-maximizing rule concentrates enforcement on lower-value properties and closely mirrors the most regressive benchmark. The takeaway is simple: objectives are not interchangeable. Taken to the limit a tax system that is progressive in its design can become effectively regressive in practice if enforcement is optimized for the wrong objective.

The rest of the paper proceeds as follows. Section 2 details the institutional setting and the door-to-door visitation program. Section 4 outlines the empirical strategy for estimating treatment effects. Section 5 presents the main evaluation results, and Section 6 examines treatment prioritization and targeting outcomes. Section 7 concludes with implications for tax administration and algorithmic governance in developing-country contexts.

2 Empirical Setting

The door-to-door intervention took place in Indonesia, in the state of gorontalo, city of gorontalo – Kota Gorontalo Map 1. Kota Gorontalo is the capital and largest urban center of Gorontalo Province, situated on the northern coast of Sulawesi along the Gulf of Tomini. With a population of roughly 250,000 residents, it serves as the province's administrative, economic, and educational hub. The city is characterized by a predominantly service-oriented economy, complemented by trade, small-scale manufacturing, and fisheries. Its coastal location and road connections make it the main gateway to surrounding districts in Gorontalo and neighboring provinces.

Figure 1: Maps



(a) Indonesia - Gorontalo in Blue

(b) Gorontalo - Kota Gorontalo in Blue

Note: Left: Map of Indonesia with the province of Gorontalo highlighted in blue. Right: Enlarged view of Gorontalo Province showing the city of Gorontalo, also highlighted in blue.

As in many developing-country cities, local governments face constraints in revenue mobilization. Following the 2013 decentralization that transferred the Land and Building Tax (PBB-P2[2]) from the central government, own-source revenue now relies on this tax, which accounts for most local tax revenue.[3]

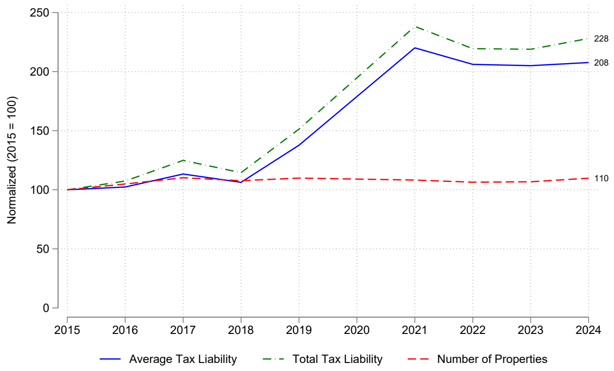
During the late 2010s, Gorontalo experienced a simultaneous rise in property tax revenues and an accumulation of tax arrears. Figures 2 and 3 trace the evolution of the city's property tax

[2] Pajak Bumi dan Bangunan Perdesaan dan Perkotaan

[3] See Section A for details.

base, liabilities, arrears, and collections from 2015 to 2024. The most substantial revenue gains followed a large-scale reassessment of property values between 2018 and 2021, which more than doubled the average tax bill. Over this period, the city's total tax potential expanded by 128 percent, driven by a 108 percent increase in average assessed property values and a 10 percent growth in the number of registered properties. However, assessed liabilities grew even faster than property values or registrations, intensifying collection pressures. Tax arrears surged by 197 percent, reflecting a 69 percent rise in the number of delinquent properties and a 75 percent increase in average arrears per delinquent taxpayer. On the collection side, nominal revenues rose by 113 percent, primarily due to higher average payments among compliant taxpayers—up 123 percent relative to 2015 levels. Yet the number of compliant taxpayers declined, revealing a widening gap between assessed liabilities and realized collections.

Figure 2: Growth in property tax base, 2015–2024

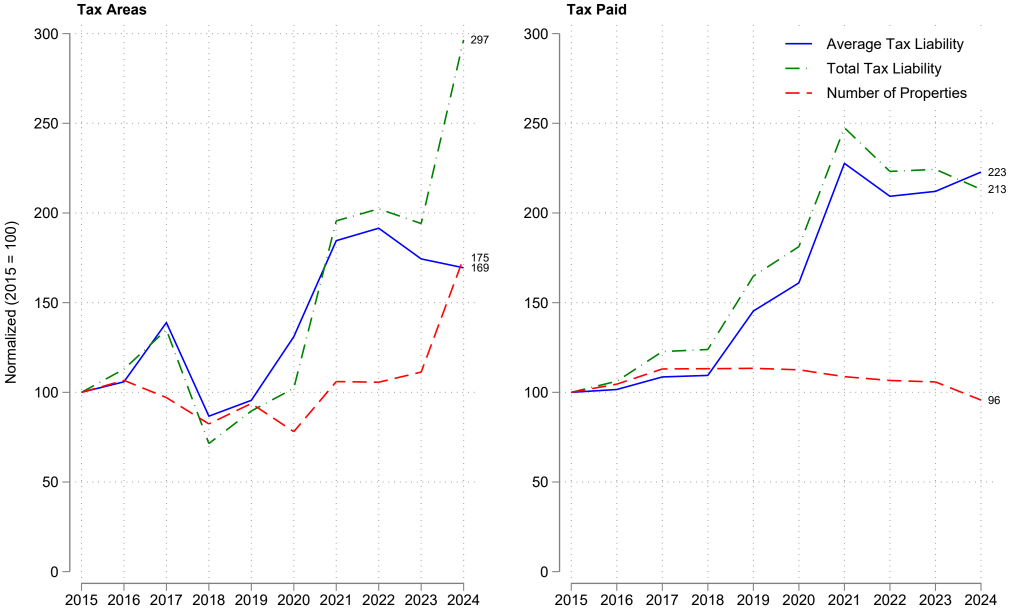


Note: The green line shows that total tax potential increased by 128%, driven primarily by reassessments between 2018 and 2021 that more than doubled the average tax bill. The blue line indicates that average property values rose by 108% over the same period, while the red line shows a 10% increase in the total number of registered properties. Includes all payments made until 03-02-2025. Values for 2020 are interpolated.

Late payment is pervasive and often occurs long after the statutory deadline. Figures **??** document the distribution of payment timeliness across all registered properties. On average, 60 percent of taxpayers paid before the deadline, 21 percent paid after it, and approximately 17 percent remained delinquent. Figure **??** further illustrates the cumulative timing of payments relative to the due date, showing that a substantial share of payments continues to arrive months after the deadline. Roughly 38 percent of all processed payments occur after the statutory deadline. Tax arrears are revenue not yet collected and their recovery timing varies, so they cannot support budget planning.

Taxpayers with arrears pay 1% per month up to 24 months and cannot transfer property until liabilities are paid. Authorities may seize property or collect through courts, but cost limits

Figure 3: Growth in tax debt and tax collection, 2015–2024.



Note: Left panel: The green line shows that total tax debt increased by 197%, largely driven by growth in the number of properties failing to pay their obligations. The blue line indicates that the average debt per delinquent property rose by 75% between 2015 and 2024. The red line reflects a 69% increase in the total number of registered properties that were delinquent. Right panel: Tax collection grew by 113% in nominal terms, driven mainly by reassessments conducted between 2018 and 2021 that more than doubled the average tax bill. The blue line shows a 123% increase in the average payment among compliant properties, while the red line highlights a decline in the number of compliant properties over time.Includes all payments made until 03-02-2025. Values for 2020 are interpolated.

use. After the cap, penalties stop and inflation further diminishes their real value, taxpayers face virtually no additional financial incentive to settle arrears. Then, enforcement rests on the transfer ban.

This institutional environment poses significant challenges for local public finance. Accumulated arrears constitute a latent revenue stream but one that is highly uncertain and unsuitable for routine budget planning. Consequently, the intervention pursued in Gorontalo had two primary objectives: (i) to increase the number of tax bills paid and (ii) to accelerate payment compliance prior to the deadline. These dual goals are central to defining the program's targeted outcome.

2.1. The Door-to-door Intervention

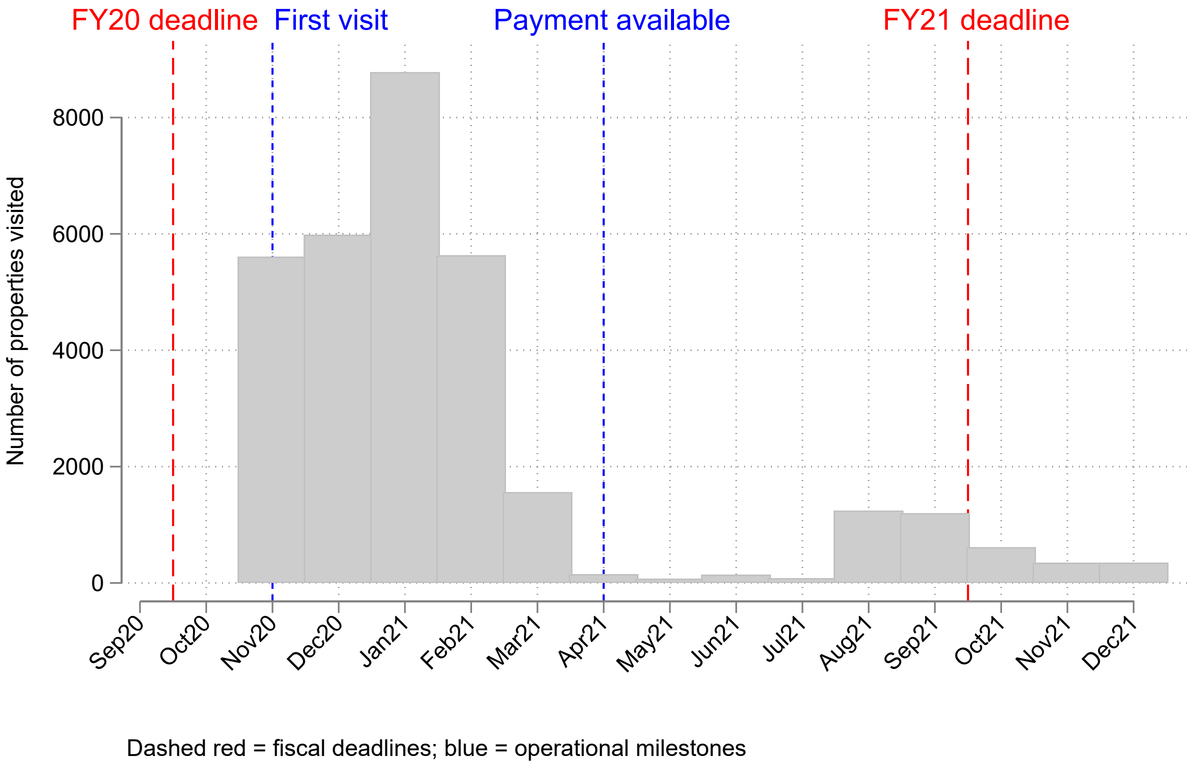
Between November 2020 and December 2021, the local tax administration in Gorontalo implemented a large-scale door-to-door enforcement campaign aimed at improving property tax compliance. Tax officials equipped with GPS-enabled tablets were instructed to visit all properties listed in the municipal tax directory and to record the outcome of each visit in real time. Upon arriving at a property, if no one was available, no record was generated. When an individual was present, the system automatically logged the official's geolocation along with the name of the person who answered the door and, when possible, the name of the individual legally responsible for the property. We define a property as *treated* only if it was successfully recorded during a visit; all remaining properties are classified as *untreated*. Each visit concluded either after contact was made with the property owner or after the official confirmed that the premises were unoccupied. Once a visit was completed, the official proceeded to the next property on the same street, ensuring that no household was visited twice.

In total, 31,730 properties were visited, representing approximately 59.9% of the 53,134 registered properties in the city. The campaign reached its peak intensity in January 2021, when 8,776 visits were completed, before declining sharply in March as administrative priorities shifted. Field activities resumed at a slower pace between August and December 2021, when the program was eventually phased out (Figure 4. All visits took place after the statutory payment deadline for the 2020 tax year, and only 478 visits occurred after the corresponding deadline for fiscal year 2021.

The intervention was not randomized. Field officers retained substantial discretion over the order and location of visits, creating the potential for selection bias in which properties were reached first. As shown in Figure 5, the share of visited properties varied widely across neighborhoods—from (31.1% to 95.4%)— while pre-intervention compliance before the deadline rates ranged between (35.8% and 99.6%). The strong spatial correlation between visit coverage and historical compliance suggests that selection into treatment was non-random, reflecting officers preference other rather than a systematic targeting rule.

Despite this limitation, the administrative data generated by the program —linking geolocated visits to property-level tax payment histories —constitute an exceptionally rich observational dataset for studying compliance behavior. In the next section, we will describe our data.

Figure 4: Monthly Visits and Key Operational Milestones

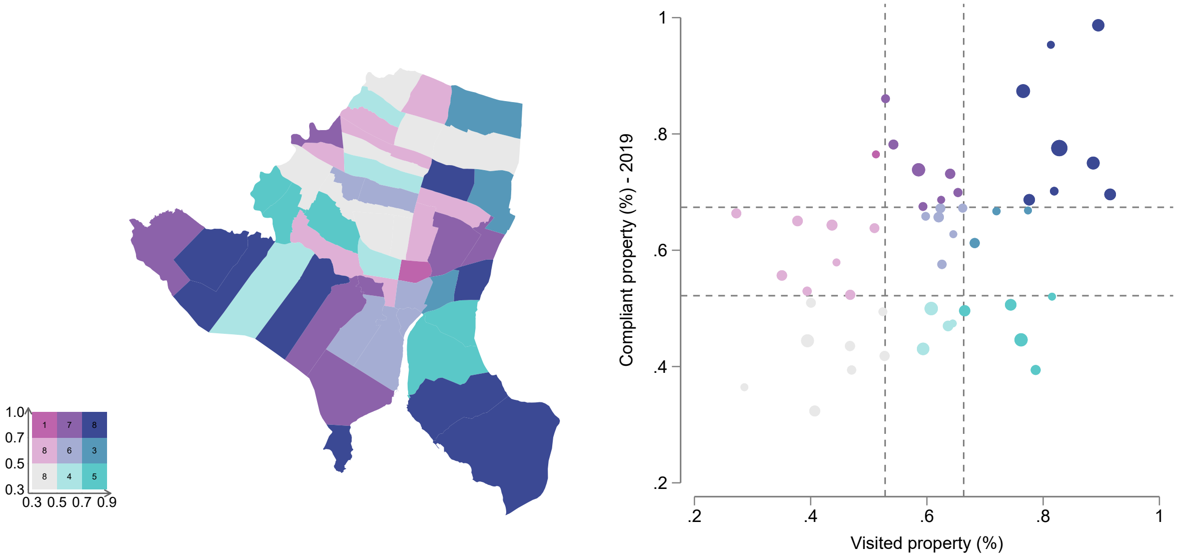


Note: The gray bar indicates the number of properties visited each month during the door-to-door intervention. Dashed red vertical lines mark fiscal year payment deadlines (FY2020 and FY2021), while blue dashed lines indicate operational milestones—the start of field visits and the date when payment processing became available.

3 Data

This section describes the administrative data underlying our analysis and outlines how we constructed the property–year panel used to estimate treatment effects. The data originate from the City of Gorontalo's property tax administration system and cover more than a decade of tax records. We first document the data sources and their coverage, followed by a discussion of quality issues and restrictions applied to ensure comparability across years. We then detail the covariates used in the empirical analysis and conclude with descriptive evidence on treatment imbalance between visited and non-visited properties.

Figure 5: Pre-Treatment Compliance and Treatment Intensity Across Neighborhoods



Note: The left panel presents a bivariate map relating the share of properties that paid the property tax before the 2019 deadline to the share of properties visited during the intervention, across 50 neighborhoods. Neighborhoods in the northern area exhibit both lower compliance and lower visitation rates. The right panel displays the corresponding scatterplot, where each point represents a neighborhood and its size is proportional to the number of properties. Vertical and horizontal dashed lines indicate the 33rd and 67th percentiles along each axis. All available observations are included. See Figure 14 for the same analysis restricted to the balanced sample.

3.1. Data Sources

Our analysis draws on a comprehensive set of administrative records from the City of Gorontalo, Indonesia, spanning the years 2013–2024. The underlying property registry contains 61,778 unique parcel identifiers, with annual totals ranging between 48,076 and 53,953, reflecting parcel subdivisions, consolidations, and new developments over time. The resulting panel is unbalanced. For analytical consistency, we restrict attention to 41,451 properties observed in all relevant years—approximately 77% of those registered in 2021, the year when the door-to-door intervention was implemented.

To measure tax compliance and treatment effects, we construct a property–year panel that integrates three complementary administrative datasets:

(1) Payment Records. The payment database contains for each year detailed information on annual property tax liabilities and payments, including assessed tax bases, nominal tax rates, discounts, fines, payment amounts, and payment dates. These variables allow us to construct precise indicators of compliance—namely, whether a property paid its full tax bill and whether payment occurred before the statutory deadline.

(2) Property Characteristics. This dataset includes land and building attributes—such as land area, building area, construction type, and assessed market values per square meter—as well as ownership characteristics, including the number of properties held by each taxpayer, whether the owner resides in Gorontalo, and whether the registered address matches the property's location. These variables are updated annually by the municipal tax office and represent the main structural determinants of both tax liability and compliance behavior. We were not provided with information on property use (residential vs. commercial), which limits our ability to differentiate behavioral responses by use type.

(3) Visitation Records. The visitation dataset documents all door-to-door enforcement activities undertaken by municipal tax officers. Each record includes the date, assigned officer, visit outcome, and GPS coordinates automatically logged through tablets during fieldwork. These data enable precise identification of treated (visited) and untreated (non-visited) properties, as well as verification of visit authenticity and spatial coverage.

3.2. Data Coverage and Quality Considerations

The property tax remains Gorontalo's largest source of local tax revenue. Although the administrative system dates back to this decentralization reform in 2013, data quality in the earliest year is inconsistent: several key variables are missing, miscoded, or recorded using nonstandard formats, particularly for property values and payments. These problems are observed to a lower extend in 2014, and from 2015 onwards are no longer present.

The year 2020 is also atypical due to the *COVID-19 fiscal stimulus*. The city government introduced substantial tax relief measures: taxpayers were required to pay only 23% of their PBB-P2 liability (a 77% discount), recorded in the system as "stimulus" or "correction" transactions. These entries complicate the interpretation of compliance indicators because they blur the distinction between partial and full payments. In addition, discrepancies arose between the *Payment Records* and *Property Characteristics* datasets, likely due to disruptions in data entry and collection processes during the pandemic.

From 2021 onward, the stimulus policy was retained in a reduced and standardized form: non-commercial properties with a taxable value (*NJOP*) below IDR 500 million received a 10% discount, while those above the threshold received 15%. These adjustments were clearly coded in the payment system and did not generate further inconsistencies.

To ensure comparability and reliability across years, the main analysis is restricted to the 2015–2024 period, excluding 2020.

3.3. Variables

Table 8 displays a comprehensive description of all variables we used in this paper. We organize the covariates used in the analysis into four main groups: (1) property characteristics, (2) Location (3) owner characteristics, and (4) historical payment behavior. All of the covariates were colected pre-treatment[4].

Property characteristics: These include features related to the assessed value and physical attributes of each parcel: (P1) Value of the tax object (*Nilai Objek Pajak*);(P2) Land value per square meter;(P3) Building (construction) value per square meter; (P4) building area; (P5) Indicator for presence of construction; (P6) Applicable tax rate; (P7) Land area.

Location: We include geographic identifiers at multiple levels of aggregation: (L1) district, (L2) subdistrict, and (L3) block capture local variation in enforcement practices and economic activity.

Owner characteristics: Properties are linked to owners through unique identifiers, allowing us to construct: (O1) Indicator for multiple-property ownership in 2021; (O2) Indicator for owner occupancy (owner lives on the property) in 2021; (O3) Indicator for owner residing within Gorontalo City in 2021; (O4) Indicator for single-property owners in 2021

Historical payments: Past compliance is one of the strongest predictors of future behavior and a central component of our analysis. For each property, we compute: (H1) Number of payments made before the annual deadline (2015–2019); (H2) Number of late payments before the start of the visitation program at 19 November 2020. (Y1) Year-specific dummies for payment before the deadline (Y2) Year-specific dummies for payment after the deadline; (Y3) Log of number of days paid before the deadline (if timely, zero otherwise) (Y4) Log of number days

[4] Some variables used as covariates come from the dataset (2) Property Characteristics of 2021. They correspond to the Fiscal Years 2021, therefore were defined in 2020, before the intervention started.

paid after the deadline (if delayed, zero otherwise)

We have a large number of control variables. In principle, there are several ways to handle this dimensionality. One could select a small subset of meaningful variables, which facilitates interpretation but risks omitting important predictors. Alternatively, one could include all available variables and rely on the estimation method to select relevant controls, though this approach may miss complex interactions and nonlinearities. Lastly, one could opt for a more flexible strategy involves including higher-order terms and interactions among the covariates, which captures richer relationships at the expense of some interpretability[5].

In this paper, we adopt an intermediate approach. We include all baseline covariates along with their first differences (relative to 2015) for time-varying attributes. While land area remains constant, other characteristics—such as assessed values and building size—may evolve over time. For example, the log of the 2021 tax base is accompanied by its change relative to 2015, capturing both level and trend information. This specification balances flexibility with parsimony and allows us to account for gradual adjustments in the property registry nonlinearity due to random forest.

Noitice that in the next section we will estimating pre-intervention confounding bias for 2018 and 2019, we further adjust the covariate set to exclude any variables that could directly or indirectly depend on post-treatment outcomes. For instance, when evaluating potential confounding in 2018, the variable Number of payments made *before* the annual deadline (2015–2019); is replaced by its truncated version computed only from 2015 through 2017.

4 Empirical Strategy

This section describes our empirical strategy. We begin by introducing the potential outcomes framework and notation. We then present the four assumptions required to identify conditional average treatment effects (CATEs). Finally, we describe the causal forest estimator and our bias-correction procedure. The analysis follows the Neyman–Rubin causal model Imbens and Rubin (2015).

Let each observational unit i denote a property. The binary treatment indicator is $Z_i \in \{0,1\}$, where $Z_i = 1$ indicates that property i received a door-to-door visit (behavioral nudge), and $Z_i = 0$ otherwise. Each unit has two potential outcomes: $Y_i(1)$ if treated and $Y_i(0)$ if untreated.

[5] See Belloni et al. (2014)

The observed outcome is

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0).$$

Our primary outcome measures compliance, defined as an indicator equal to one if the property paid its tax obligations before the deadline and zero otherwise:

$$Y_i^t = \begin{cases} 1, & \text{if the property paid before the deadline,} \\ 0, & \text{if delinquent.} \end{cases}$$

The Average Treatment Effect (ATE) is defined as

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)],$$

which captures the average increase in the probability of timely payment if all properties were visited relative to none. Equivalently,

$$\tau = \mathbb{P}\big[Y_i(\text{Visited}) = \text{Pay}\big] - \mathbb{P}\big[Y_i(\text{Not Visited}) = \text{Pay}\big],$$

interpretable as the overall probability of behavioral change induced by the visit. We are also interested in treatment heterogeneity across observable characteristics $X_i \in \mathscr{X} \subseteq \mathbb{R}^p$. The Conditional Average Treatment Effect (CATE) is

$$\tau(X_i) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i\big] = \mathbb{P}\big[\text{Behavioral change} \mid X_i\big].$$

To identify $\tau(X_i)$ from observational data, we require four standard assumptions: Unconfoundedness, Overlap, Stable Unit Treatment Value Assumption (SUTVA)

1. Unconfoundedness (Selection on Observables): We assume that, conditional on observable characteristics X_i, treatment assignment is as good as random:

$$(Y_i(0), Y_i(1)) \perp Z_i \mid X_i.$$

This assumption is inherently untestable, since only one potential outcome is observed for each unit. Nevertheless, its plausibility can be indirectly evaluated using pre-treatment (placebo) outcomes. If unobserved confounders drove treatment assignment, we would observe spurious "effects" even before the intervention. Specifically, any estimated treatment effect in pre-treatment years (2018–2020) would indicate bias. In our data, treatment assignment shows small yet statistically significant predictive power for outcomes in these years. We propose a pre-estimation debiasing to correct for the confounding bias.

2. Strict Overlap (Positivity): For all values of X_i, the probability of treatment is strictly between zero and one:

$$0 < \mathbb{P}(Z_i = 1 \mid X_i) < 1.$$

This assumption ensures sufficient common support between treated and untreated units. It can be empirically verified using the estimated propensity score[6].

3. Stable Unit Treatment Value Assumption (SUTVA): Potential outcomes for unit i depend only on its own treatment status Z_i and not on the treatment status of others:

$$Y_i(z) \text{ is invariant to } Z_j,\ \forall\, j \neq i.$$

In practice, spillovers may arise if untreated properties are influenced by nearby treated neighbors. Evidence from the tax nudging literature suggests that such spillovers are typically positive, meaning untreated properties might also improve compliance. In this case, our estimates would be conservative, understating the true effect of the intervention.

4.1. The causal forest approach to estimating CATE

Integrating machine learning with causal inference enable flexible and credible estimation of conditional average treatment effects (CATE). Approaches span Bayesian nonparametrics (Hill, 2011; Hahn et al., 2020), meta-learners that recast CATE estimation as supervised learning task (Künzel et al., 2019), representation learning for counterfactual prediction (Shalit et al., 2017), forest-based methods tailored to heterogeneity (Wager & Athey, 2018), and learned-distance matching (Parikh et al., 2022). See Caron et al. (2022) for a recent review.

In this paper we rely on the Causal Forest estimator of Wager and Athey (2018), which extends the random forest framework to the estimation of heterogeneous treatment effects. Formally, the object of interest is the Conditional Average Treatment Effect (CATE),

$$\tau(X_i) := \mathbb{E}[\, Y_i(1) - Y_i(0) \mid X_i\,],$$

where $Y_i(1)$ and $Y_i(0)$ denote potential outcomes under treatment and control, respectively, and $X_i \in \mathscr{X} \subseteq \mathbb{R}^p$ denotes the covariate vector.

[6] For further discussion see section 5.1

The causal forest can be viewed as a nonparametric implementation of the *R-learner* of Nie and Wager (2021), which reformulates treatment effect estimation as a residual-on-residual regression problem based on the orthogonal decomposition of Robinson (1988). Let $W_i = (Y_i, Z_i, X_i)$ denote the observed data, where $Z_i \in \{0,1\}$ indicates treatment assignment. The structural model can be written as

$$Y_i = \mu(X_i) + \tau(X_i) Z_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X_i, Z_i] = 0, \tag{1}$$

where $\mu(X_i) = \mathbb{E}[Y_i \mid X_i, Z_i = 0]$ denotes the baseline outcome function. Let $m(X_i) := \mathbb{E}[Y_i \mid X_i]$ be the unconditional outcome regression, and $e(X_i) := \mathbb{E}[Z_i \mid X_i]$ the propensity score. By iterated expectations,

$$Y_i - m(X_i) = \big(Z_i - e(X_i)\big)\tau(X_i) + \varepsilon_i, \tag{2}$$

which expresses the treatment effect as the slope in a regression of residualized outcomes on residualized treatments. Equation (2) is the foundation of the *R-learner*:

$$\hat{\tau}(\cdot) = \arg\min_{\tau(\cdot)} \frac{1}{n} \sum_{i=1}^{n} \Big[\big(Y_i - \hat{m}^{(-i)}(X_i)\big) - \big(Z_i - \hat{e}^{(-i)}(X_i)\big)\tau(X_i) \Big]^2, \tag{3}$$

where $(\hat{m}^{(-i)}, \hat{e}^{(-i)})$ are out-of-fold predictions obtained from an auxiliary sample (*cross-fitting*) to ensure that the residuals are approximately orthogonal to estimation errors in the nuisance functions. Under mild regularity conditions, this orthogonalization property yields $\sqrt{n}$-consistency and asymptotic normality even when (m, e) are estimated via flexible machine-learning methods Chernozhukov et al. (2018) and Semenova and Chernozhukov (2020).

The causal forest operationalizes the objective in (3) via a recursive partitioning scheme that adaptively estimates heterogeneous slopes $\tau(X)$ in a nonparametric, data-driven way. It belongs to the class of Generalized Random Forests (GRFs) introduced by Athey and Wager (2019), where the forest weights $\alpha_i(x)$ implicitly define a local neighborhood around each evaluation point x:

$$\hat{\tau}(x) = \sum_{i=1}^{n} \alpha_i(x)\,\Gamma_i, \qquad \Gamma_i := \hat{\tau}^{(-i)}(X_i) + \frac{Z_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)[1 - \hat{e}^{(-i)}(X_i)]} \Big[Y_i - \hat{\mu}^{(-i)}(X_i, Z_i) \Big], \tag{4}$$

where Γ_i denotes the orthogonalized pseudo-outcome. Each tree in the forest is grown "honestly": one subsample determines the splitting structure, and a disjoint subsample estimates treatment effects within each leaf. This honesty constraint ensures asymptotic unbiasedness, consistency, and valid inference for $\tau(X)$ Wager and Athey (2018) and Athey and Imbens (2016).

In our application, each treatment effect $\tau(X_i)$ can be expressed in monetary terms by multiplying it by the tax liability W_i associated with property i. We do it directly at the score function.

Causal forests are thus particularly well-suited for rich administrative datasets such as ours, combining (i) robustness to high-dimensional confounding through orthogonalization, (ii) flexibility to approximate nonlinear heterogeneity, and (iii) sample-splitting to avoid overfitting. All estimators in this paper follow the cross-fitted and honest GRF implementation of Athey and Wager (2019), as implemented in the `grf` package for R.

4.2. Debiasing Procedure and Bias-Corrected Outcomes

A key challenge in our observational setting is residual confounding: even after controlling for observed X_i, there may be systematic outcome differences between treated and control units due to non-random selection into the program. To address this, we employ a bias-correction strategy using pre-intervention data. The idea is to use the year before the program started (2018 in our case) to measure any pre-existing outcome differences associated with treatment status, and then adjust our post-intervention outcomes accordingly. Since the program had not begun in 2018, any difference in outcomes between units that eventually received the treatment ($Z_i = 1$) and those that did not ($Z_i = 0$) in that year must be due to selection bias rather than a true treatment effect. We can thus estimate a bias function $B(x)$ that captures this spurious effect of Z on outcomes as a function of covariates. Formally, let $Y_i^{(t)}$ denote the outcome for unit i in year t and $X_i^{(17)}$ the vector of baseline characteristics (including covariates and past outcomes) up to 2017. We posit the following model for the pre-intervention year 2018:

Here $f(X_i)$ represents the baseline outcome in 2018 as a function of X_i (analogous to $\mu(X_i)$ for the control group), and $B\left(X_i^{(17)}\right)$ represents the bias term: the systematic difference in outcomes in 2018 between treated and untreated units with the same covariates. By definition, there was no actual treatment effect in 2018 (the program had not started), so $\tau^{(18)}(x) = 0$ for all x and any correlation between Z_i and $Y_i^{(18)}$ is purely due to latent differences. The function $B(x)$ captures those differences. We estimate $B(x)$ using a causal forest on the 2018 data, treating Z_i as the "pseudo-treatment" and $Y^{(18)}$ as the outcome. Specifically, we estimate the conditional mean outcomes by treatment status and define:

To account for residual confounding arising from pre-intervention selection into treatment, we estimate and remove a bias component $B(X)$ measured using pre-treatment data. Let $X_i^{(17)}$ denote the vector of baseline covariates and historical payment behavior up to 2017. We postulate the following reduced-form model for the pre-intervention period (2018):

$$Y_i^{(18)} = \big(\underbrace{\tau^{(18)}(X_i)}_{:=0} + B(X_i^{(17)})\big) Z_i + f(X_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X_i, Z_i] = 0, \tag{5}$$

where $\tau^{(18)}(X_i)$ represents the (zero) treatment effect before the program, and $B(X_i^{(17)})$ captures systematic bias arising from correlation between Z_i and potential outcomes in the pre-period. Since no actual intervention occurred in 2018, we have $\tau^{(18)}(X_i) = 0$, and therefore the CATE estimator from that year identifies the bias function:

$$\hat{B}(X_i^{(17)}) = \hat{\mu}^{(-i)}(X_i^{(17)}, 1) - \hat{\mu}^{(-i)}(X_i^{(17)}, 0),$$

where $\hat{\mu}^{(-i)}(\cdot,\cdot)$ are out-of-fold outcome regressions estimated via a causal forest using only pre-treatment data. We assume that this bias function remains stable over time, i.e. $B(X_i^{(17)}) = B(X_i^{(t)})$ for later periods $t \geq 2019$. For post-intervention years $t \in \{2019, 2021, 2022, 2023, 2024\}$, we define the bias-corrected (debiased) outcome as

$$\tilde{Y}_i^{(t)} := Y_i^{(t)} - \hat{B}(X_i^{(17)}) Z_i. \tag{6}$$

Substituting into (1) yields the bias-adjusted model

$$\tilde{Y}_i^{(t)} = \tau^{(t)}(X_i) Z_i + f(X_i) + \tilde{\varepsilon}_i, \qquad \mathbb{E}[\tilde{\varepsilon}_i \mid X_i, Z_i] = 0, \tag{7}$$

where $\tilde{\varepsilon}_i := \varepsilon_i - (B(X_i^{(17)}) - \hat{B}(X_i^{(17)})) Z_i$. Equation (7) now satisfies strict unconfoundedness by construction, allowing causal forest estimation to proceed as in the standard R-learner setting. We then estimate the period-specific CATEs via cross-fitted causal forests:

$$\hat{\tau}^{(t)}(X_i) = \mathrm{CausalForest}\{(\tilde{Y}_i^{(t)}, Z_i, X_i)\}.$$

As a consistency check, we expect that in placebo years (before the treatment actually took effect) the causal forest should find no significant effect. Indeed, using 2019 as a placebo test (the program had not yet been rolled out in 2019), our procedure yields $\hat{\tau}^{(2019)}(x) \approx 0$ for all x. By contrast, for the years after the intervention starts (2021 onward), we find $\hat{\tau}^{(t)}(x) > 0$ for many x, reflecting the program's positive impact. This aligns with the intervention timeline and increases our confidence that the bias removal was effective. All estimated CATEs $\hat{\tau}^{(t)}(x)$ are obtained via the described bias-corrected causal forest procedure. Thanks to the orthogonalization (residualization) and honest sample splitting, these estimates are asymptotically unbiased for the true heterogeneous effects and come with valid confidence intervals Wager and Athey (2018). For further technical details on the orthogonal estimating procedure and its statistical properties, see the Appendix.

5 Treatment Effects

The empirical analysis proceeds in three steps. We first conduct an exploratory assessment of the data and estimate the propensity score—the probability that a property was visited given its pre-treatment characteristics—using a boosted regression forest. The resulting distribution shows broad common support between treated and control units, indicating that treatment assignment was plausibly stochastic conditional on observables.

Next, we use these estimated propensity scores to identify the causal effects of the door-to-door intervention and their heterogeneity. Average treatment effects (ATEs) before and after the campaign are estimated with the Augmented Inverse Propensity Weighting (AIPW) method combined with causal forests, which corrects for selection bias and model misspecification. Finally, we analyze treatment effect heterogeneity using Best Linear Projection (BLP) estimates to determine which types of taxpayers responded most to the intervention.These are the subsidies to the treatmetn priorization analisys in next Section 5.3.

5.1. Propensity Score Estimation

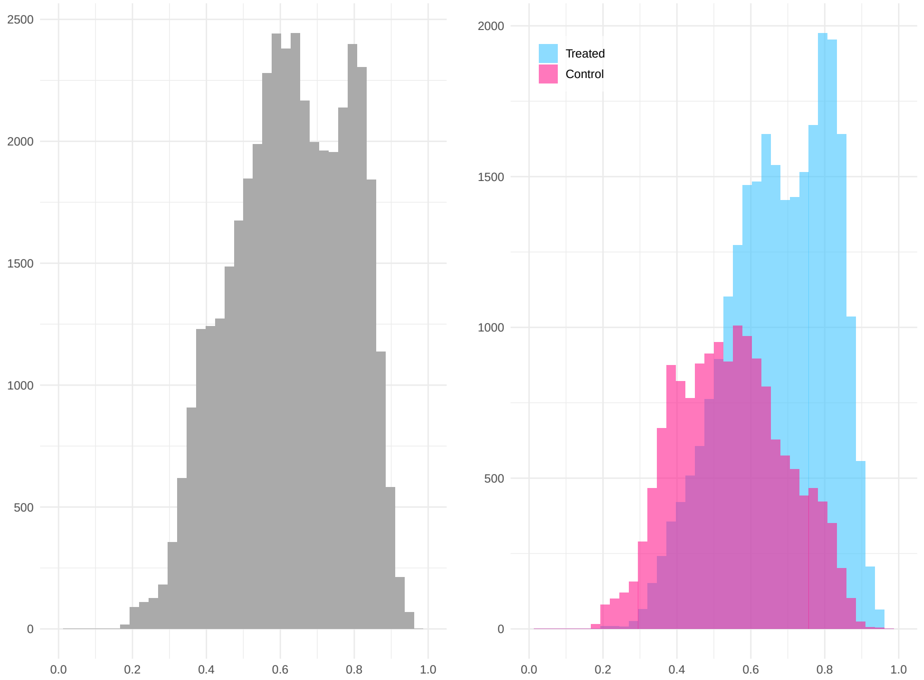
Our cleaned dataset[7] comprises 41,448 property-level observations spanning 2015–2024. Among these, 26,036 properties (approximately 62%) were visited as part of the door-to-door enforcement campaign. Visited and unvisited properties differ systematically across several pre-treatment dimensions: (1) physical property characteristics, (2) location, (3) owner attributes, and (4) historical payment behavior. A detailed description of all variables is provided in Table 8.

On average, treated (visited) properties exhibit a lower assessed tax base, largely reflecting smaller land areas and lower land values per square meter. Nevertheless, these properties tend to be more intensively used, with a higher prevalence of both new and old constructions, larger total building areas, and higher average building values per square meter.

In terms of ownership characteristics, treated properties are more likely to be owner-occupied, more likely to be located within Gorontalo city, and less likely to belong to individuals who own multiple properties. Finally, across all measures, visited properties display higher baseline tax compliance. For instance, when focusing on timely payments (before the statutory deadline), the treated group shows between 12.7% and 16.9% higher compliance rates relative to the control group in the pre-treatment years 2018–2024.

[7] See Appendix **??** for details on the data cleaning process.

Figure 6: Propensity Score Estimates



Note: The figure shows the distribution of estimated propensity scores overall (left) and by treatment status (right, blue = visited, pink = not visited). Scores are concentrated between 0.3 and 0.9, with minimal mass in the tails, indicating strong overlap and adequate common support for causal-forest estimation.

We control for an extensive set of pre-treatment covariates covering property, owner, and locational attributes as well as past payment history. Conditioning on these variables is consistent with the literature on tax compliance and default prediction, and—under the standard stric unconfoundedness assumption—suffices to account for differences in treatment assignment between visited and non-visited properties. We further argue that we have strong indications that strict unconfoundedness assumptions is adequate.

First, we further argue that the strict overlap assumption holds in our data, ensuring that treatment assignment is probabilistic rather than deterministic for all units. We estimate the propensity score $\hat{e}(X_i) = P(Z_i = 1 \mid X_i)$ using the boosted regression forest algorithm implemented in the `grf` package Tibshirani et al. (2025). Figure 6 and Table 1 illustrate the resulting distribution of estimated propensity scores. The overlap condition appears well satisfied: estimated probabilities are bounded away from 0 and 1 for both treated and control groups, with the overall range spanning from 0.174 to 0.961. No observation has an estimated probability below 0.05, and only ten units (0.02%) have scores above 0.95.

The 5th and 95th percentiles for treated (0.424–0.871) and control (0.322–0.802) units largely overlap, suggesting that for most values of X_i, both treated and untreated observations exhibit similar treatment propensities. The mean propensity scores of the treated and control groups (0.677 and 0.547, respectively) are close and both well within the interior of the unit interval. These features indicate a broad region of common support and rule out regions of perfect prediction, supporting the plausibility of the overlap assumption necessary for causal identification.

Table 1: Descriptive Statistics of Estimated Propensity Scores

Group	N	% Total	Min	P(5)	P(50)	P(95)	Max	Mean	SD	<5%	>95%
Overall	41,448	100.00	0.174	0.363	0.633	0.858	0.961	0.628	0.156	0	10 (0.02%)
Treated	26,036	62.82	0.182	0.424	0.687	0.871	0.961	0.677	0.140	0	10 (0.04%)
Control	15,412	37.18	0.174	0.322	0.544	0.802	0.950	0.547	0.148	0	0.00%

Notes: Propensity scores $\hat{e}_i = P(Z_i{=}1 \mid X_i)$ are bounded in $[0,1]$. P(5), P(50), and P(95) denote the 5th, 50th (median), and 95th percentiles. “ <5%” and “ >95%” are number and the shares the percentages of observations with $\hat{p}_i < 0.05$ and $\hat{p}_i > 0.95$, respectively. Percentages rounded to two decimals; probabilities to three decimals.

5.2. Treatment Effect Estimates

Table 2 reports the estimated treatment effects on property-tax compliance before and after the implementation of the door-to-door intervention. The estimates are presented separately for the pre-intervention period (2018–2019) and for the post-intervention period (2021–2024). The year 2020 is excluded from the analysis due to the exceptional disruptions caused by the COVID-19 pandemic, which temporarily altered payment rules and enforcement procedures (see Section 3.2).

We present our main results using the debiased estimation methods proposed in the previous section. In addition, as a robustness check, all results from this section onward are also replicated using causal forests estimated without the debiasing correction and restricting covariates to the pre-treatment period (2015–2018). The findings remain consistent across specifications, and equivalent tables are reported in Appendix K.

Before the intervention, treatment effects were negligible. In 2019, the estimated ATE was only 0.004 and statistically indistinguishable from zero, indicating that the model detected no systematic relationship between the treatment indicator and tax compliance prior to the program rollout. This absence of effect provides a useful falsification check: if confounding or model-induced bias were driving the results, we would expect to observe similar magnitudes of treatment effects in pre-intervention years as well. The fact that the estimated effects are essentially null reinforces that the post-2021 treatment effects capture genuine behavioral responses rather than spurious correlations.

After the full roll-out, treatment effects increase sharply and remain persistently large. The ATE rises to 7.8% in 2021 . It remains remains highly significant in all subsequent years, at 7.3% in 2022, 6.4% in 2023, and 5.1% in 2024. When compares to the baseline there has been an increase between 12.3% in 2021 and 7.4% in 2024. Treatment and control groups present different results. In all years, the treatment effects on the visited unities is superior to the estimates of treatment effects among non-visited. The difference is statistically significant but the magnitude is small, less than 10% the ATE.

The persistence of large and statistically significant effects over multiple years demonstrates that the door-to-door intervention produced durable improvements in tax compliance. Moreover, the systematic heterogeneity in treatment responses underscores the value of forest-based methods for identifying high-return segments of the taxpayer population. These results collectively support the conclusion that personalized or targeted enforcement strategies-guided by machine-learning predictions of responsiveness-can enhance the efficiency and fairness of local tax administration while maintaining sustained compliance gains over time.

Table 2 Panel B examines treatment heterogeneity. Taxpayers are divided into High-$\hat{\tau}$ and Low-$\hat{\tau}$ groups according to their predicted responsiveness. A property is classified as High-$\hat{\tau}$ if its predicted treatment effect is above the sample median. The differences between these groups are substantial. In 2021, the estimated effect among high-$\hat{\tau}$ taxpayers is 13.1% (s.e. 0.007, $p < 0.01$), compared with 0.025 (s.e. 0.006, $p < 0.01$) among low-$\hat{\tau}$ taxpayers, yielding a gap of 0.106 percentage points (s.e. 0.009, $p < 0.01$). This pattern persists across all post-intervention years, with high–low differences ranging between 6 and 11 percentage points.

The persistence of large and statistically significant effects over multiple years demonstrates that the door-to-door intervention produced durable improvements in tax compliance. Moreover, the systematic heterogeneity in treatment responses underscores the value of forest-based methods for identifying high-return segments of the taxpayer population. These results collectively support the conclusion that personalized or targeted enforcement strategies—guided by machine-learning predictions of responsiveness—can enhance the efficiency and fairness of local tax administration while maintaining sustained compliance gains over time.

The persistence of large and statistically significant effects over multiple years demonstrates that the door-to-door intervention generated durable improvements in tax compliance. The systematic heterogeneity in treatment responses further highlights the value of forest-based methods for identifying high-return segments of the taxpayer population. Together, these find-

Table 2: Treatment Effect Estimates Before and After Intervention — Debiased

	Placebo	**Treatment Effects**			
	2019	2021	2022	2023	2024
Panel A: Average Treatment Effects (AIPW Estimates)					
Average (ATE)	0.004	0.078^{***}	0.073^{***}	0.064^{***}	0.051^{***}
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Treated (ATT)	0.003	0.080^{***}	0.072^{***}	0.068^{***}	0.051^{***}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Untreated (ATU)	0.003	0.074^{***}	0.069^{***}	0.056^{***}	0.048^{***}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Outcome Mean:					
Average	0.647	0.634	0.663	0.691	0.682
Treated	0.697	0.693	0.726	0.742	0.729
Untreated	0.563	0.533	0.557	0.605	0.602
Panel B: Heterogeneous Effects by Predicted $\hat{\tau}$ (High vs. Low)					
High (τ_{High})	0.023^{***}	0.131^{***}	0.118^{***}	0.108^{***}	0.081^{***}
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Low (τ_{Low})	-0.015	0.025^{***}	0.025^{***}	0.019^{***}	0.020^{***}
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Difference (High–Low)	0.038^{**}	0.106^{***}	0.093^{***}	0.088^{***}	0.061^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Outcome Mean:					
High	0.635	0.618	0.629	0.662	0.660
Low	0.635	0.638	0.687	0.709	0.692

Notes: All treatment effects are estimated using out-of-bag, forest-based Augmented Inverse Probability Weighting. Year 2020 is excluded (see Section 3.2). Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. See Table **??** for the same table under alternative specifications.

ings support the idea that personalized or targeted enforcement strategies, guided by machine-learning predictions of taxpayer responsiveness, can be used by tax authorities. In the next section, we investigate the underlying characteristics driving this heterogeneity in treatment effects, examining which taxpayer and property attributes predict stronger behavioral responses to enforcement.

5.3. Treatment Effect Heterogeneity

The previous section established that the door-to-door intervention produced large and persistent improvements in property-tax compliance, raising payment rates by roughly 7.8 percentage points relative to an average baseline compliance of 63.4 percent. We now turn to the question of who responded to the intervention and why. Understanding the sources of treatment effect heterogeneity is essential for designing enforcement strategies that are both effective and equitable.

We explore these patterns using the Best Linear Projection (BLP) framework Semenova and Chernozhukov (2020) and Chernozhukov, Demirer, et al. (2025), which provides a doubly robust estimate of the linear relationship

$$\tau(X_i) = \beta_0 + A_i\beta, \tag{8}$$

where $\tau(X_i)$ denotes the conditional average treatment effect (CATE) for taxpayer i, and A_i is a vector of observable property and ownership characteristics. In this context, the BLP coefficients can be interpreted analogously to those from an OLS regression: they represent the best linear approximation of how treatment effects vary with observable attributes. A full derivation of the BLP estimator, along with a proof of the Neyman-orthogonality of our score function, is provided in Appendix F.

Table 3 reports the BLP regressions relating the estimated heterogeneous treatment effects from the causal forest, $\hat{\tau}_i$, to observable property, owner, and behavioral characteristics. The coefficients summarize how the probability of shifting from non-compliance to compliance varies with observable taxpayer and property attributes. The results reveal systematic and economically meaningful heterogeneity. Results are shown separately for the pre-intervention year (2019) and the first full year after the door-to-door program (2021), both using debiased estimates of the conditional treatment effects.

Before the program's rollout, treatment heterogeneity was essentially absent. The 2019 coefficients are small and statistically insignificant across all specifications, confirming that prior to the intervention there were no systematic differences in predicted responsiveness associated with property, ownership, or payment characteristics. This null finding serves as a placebo test, indicating that the heterogeneity captured by the model after 2021 does not reflect pre-existing structural correlations or model-induced bias. In contrast, the 2021 results reveal clear and theoretically consistent patterns of heterogeneity aligned with the behavioral mechanisms

Table 3: Best Linear Projection Results by Variable Set (2019 vs. 2021) - Debiased

	2019					**2021**				
Variable	(i)	(ii)	(iii)	(iv)	all	(i)	(ii)	(iii)	(iv)	all
Intercept	0.056	0.073	0.034	-0.002	0.099	0.628***	0.500***	0.108***	0.073***	0.570***
	(0.071)	(0.076)	(0.034)	(0.011)	(0.085)	(0.074)	(0.080)	(0.036)	(0.012)	(0.091)
A. Property Characteristics										
ln(Tax Base, 2021)	-0.003					-0.030***				
	(0.004)					(0.004)				
ln(Land Area)		-0.004			-0.003		-0.020***			-0.018***
		(0.005)			(0.005)		(0.005)			(0.005)
ln(Land Value)		-0.005			-0.005		-0.027***			-0.028***
		(0.006)			(0.006)		(0.006)			(0.006)
ln(Building Area)		0.002			0.002		-0.011*			-0.010*
		(0.008)			(0.008)		(0.008)			(0.008)
ln(Building Value)		0.010*			0.010*		-0.004			-0.004
		(0.006)			(0.006)		(0.008)			(0.008)
Dummy: New Construction		-0.113			-0.116		0.132			0.120
		(0.082)			(0.082)		(0.096)			(0.097)
Dummy: Old Construction		-0.126*			-0.127*		0.118			0.111
		(0.080)			(0.080)		(0.094)			(0.094)
B. Owner Characteristics (2021)										
Lives in Gorontalo			-0.029		-0.029			-0.043		-0.057*
			(0.034)		(0.034)			(0.036)		(0.037)
Multiple Properties			-0.018		-0.012			-0.046**		-0.024
			(0.018)		(0.018)			(0.018)		(0.018)
Same Address as Property			0.001		-0.000			0.009		0.008
			(0.009)		(0.009)			(0.009)		(0.009)
C. Historical Payment Behavior										
Paid 1 bill on time				0.002	0.005				-0.009	-0.004
				(0.018)	(0.018)				(0.019)	(0.019)
Paid 2 bills on time				0.021*	0.022*				0.002	0.004
				(0.017)	(0.017)				(0.018)	(0.018)
Paid 4 bills on time				0.012	0.011				0.010	0.006
				(0.014)	(0.014)				(0.016)	(0.016)
Paid 5 bills on time				0.003	0.002				0.017	0.014
				(0.012)	(0.013)				(0.015)	(0.015)
Paid 6 bills on time									-0.035**	-0.037***
									(0.014)	(0.014)

Notes: Coefficients on top, robust standard errors in parentheses below. Appendix K
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Omitted category: "Paid 3 bills on time."

discussed in the main analysis.

Among property characteristics, treatment effects are strongly and negatively associated with the taxable base, land area, and land value, indicating that smaller and lower-value properties benefited most from the intervention. The coefficient on log land area is -0.018, while that on log land value is -0.028, both statistically significant at the one-percent level. These magnitudes are economically meaningful. Given that the observed range of log land area in the sample is approximately 3.68 to 5.34 (corresponding to parcel sizes between 40 and 210 square meters), the estimated coefficient implies that moving from the smallest to the largest parcel is associated with roughly a 3 percentage-point lower treatment effect. Similarly, log land value ranges from about 10.7 to 14.6 (roughly Rp 45 million to Rp 2.2 billion), and the coefficient of -0.028 implies that properties at the top of the land value distribution experience treatment effects roughly 11 percentage points smaller than those at the bottom. These gradients suggest that the door-to-door program was particularly effective in improving compliance among owners of smaller and less valuable plots—taxpayers who are more likely to operate near the margin of compliance and for whom direct enforcement visits substantially increased the salience and perceived cost of nonpayment.

Owner characteristics also display meaningful behavioral gradients. Responsiveness is lower among owners residing in Gorontalo and among those with multiple properties, with coefficients ranging from -0.04 to -0.06 for residency and approximately -0.05 for multiple ownership. These results are consistent with theoretical expectations that wealthier or better-informed owners are less susceptible to behavioral nudges from enforcement visits, whereas absentee or single-property owners—who may be less familiar with tax procedures—respond more strongly to direct engagement by tax officers.

Historical payment behavior reinforces this interpretation. In 2019, past payment history shows no association with treatment effects. By 2021, however, properties with consistent on-time payment histories exhibit smaller treatment effects, while those with irregular or delayed payments display the largest gains. The negative and statistically significant coefficient on the indicator for "Paid six bills on time" (around -0.035, $p < 0.05$) suggests that the marginal impact of the intervention was smallest among already compliant taxpayers. In other words, households that had historically paid reliably changed little in response to the visits, while previously non-compliant taxpayers were the most responsive.

Table 4 reports the results of the Best Linear Predictor (BLP) regressions estimated separately for each year from 2019 to 2024 using all covariates jointly. Each specification projects the individual-level treatment effects estimated by the debiased causal forest, $\hat{\tau}_i$, onto a comprehensive set of property, owner, and historical payment characteristics. The coefficients represent the best linear approximation to how treatment responsiveness varies with observable taxpayer and property attributes over time.

The evolution of coefficients across years reveals a clear transition from an absence of systematic patterns before the program to well-structured and economically meaningful hetero-

Table 4: Best Linear Projection — All Covariates (Vertical by Year)

Variable	**2019**	**2021**	**2022**	**2023**	**2024**
Intercept	0.099	0.570***	0.512***	0.456***	0.484***
	(0.085)	(0.091)	(0.093)	(0.094)	(0.093)
A. Property Characteristics					
ln(Land Area)	-0.003	-0.018***	-0.029***	-0.012**	-0.019***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
ln(Land Value)	-0.005	-0.028***	-0.016***	-0.016***	-0.019***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Dummy: New Construction	-0.116	0.120	-0.090	0.017	-0.097
	(0.082)	(0.097)	(0.094)	(0.094)	(0.092)
Dummy: Old Construction	-0.127	0.111	-0.065	0.002	-0.083
	(0.080)	(0.094)	(0.092)	(0.092)	(0.090)
ln(Building Area)	0.002	-0.010	0.003	-0.011	-0.001
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
ln(Building Value)	0.010*	-0.004	0.002	0.003	0.005
	(0.006)	(0.008)	(0.008)	(0.008)	(0.007)
B. Owner Characteristics (2021)					
Lives in Gorontalo	-0.029	-0.057	-0.048	-0.104***	-0.067*
	(0.034)	(0.037)	(0.039)	(0.039)	(0.039)
Multiple Properties	-0.012	-0.024	-0.043**	-0.028	-0.033*
	(0.018)	(0.018)	(0.019)	(0.019)	(0.020)
Same Address as Property	-0.000	0.008	0.016*	-0.004	0.001
	(0.009)	(0.009)	(0.009)	(0.010)	(0.009)
C. Historical Payment Behavior					
Paid 1 bill on time	0.005	-0.004	-0.003	-0.059**	-0.023
	(0.018)	(0.019)	(0.023)	(0.024)	(0.025)
Paid 2 bills on time	0.022	0.004	-0.002	-0.035*	-0.007
	(0.017)	(0.018)	(0.019)	(0.020)	(0.020)
Paid 4 bills on time	0.011	0.006	-0.000	-0.003	-0.006
	(0.014)	(0.016)	(0.017)	(0.017)	(0.017)
Paid 5 bills on time	0.002	0.014	0.023	-0.006	-0.001
	(0.013)	(0.015)	(0.016)	(0.016)	(0.016)
Paid 6 bills on time		-0.037***	-0.037**	-0.036**	-0.024
		(0.014)	(0.015)	(0.015)	(0.015)

Notes: Coefficients on top; robust standard errors in parentheses below.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Omitted category for payment history: “Paid 3 bills on time.”

geneity after the intervention. In 2019, the coefficients are small, imprecisely estimated, and jointly insignificant.

Beginning in 2021, strong and consistent patterns emerge. The intercept increases sharply from 0.099 in 2019 to 0.570 in 2021, and remains in the range of 0.45–0.50 in subsequent years, indicating a substantial and persistent upward shift in the average predicted treatment effect following the full implementation of the intervention. This result aligns with the earlier average treatment effect (ATE) estimates, confirming that the door-to-door visits generated both large average effects and meaningful variation across taxpayers.

Overall, the year-by-year BLP estimates reveal patterns that are fully consistent with the findings for 2021. Property characteristics continue to show strong negative gradients with respect to land area and value, indicating that smaller and lower-value parcels remain the most responsive to the intervention. Owner characteristics and payment histories exhibit similarly stable relationships: absentee and single-property owners consistently display larger compliance gains, while taxpayers with strong prior payment records respond less. These persistent associations confirm that the heterogeneous treatment effects identified in 2021 reflect enduring behavioral mechanisms rather than short-term fluctuations.

In summary, the BLP results reveal a clear and persistent structure of treatment effect heterogeneity that emerges only after the implementation of the door-to-door program. Following 2021, the largest compliance gains are consistently concentrated among small, low-value properties, single-property or absentee owners, and taxpayers with irregular payment histories—groups that lie near the extensive margin of compliance. The stability of these patterns through 2024 indicates that the behavioral mechanisms underlying the program's effectiveness are both durable and predictable. The absence of any systematic relationship before 2021 further strengthens the causal interpretation of the estimated heterogeneity, demonstrating that the causal forest captures genuine behavioral responses rather than pre-existing correlates of compliance.

More broadly, the observed heterogeneity provides the empirical foundation forpolicy learning approaches to enforcement targeting. Following the frameworks of Athey and Wager (2021) and Yadlowsky et al. (2025), the estimated CATEs can be used to learn optimal assignment rules that maximize expected compliance gains subject to administrative costs or equity constraints. In this sense, the BLP results not only describe which taxpayers benefited most from the intervention, but also inform how future door-to-door programs can be optimized through data-driven policy assignment.

6 Treatment Prioritization and Policy Learning

The previous section documented heterogeneity in treatment effects, raising a practical question under capacity constraints: given available information, how should limited visits or reminders be allocated? This section lays out the framework and empirical strategy. We specify alternative enforcement objectives, derive the implied prioritization rules, and evaluate each rule on two dimensions: (i) How well this priorization increases revenue and tax participation (ii) how it distributes administrative effort across taxpayers.

We begin by defining objectives through three stylized roles, each associated with a distinct welfare criterion. The Teacher aims to maximize deterrence by moving as many taxpayers as possible into compliance and therefore ranks units by their conditional average treatment effect. The Businessman seeks to maximize fiscal returns and ranks units by the product of responsiveness and value, concentrating resources where expected revenue gains are largest. The Police concentrates enforcement on noncompliance risk, ranking units by a prognostic score for delinquency regardless of their estimated treatment effect. Two distributional benchmarks complete the spectrum: the most progressive rule favors lower-value properties and the most regressive rule favors higher-value properties, both irrespective of responsiveness.

Each welfare criterion -including the two benchmarks—induces a prioritization rule that answers "who should be targeted first." The three stylized rules map to standard goals in the economics of enforcement and to objectives commonly articulated in practitioner guidance while the progressive and regressive benchmarks provide conceptual bounds for equity–efficiency trade-offs.

Next, we assess how well each priorization rule targets causal payoffs. We evaluate performance with the Targeting Operator Characteristic (TOC) curve and its area, the Rank-Weighted Average Treatment Effect (RATE). The TOC traces the improvement that implementing the treatment priorization may yield over the average treatment effect. The RATE aggregates these gains across all feasible program sizes. Because the evaluation depends on weights, we report results on two margins that matter for tax policy: participation, which treats all taxpayers symmetrically, and revenue, which weights participation gains by a the expected revenue. Positive TOC values at low coverage and a positive RATE indicate that a rule successfully concentrates treatment on units with above-average causal payoffs for the relevant margin.

To ensure credible out-of-sample assessment, we estimate scores on a training subsample and evaluate TOC and RATE on a holdout test subsample. Sixty percent of the data are used for training and forty percent for testing[8]. Table 9 shows that AIPW estimates computed separately on the train and test sets fall within the 95 percent confidence intervals of the full-sample estimates, indicating negligible loss of precision from sample splitting and supporting the stability of the effects used to construct the scores.

We then examine distributional implications. Even if two rules perform similarly on TOC and RATE, they may allocate administrative effort very differently across taxpayers. We therefore study concentration curves that plot the cumulative share of an outcome—visits, payments, or liabilities—against the cumulative share of the population when taxpayers are ranked by a prioritization score. Curves above the 45-degree line indicate progressive allocations that tilt effort toward the lower ranks of the score distribution, while curves below the diagonal indicate regressive allocations that concentrate effort among higher ranks. To summarize these patterns in a single, comparable statistic, we report a Policy Concentration Index (PCI) that rescales the traditional concentration index to the interval $[1,1]$, anchoring the extremes at the most progressive and most regressive feasible allocations. Values near zero indicate proportional allocation; values near the bounds indicate sharp departures from proportionality. Inference for curves and indices relies on a nonparametric bootstrap, and we report means, standard deviations, and percentile confidence intervals.

The presentation proceeds as follows. Section 6 formalizes the objectives, defines the induced scores, and reports TOC curves and RATE on deterrence and revenue. Section 6.2 turns to distributional patterns, presenting concentration curves and the PCI for each rule, with disaggregated figures by treatment status in the appendix. Anticipating the main results, prioritization by the conditional average treatment effect delivers broad-based deterrence gains and nontrivial revenue improvements with a progressive allocation; scaling by value yields the largest fiscal gains but concentrates effort in the upper ranks; risk targeting is closer to uniform enforcement and produces modest improvements on both margins. Together, these findings make the equity–efficiency trade-offs explicit and provide a transparent basis for policy choice.

6.1. Welfare–induced prioritization rules

When resources are scarce, the central policy question is whom to treat first—in our case, which taxpayers are most responsive to a visit. We build on the policy–learning literature (Manski, 2004; Chernozhukov, Lee, et al., 2025; Athey & Wager, 2021; Kitagawa & Tetenov, 2018) to show how a welfare criterion induces an operational prioritization rule, yielding a score $S(X)$

[8] This split is mainly heuristics and follows observed practice in policy learning applications. Given the sample size, results are insensitive to reasonable alternatives such as 50/50.

that ranks units by treatment priority. We then evaluate any candidate S—whether derived from estimated treatment effects, predicted payment probabilities, or simple heuristics—using the Targeting Operator Characteristic (TOC) and its scalar summary, the Rank-Weighted Average Treatment Effect (RATE; Yadlowsky et al. (2025)), which measure the causal gain from targeting the top q share of the score distribution relative to random allocation and aggregate this curve into a single, comparable statistic.

A deterministic policy is a measurable map $\pi : \mathscr{X} \to \{0,1\}$ that treats unit i when $\pi(X_i) = 1$. Let the welfare weights $\lambda(x) \geq 0$ encode the policy margin and let $c(x) > 0$ be the per-unit cost. With a budget (or coverage) constraint $\mathbb{E}[c(X)\pi(X)] \leq B$ (coverage q corresponds to $c \equiv 1$, $B = q$), the planner solves

$$\max_{\pi:\mathscr{X}\to\{0,1\}} \mathbb{E}\big[\lambda(X)\,\tau^{(t)}(X)\,\pi(X)\big] \qquad \text{s.t.} \qquad \mathbb{E}\big[c(X)\,\pi(X)\big] \leq B. \tag{9}$$

The Lagrangian for (9) implies a pointwise threshold rule: treat unit x iff

$$\lambda(x)\,\tau^{(t)}(x) - \eta\,c(x) \geq 0 \quad \Longleftrightarrow \quad \underbrace{S_t^\star(x)}_{\text{welfare score}} = \frac{\lambda(x)\,\tau^{(t)}(x)}{c(x)} \geq \eta,$$

where η is chosen so that the constraint binds. Thus a welfare criterion with weights $\lambda(\cdot)$ *induces a prioritization rule*: rank by $S_t^\star(x)$ and treat those above the cutoff. When costs are homogeneous ($c \equiv 1$), this reduces to ranking by $\lambda(X)\tau^{(t)}(X)$ and treating the top q share. Only the induced ranking matters (strictly monotone transformations of $S_t^\star$ are equivalent); ties may be broken arbitrarily or at random.

Examples of welfare weights:

1. *Participation objective*: $\lambda(x_i) \equiv 1$, $c \equiv 1 \Rightarrow$ rank by $\tau^{(t)}(x_i)$ and treat the top q.

2. *Revenue objective*: $\lambda(x_i) \equiv w_i$ (e.g., assessed liability), $c \equiv 1 \Rightarrow$ rank by $w_i\,\tau^{(t)}(x_i)$.

3. *Budgeted targeting with heterogeneous costs*: $c(x_i)$ varies (e.g., distance/travel time) $\Rightarrow$ rank by $\tau^{(t)}(x_i)/c(x_i)$ for participation, or $w_i\tau^{(t)}(x_i)/c(x_i)$ for revenue.

In practice, $\tau^{(t)}(X)$ and possibly $c(X)$ are unknown. We estimate $\hat{\tau}^{(t)}(X)$ via the debiased causal-forest procedure, assume $c(X) = 1$ and construct the implementable score:

$$\hat{S}_t(X) = \lambda(X)\,\hat{\tau}^{(t)}(X)$$

Given a coverage target $q \in (0,1]$, the induced policy treats the top q share by $\hat{S}_t$:

$$\pi_q(X) \;=\; \mathbf{1}\Big\{\,\hat{S}_t(X) \geq F^{-1}_{\hat{S}_t}(1-q)\,\Big\}.$$

where $F_{\hat{S}}$ is the marginal distribution (cdf) of $\hat{S}(X_i)$ and tied scores are broken arbitrarily.

$$\mathrm{TOC}_\lambda(q;S) = \mathbb{E}\Big[\;Y_{\lambda,i}(1) - Y_{\lambda,i}(0) \mid S(X_i) \geq F^{-1}_{S(X_i)}(1-q)\Big] - \underbrace{\mathbb{E}\big[\;Y_{\lambda,i}(1) - Y_{\lambda,i}(0)\,\big]}_{ATE} \tag{10}$$

By construction, $\mathrm{TOC}_\lambda(1;S) = 0$. If $\mathrm{TOC}_\lambda(q;S) > 0$ for small and moderate q, high-ranked units exhibit larger effects than the average unit. The RATE aggregates (10) over coverage levels; with uniform weights it equals the area under the TOC curve:

$$\mathrm{RATE}_\lambda(S) \;=\; \int_0^1 \mathrm{TOC}_\lambda(q;S)\,dq. \tag{11}$$

Large positive $\mathrm{RATE}_\lambda(S)$ indicates effective targeting; values near zero imply little gain over random allocation; negative values indicate systematic misranking.

$\mathrm{RATE}(S)$ captures two ingredients: the available "signal" —dispersion in conditional treatment effects $\tau(X_i)$—and the "extraction" of that signal by the score $S(X_i)$ via its induced ordering. If effects are homogeneous, ordering is irrelevant and $\mathrm{TOC}(q;S) = 0$ for all q, hence $\mathrm{RATE}(S) = 0$. With heterogeneity, $\mathrm{RATE}(S)$ is large and positive when S closely aligns with $\tau(X_i)$ (high rank correlation), near zero if S is weakly informative, and negative under systematic misranking that elevates low-effect units.

We evaluate targeting performance along two policy margins. The tax *participation* margin treats all taxpayers symmetrically (equal weights), while the *revenue* margin weights outcomes by monetary potential (e.g., assessed tax liability). We then compare three prototypical government objectives—participation, revenue, and a welfare-weighted hybrid—against two normative baselines that anchor the distributional spectrum (most progressive vs. most regressive targeting of door-to-door visits).

For each objective, we derive the implied prioritization rule $S(X_i)$ and assess whether implementing it would yield statistically and economically meaningful gains. Performance is summarized by the RATE (area under the TOC), and by pointwise policy gains at selected coverage levels, such as $P(60)$, defined as the improvement in the average treatment effect among the treated (ATT) when the top 60% of the score distribution is targeted rather than assigned at random. This framework allows transparent, apples-to-apples comparisons of alternative rules on

Table 5: Value — Tax Participation and Revenue Effects by Strategy and Percentile Cutoffs (Panels A–C)

	P(5)	P(10)	P(25)	P(50)	P(75)	P(90)	P(95)	P(60)^1	**RATE**
Panel A. Teacher Strategy									
Participation	0.173^{***}	0.127^{***}	0.096^{***}	0.056^{***}	0.024^{***}	0.012^{***}	0.007^{***}	0.041^{***}	0.067^{***}
	(0.031)	(0.022)	(0.013)	(0.007)	(0.004)	(0.002)	(0.002)	(0.006)	(0.007)
Revenue	0.751^{***}	0.667^{***}	0.479^{***}	0.336^{***}	0.164^{***}	0.166^{***}	0.070^{***}	0.209^{***}	0.343^{***}
	(0.226)	(0.168)	(0.125)	(0.097)	(0.075)	(0.054)	(0.045)	(0.089)	(0.082)
Panel B. Businessman Strategy									
Participation	0.014	0.011	−0.006	−0.010	−0.004	−0.001	−0.001	−0.006	−0.001
	(0.029)	(0.020)	(0.012)	(0.007)	(0.004)	(0.002)	(0.002)	(0.006)	(0.007)
Revenue	2.486^{***}	1.822^{***}	0.638^{***}	0.246^{***}	0.104^{***}	0.015	−0.002	0.210^{***}	0.595^{***}
	(0.988)	(0.590)	(0.245)	(0.096)	(0.044)	(0.031)	(0.028)	(0.069)	(0.190)
Panel C. Police Strategy									
Participation	0.017	−0.008	−0.004	0.013	0.009^{*}	0.002	0.001	0.018^{**}	0.008
	(0.029)	(0.021)	(0.012)	(0.007)	(0.004)	(0.002)	(0.002)	(0.006)	(0.007)
Revenue	0.206	−0.009	0.008	0.181	0.029	0.008	0.003	0.108	0.104
	(0.433)	(0.286)	(0.163)	(0.097)	(0.057)	(0.029)	(0.018)	(0.080)	(0.095)

Notes: Standard errors in parentheses. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Within each panel, the *Participation* row reports incremental tax participation effects; the *Revenue* row reports incremental revenue effects (USD) at the indicated percentiles. The final column reports the corresponding RATE for each outcome across the evaluation sample.

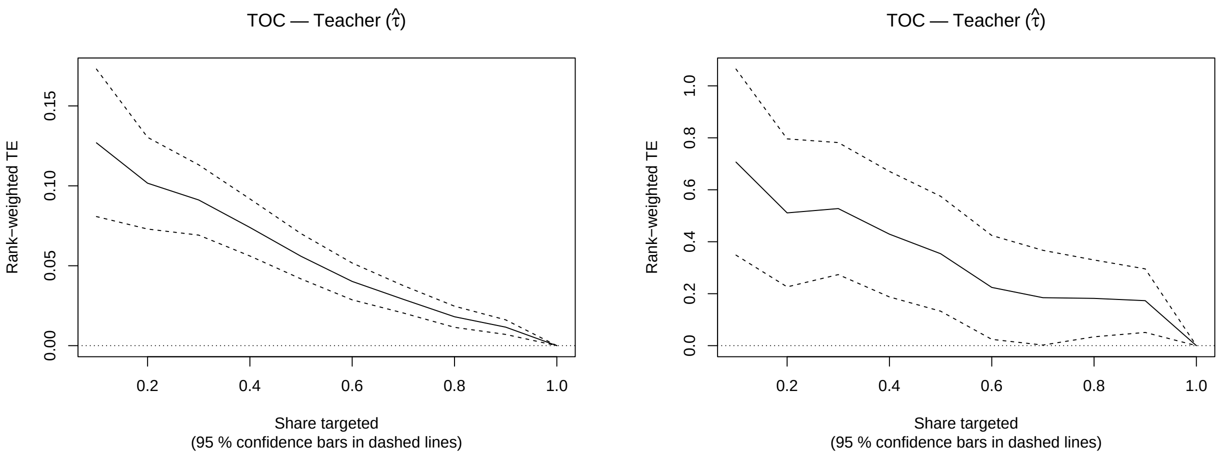
both participation and revenue margins.

Let $r \in \{\mathrm{T}, \mathrm{B}, \mathrm{P}\}$ index the three roles Teacher, Businessman, and Police. For each role we define a welfare function $V^r(\pi)$ and an associated prioritization rule $S^r(X_i)$ with evaluation score Γ_i. Table 5 contain all gains in terms of participation and revenue mobilization of the proper targeting vis a vis the random assingment of treatment. As reference, ATT on deterance is 7.8 while the average treatment on deterance is 1.6USD.

Government as Teacher: Participation–maximizing

This formulation adopts a behavioral perspective in which the state seeks to maximize deterrence by raising the number of compliant taxpayers. The idea of teacher strategy comes from the literature of tax education. This is also implicit in the literature og tax nudges that looks at the taxation problems where evasion is not possível as an efort to increase the number of

Figure 7: TOC — Teacher's Strategy



Note: Left panel: TOC by participation. Right panel: TOC by value (participation × tax liability). Confidence bands (if shown) are pointwise 95% intervals from a nonparametric bootstrap with 1,000 resamples.

compliance.

Each taxpayer contributes equally to welfare, regardless of tax liability:

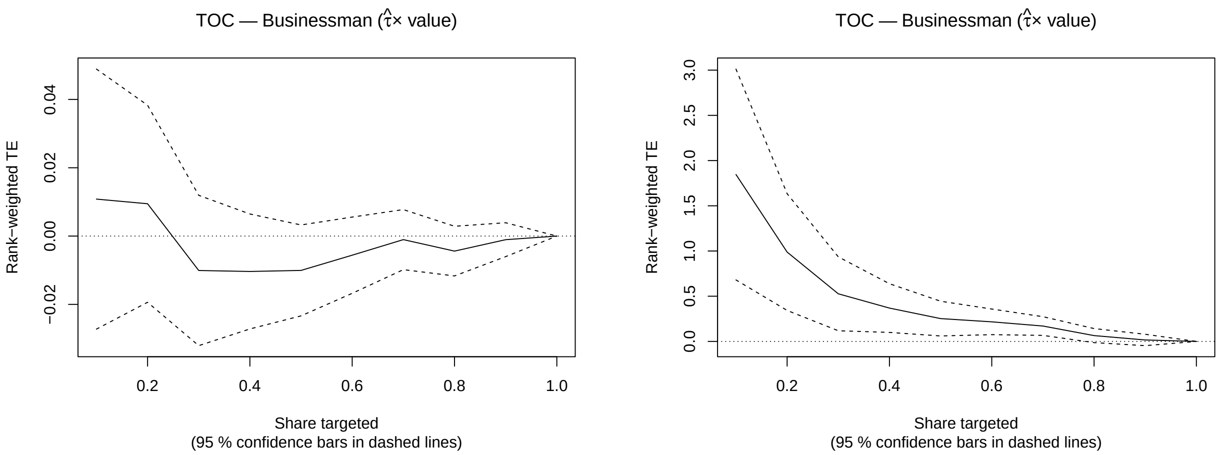
$$V^{\text{Teacher}}(\pi) = \mathbb{E}\big[\,Y_i\big(\pi(X_i)\big) - Y_i(0)\big]\,. \tag{12}$$

$\lambda(x_i) \equiv 1$, $c \equiv 1 \Rightarrow$ rank by $\tau^{(t)}(x_i)$ and treat the top q so that taxpayers with the largest predicted compliance gains are targeted first.

Figure 7 displays the targeting operator characteristic (TOC) under the Teacher strategy. The left panel shows gains on the deterrence margin and the right panel shows gains on the value margin. Table 5 Panel A summarizes the corresponding quantification. There are strong participation gains: 17.3 percentage points at the top 5 percent, 12.7 at 10 percent, and 9.6 at 25 percent, tapering toward zero in the right tail. The RATE of 6.7 percentage points indicates substantial aggregate improvement in compliance when implementing this prioritization. There are also gains in revenue the top 5 percent yields an average gain of 0.75 USD per property and effects remain positive down to the 95th percentile (0.07 USD).

In summary, adopting the teacher's strategy yield gains in deterance and in total value.

Figure 8: TOC — Businessman's Strategy



Note: Left panel: TOC by participation. Right panel: TOC by value (participation × tax liability). Confidence bands (if shown) are pointwise 95% intervals from a nonparametric bootstrap with 1,000 resamples.

The Government as Businessman: Revenue–maximizing

A revenue–maximizing government allocates visits to maximize expected fiscal returns. Let w_i denote the property's tax liability (or predicted payment). The welfare function places more weight on units with larger expected payments:

$$V^{\text{Business}}(\pi) = \mathbb{E}\big[\, w_i \cdot \big(Y_i(\pi(X_i)) - Y_i(0)\big)\big]. \tag{13}$$

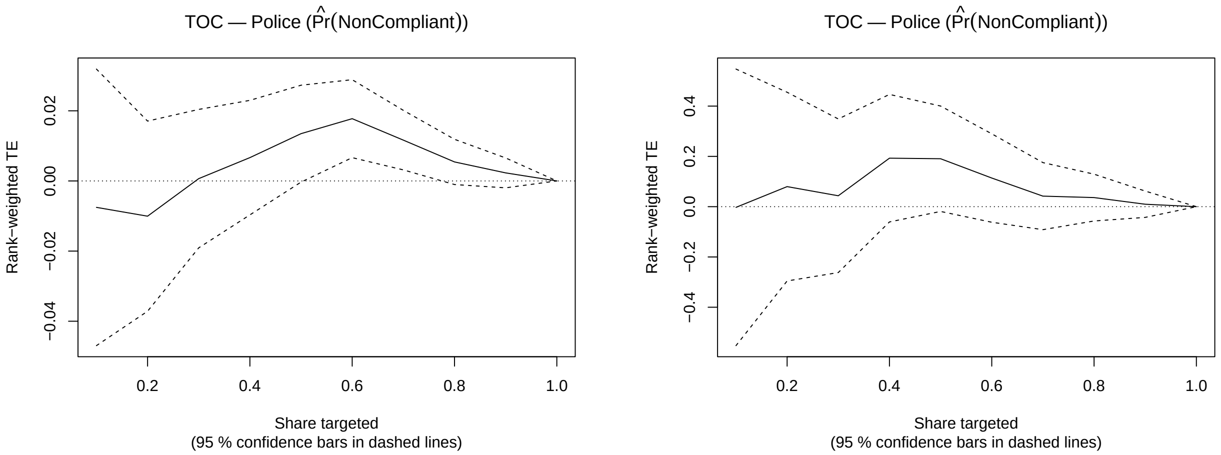
Then, $\lambda(x_i) \equiv w_i \; c \equiv 1 \Rightarrow$ rank by $w_i \, \tau(x_i)$.

Table 5 shows that this ranking performs weakly on the deterrence margin, with estimates close to zero or negative and an effectively null RATE. In contrast, it greatly increase revenue. The top 5 percent delivers an incremental 2.49 USD per property, more than 50 percent above the 1.6 USD sample average, with positive and significant gains extending to the 75th percentile. The RATE of 0.60 USD suggests an improvement of nearly 40 percent relative to a uniform policy. Revenue-maximizing targeting thus delivers strong fiscal returns, even though its deterrence effects are limited.

The Government as Police: Risk based targeting

An enforcement-oriented government targets the riskiest taxpayers, defined as those most likely to be noncompliant. Let $g_{-1}(X_i)$ denote the prognostic score, the predicted probability of

Figure 9: TOC — Police Strategy



Note: Left panel: TOC by participation. Right panel: TOC by value (participation × tax liability). Confidence bands (if shown) are pointwise 95% intervals from a nonparametric bootstrap with 1,000 resamples.

noncompliance. The welfare function penalizes failing to audit high-risk units:

$$V^{\text{Police}}(\pi) = \mathbb{E}\big[-g_{-1}(X_i)\,\pi(X_i)\big]. \tag{14}$$

We estimate $g_{-1}(X_i)$ with a boosted regression forest trained on pre-intervention data, using compliance in the intervention year as the outcome. The resulting prioritization rule is $S_i^{\text{P}} = g_{-1}(X_i)$ with $\Gamma_i^{\text{P}} = g_{-1}(X_i)$. This criterion focuses on detection rather than behavioral response and does not depend on $\tau(X_i)$.

Table 5 indicates small and statistically insignificant revenue effects overall, with a modest increase at the 50 percent cutoff (0.18 USD) and a RATE of approximately 0.10 USD. In Table **??**, deterrence effects are modest as well; statistically significant improvements appear at the 75th percentile and at the operational 60 percent cutoff, but aggregate gains are limited (RATE of 0.8 percentage points). Enforcement-oriented targeting may serve fairness or rule-of-law goals but is relatively inefficient for inducing behavioral change or revenue growth.

Most Progressive Policy

The progressive benchmark favors taxpayers with lower property values. Let A_i denote the assessed property value, a proxy for fiscal capacity. The welfare function and ranking rule are

$$V^{\text{Prog}}(\pi) = \mathbb{E}[-A_i\,\pi(X_i)], \qquad S_i^{\text{Prog}} = -A_i. \tag{15}$$

Table 6: Value — Tax Participation and Revenue Effects by Strategy and Percentile Cutoffs (Panels D–E)

	P(5)	P(10)	P(25)	P(50)	P(75)	P(90)	P(95)	P(60)^{1}	**RATE**
Panel D. Most Progressive Policy									
Participation	-0.107^{***}	-0.065^{***}	-0.057^{***}	-0.032^{***}	-0.016^{***}	-0.005^{**}	-0.003^{**}	-0.023^{***}	-0.036^{***}
	(0.028)	(0.020)	(0.012)	(0.007)	(0.005)	(0.003)	(0.002)	(0.006)	(0.006)
Revenue	-1.171^{***}	0.230	0.237	0.194	0.102	0.046	0.023	0.166	0.080
	(1.258)	(0.774)	(0.285)	(0.100)	(0.034)	(0.011)	(0.005)	(0.068)	(0.222)
Panel E. Most Regressive Policy									
Participation	0.059^{***}	0.045^{**}	0.047^{***}	0.032^{***}	0.019^{***}	0.007^{**}	0.006^{**}	0.025^{***}	0.035^{***}
	(0.033)	(0.023)	(0.013)	(0.007)	(0.004)	(0.002)	(0.002)	(0.005)	(0.007)
Revenue	-0.445^{***}	-0.411^{***}	-0.305^{***}	-0.194^{***}	-0.079^{**}	-0.026	0.062	-0.160^{***}	-0.203^{***}
	(0.096)	(0.096)	(0.096)	(0.095)	(0.091)	(0.083)	(0.066)	(0.094)	(0.086)

Notes: Standard errors in parentheses. Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Within each panel, the *Participation* row reports incremental tax participation effects; the *Revenue* row reports incremental revenue effects (USD) at the indicated percentiles. The final column reports the corresponding RATE for each outcome across the evaluation sample.

This rule represents an extreme redistributive objective, allocating enforcement or assistance toward small properties regardless of predicted responsiveness. It is informative for the equity frontier rather than for efficiency.

Table 6 shows mixed and largely insignificant value effects, with a small RATE (0.08 USD). On the deterrence margin, effects are negative across percentiles and the RATE is −3.6 percentage points, consistent with weaker responsiveness among small taxpayers or higher administrative frictions.

Most Regressive Policy

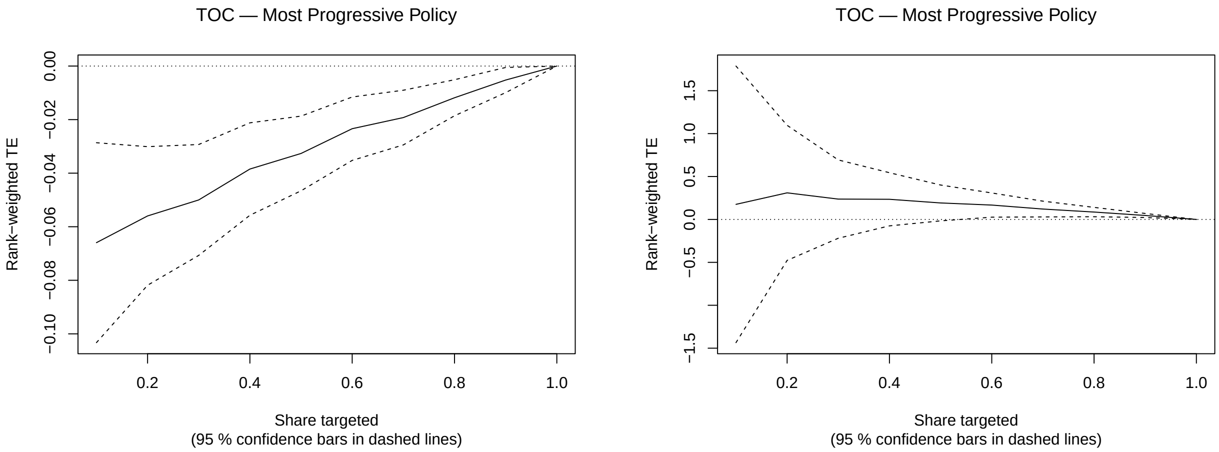
The regressive benchmark is the mirror image, prioritizing higher-value properties regardless of compliance risk or behavioral response:

$$V^{\text{Regr}}(\pi) = \mathbb{E}[A_i \, \pi(X_i)], \qquad S_i^{\text{Regr}} = A_i. \tag{16}$$

This rule anchors the opposite bound of the equity spectrum.

On the value margin Table 6, coefficients are predominantly negative with a RATE of −0.20 USD, suggesting limited marginal returns among the very top of the value distribution, possibly due to preexisting compliance. In contrast, there are positive deterrence effects, including 5.9

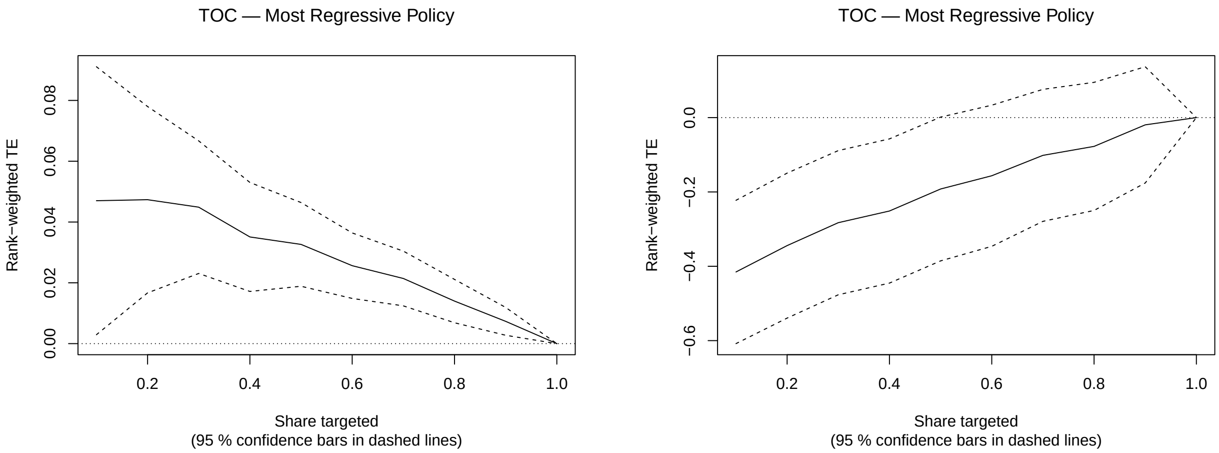
Figure 10: TOC — Most Progressive Policy



Note: Left panel: TOC by participation. Right panel: TOC by value (participation × tax liability). Confidence bands (if shown) are pointwise 95% intervals from a nonparametric bootstrap with 1,000 resamples.

percentage points at the top 5 percent and a RATE of 3.5 percentage points. Although smaller than the gains obtained by ranking on predicted responsiveness, these effects exceed random assignment, indicating that some high-value taxpayers do respond on the compliance margin.

Figure 11: TOC — Most Regressive Policy



Note: Left panel: TOC by participation. Right panel: TOC by value (participation × tax liability). Confidence bands (if shown) are pointwise 95% intervals from a nonparametric bootstrap with 1,000 resamples.

These five formulations highlight the flexibility of the policy-learning framework: by redefining $V(\pi)$, policymakers can align statistical optimization with deterrence, revenue, fairness, or normative distributional goals. While the first three formulations correspond to objectives commonly discussed in the empirical and theoretical literature, the last two serve as natural policy bounds—representing, respectively, purely progressive/regressive. In practice, these objectives often conflict—a policy that maximizes revenue may be inequitable, while one that enforces uniform audits or favors the poor may be fiscally inefficient. The unified framework presented here allows us to quantify and compare these trade-offs using identical data and estimation procedures.

The most regressive policy and the teacher's priorization strategy both prioritize targeting population with better registered records.

6.2. Distributional Patterns of Targeting: Concentration Curves and the Policy Concentration Index

This subsection characterizes how alternative prioritization rules distribute administrative effort across taxpayers and how those distributions translate into observed compliance. The analysis is conducted on the held-out test sample pooling treated and untreated properties; We assess distributional orientation—progressive versus regressive—using concentration curves and a scalar Policy Concentration Index (PCI). For further details on construction, estimation, and inference for the concentration curves and PCI are provided in Appendix H.

Let $R_i \in [0,1]$ be the fractional rank of unit i in ascending order of $S(X_i)$, and let $y_i \geq 0$ denote the outcome of interest (for example, visits, payments, or liabilities). The concentration curve associated with S is

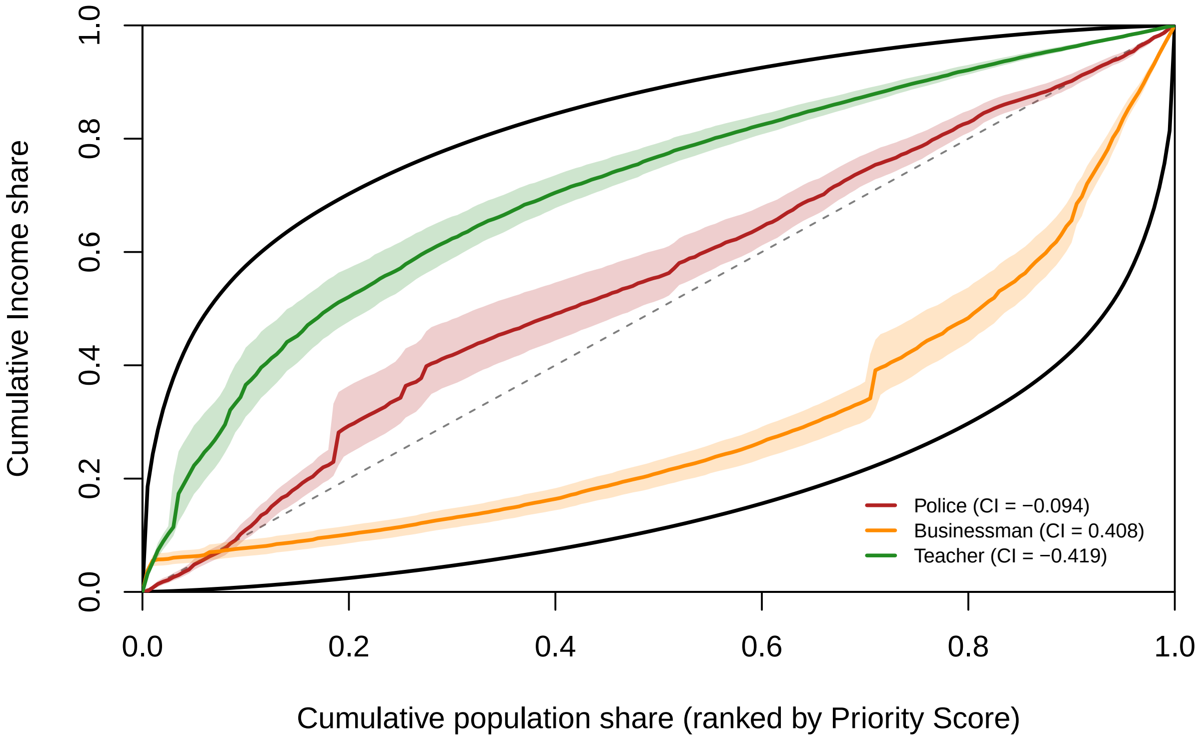
$$L(p;S) \equiv \frac{\sum_{i:R_i \leq p} y_i}{\sum_i y_i}, \qquad p \in [0,1].$$

If administrative effort is distributed proportionally to population, $L(p;S) = p$ and the curve coincides with the 45-degree line. Curves above the diagonal indicate progressive allocations that tilt effort toward lower-ranked units; curves below the diagonal indicate regressive allocations that concentrate effort among higher-ranked units.

The Lorenz curve is a special case of a concentration curve in which individuals are ranked by their own outcome - income - y_i. In that case the curve lies on or below the 45° line, and the Gini coefficient G_y equals the concentration index for that ranking.

We report concentration curves for the main prioritization rules together with conceptual bounds that correspond to the most progressive and most regressive feasible allocations given the observed outcome distribution. To summarize each curve in a single statistic, we compute the traditional concentration index which is negative for progressive allocations and positive for regressive allocations.

Figure 12: Concentration Curves



Note: Green: *Teacher* (participation maximization); Yellow: *Businessman* (revenue maximization); Red: *Police* (delinquency targeting). The upper black envelope marks the most regressive feasible policy; the lower black envelope (the Lorenz curve) marks the most progressive. The gray region is the feasible policy set. Values in parentheses report concentration indices (CI) prior to normalization. Confidence bands for the Green, Yellow, and Red curves are pointwise 95% intervals computed via bootstrap (1,000 resamples).

In practice the CI are not bounded by -1 and 1, but by the -GINI and the GINI. Concretely, there is no allocation that can be more progressive that explicitly ordering people by their income. Because the feasible range of CI(S) varies with the outcome distribution, we also report a normalized Policy Concentration Index that rescales out measure to $[-1,1]$. Let $\text{CI}^- < 0$ and $\text{CI}^+ > 0$ denote the indices for the most progressive and most regressive feasible allocations. Then, PCI $= -1$ and $+1$ are the progressive and regressive bounds, and PCI $= 0$ indicates proportional allocation. Inference for curves and indices relies on bootstrap; we report means, standard deviations, and percentile confidence intervals.

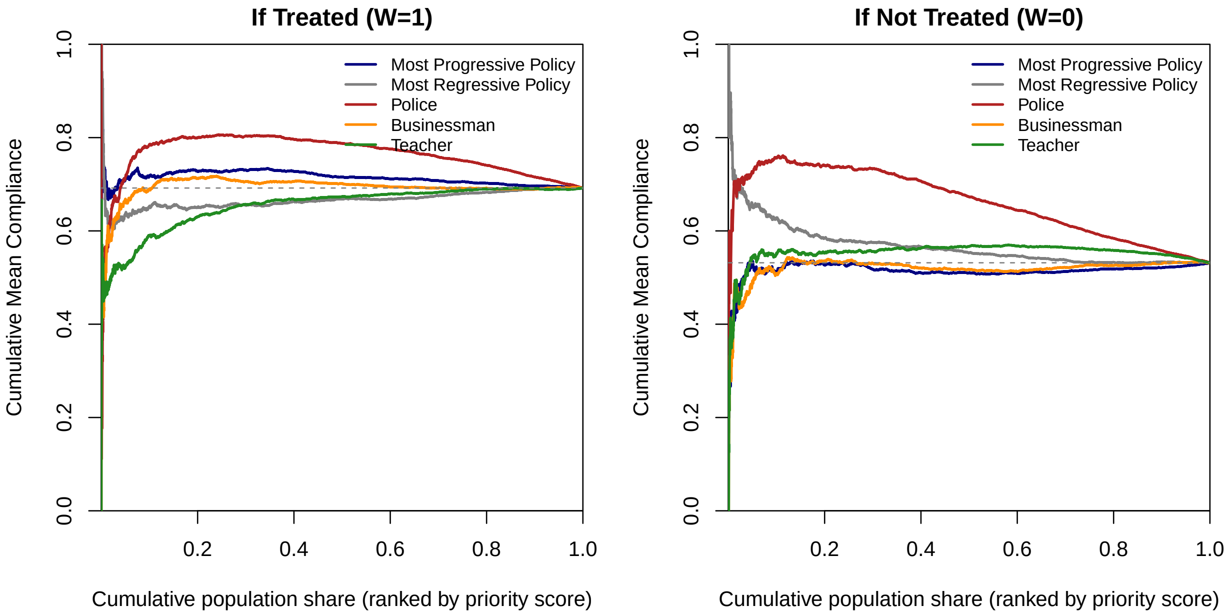
Table 7 reports the concentration index and normalized PCI by rule. The *Teacher* (deterrence-maximizing) rule yields a strongly negative CI of about -0.42, placing it near the progressive bound on the PCI scale. The *Businessman* (revenue-maximizing) rule shows a large positive CI

Table 7: Concentration Index and Policy Concentration Index Estimates

	Concentration Index	Policy Concentration Index			
Policy	Test Data	Test Data	2.5%	Median	97.5%
Police	-0.094	-0.141	-0.218	-0.139	-0.076
Businessman	0.408	0.620	0.574	0.620	0.664
Teacher	-0.419	-0.636	-0.681	-0.635	-0.599
Gini (Most Progressive Policy) = 0.659					

Notes: The table reports bootstrap percentile confidence intervals (2.5%, 50%, 97.5%) for the Concentration Index (CI) and Policy Concentration Index (PCI) across prioritization rules with 1000 repetitions. The "Test Data" column shows the point estimate computed from the held-out sample.

Figure 13: Mean compliance by treatment prioritization



Note: The figure plots cumulative mean compliance in the held-out test sample as coverage increases from 0 to 100% along each rule's ranking $S(X)$. At coverage p, the value is the average compliance among the lowest-ranked p share of properties. Curves below (above) the sample mean at low coverage indicate targeting of low- (high-) compliance units. Lines are by prioritization rule (and, where applicable, by treatment status). Shaded areas, if shown, are 95% bootstrap confidence bands.

near +0.41, close to the regressive bound. The *Police* (risk-based) rule has a small negative CI (around −0.09), implying near-proportional allocation with a slight tilt toward lower ranks. For context, the Gini coefficient of the outcome distribution is approximately 0.66, indicating substantial baseline inequality and motivating normalization via the PCI. Bootstrap intervals for Teacher and Businessman exclude zero and do not overlap, highlighting robust distributional differences; dispersion is modest (bootstrap SDs < 0.03).

Figure 13 plots cumulative mean compliance among units prioritized by each rule as coverage increases (i.e., as the share of ranked properties considered rises from 0 to 1). In the test sample, the learned policy scores successfully identify out-of-sample low-compliance properties early in the ranking, validating predictive content. The Teacher rule prioritizes low-compliance ranks—consistent with a deterrence objective—so the average pre-visit compliance of the visited set initially lies below the sample mean and rises with coverage. By contrast, the Businessman rule and the most progressive value-based benchmark track the sample mean more closely across coverage profiles, reflecting their focus on value rather than on raising the number of compliers. Results disaggregated by treatment status show the same ordering (Appendix H).

The concentration-curve and PCI evidence place the rules at distinct points along the equity–efficiency frontier. *Teacher* delivers a progressive allocation consistent with broad-based deterrence. *Businessman* concentrates resources where expected revenue payoffs are largest, yielding a regressive allocation by construction. *Police* lies closer to uniform enforcement across ranks. These distributional facts complement the targeting-performance results in Section 6 and clarify the trade-offs policymakers face between inclusionary deterrence and revenue concentration.

7 Conclusion

This paper evaluates a large-scale door-to-door enforcement program for property taxation and develops a policy-learning framework to allocate scarce enforcement capacity. Two sets of results anchor the conclusions. First, the intervention produced sizable and persistent average gains in compliance, with clear and economically meaningful heterogeneity across taxpayers. Second, alternative targeting rules trace out a transparent equity–efficiency frontier: rules that maximize broad participation differ systematically from rules that maximize fiscal yields, and both differ from conventional risk-based selection. Together, these findings provide an empirically grounded basis for designing enforcement that is effective, predictable, and aligned with local policy objectives.

On average effects, the program substantially raised timely payment after the 2021 rollout. Placebo estimates in 2019 are negligible, ruling out spurious correlations and supporting a causal interpretation. Following implementation, the average treatment effect rises to 7.8 percentage points in 2021 and remains large thereafter—7.3, 6.4, and 5.1 percentage points through 2024—equivalent to roughly a 12 percent increase relative to a 63.4 percent baseline in the first post period, tapering to about 8 percent by year three. Effects are consistently larger for visited properties than for unvisited ones, with a statistically significant but small gap, consistent with strong direct impacts and limited spillovers. These estimates are robust to alternative causal-forest specifications without the debiasing correction and with covariates restricted to pre-treatment years, indicating that the main conclusions are not artifacts of model choice.

Heterogeneity is first-order. Splitting taxpayers at the median predicted treatment effect, high-$\hat{\tau}$ units exhibit gains that are multiple times larger than those of low-$\hat{\tau}$ units, with gaps on the order of 6–11 percentage points across post-intervention years. Best linear projections show that responsiveness declines with parcel size and value and is lower for residents and multi-property owners, while prior on-time payers exhibit smaller treatment gains. These patterns are absent in the pre-period and persist from 2021 through 2024, implying that the underlying behavioral mechanisms are durable and that the estimated conditional treatment effects are stable predictors for targeting.

Turning from evaluation to optimization, the policy-learning analysis compares five prioritization rules estimated on training data and evaluated out of sample. A rule that ranks by predicted treatment effects (the Teacher strategy) delivers broad deterrence gains and nontrivial revenue improvements: the targeting operator characteristic lies well above zero at low coverage and the corresponding area, the rank-weighted average treatment effect, is positive on both margins. At five percent coverage, the Teacher rule increases the average treatment effect on compliance by roughly 17 percentage points relative to a 7.8 point baseline and yields positive value gains, with an area under the curve around a fifth of the average value-margin effect. A rule that scales treatment effects by value (the Businessman strategy) produces the largest revenue gains in both levels and area but little movement in participation, consistent with the intensive margin dominating among high-value accounts. A risk-based rule that targets predicted non-compliance (the Police strategy) sits near uniform enforcement, with modest improvements on both margins and much smaller areas under the TOC. Two distributional benchmarks, which prioritize the lowest-value or highest-value properties irrespective of responsiveness, perform as conceptual bounds rather than practical policies: the progressive benchmark underperforms on deterrence and has weak value gains, while the regressive benchmark shows negative value gains on net but some deterrence at the very top of the ranking.

Distributional diagnostics make the trade-offs explicit. Concentration curves and a normalized policy concentration index reveal that the Teacher rule allocates effort progressively, with a large negative concentration index and a PCI close to the progressive bound. The Businessman rule concentrates on upper ranks by construction, with a large positive concentration index near the regressive bound. The Police rule is near proportional with a mildly negative concentration index. Bootstrap dispersion is small and intervals for the Teacher and Businessman rules do not overlap, underscoring that distributional orientation is a stable feature of the underlying scores rather than sampling noise. These results mirror the cumulative-compliance profiles: participation gains under the Teacher rule arrive earlier in the score distribution, whereas revenue gains under the Businessman rule arrive later among higher-value ranks.

The combined evidence yields clear implications for tax administration under binding capacity constraints. When the policy objective is to broaden compliance and strengthen social norms, ranking by predicted treatment effects is a robust default: it delivers the largest participation gains, preserves meaningful revenue improvements, and does so with a progressive allocation of administrative effort. When short-run revenue mobilization is the binding constraint, scaling by value achieves the largest fiscal returns but concentrates enforcement among higher-value taxpayers and has limited effects on participation, a trade-off that may or may not be acceptable depending on distributional and political economy considerations. Conventional risk targeting offers a plausible status quo but leaves efficiency gains on the table relative to rules that directly prioritize causal payoffs.

These choices matter for equity as well as efficiency. Because responsiveness is highest among smaller, lower-value properties while the tax base is concentrated among higher-value accounts, optimizing for different objectives will tilt enforcement in opposite directions. A policy that is progressive in design can become effectively regressive in implementation if the assignment rule is optimized solely for revenue; conversely, a deterrence-oriented rule can achieve broad-based gains without forgoing fiscal space. Concentration curves and the PCI provide administrators with a compact way to audit these consequences ex ante and to set explicit guardrails or quotas if distributional concerns warrant them.

The results also speak to algorithmic governance. All targeting rules are evaluated out of sample, and their gains are statistically validated with bootstrap inference, illustrating how modern machine learning can be operationalized within credible evaluation pipelines. The framework is transparent about objectives, score construction, and welfare criteria, and it naturally accommodates administrative constraints such as fixed coverage or geographic quotas. In practice, agencies can implement Teacher-style prioritization as a baseline, overlay Businessman-style scaling during fiscal crunches, and retain human review for cases where fairness or legal considerations are paramount. Because the causal objects used for prioritization are stable across years, the approach supports medium-run planning rather than one-off campaigns.

Two caveats warrant attention and suggest channels for future work. First, the analysis focuses on compliance outcomes and does not fully incorporate administrative costs beyond coverage constraints; integrating cost heterogeneity would sharpen fiscal trade-offs and may further favor progressive targeting if low-value visits are cheaper to execute. Second, external validity depends on institutional details such as transfer rules and payment frictions; applying the same framework to other tax bases and jurisdictions will help map where causal heterogeneity and distributional impacts differ.

In sum, direct taxpayer engagement is an effective instrument for raising compliance, and modern policy learning can make it substantially more efficient and transparent. By aligning targeting rules with clearly stated objectives and auditing their distributional consequences, tax authorities can mobilize revenue, broaden participation, and maintain fairness in enforcement. The evidence shows that these aims need not be in conflict when design choices are explicit, evaluation is out of sample, and welfare trade-offs are quantified rather than assumed.

References

Khwaja, M., Awasthi, R., & Loeprick, J. (2011). *Risk-based tax audits: Approaches and country experiences*. World Bank Publications. https://books.google.com/books?id=bQ4mlYTNft8C

Battaglini, M., Guiso, L., Lacava, C., Miller, D. L., & Patacchini, E. (2025). Refining public policies with machine learning: The case of tax auditing. *Journal of Econometrics*, *249*, 105847. https://doi.org/10.1016/j.jeconom.2024.105847

Haseeb, M., & Vyborny, K. (2022). Data, discretion and institutional capacity: Evidence from cash transfers in pakistan. *Journal of Public Economics*, *206*, 104535. https://doi.org/10.1016/j.jpubeco.2021.104535

Knebelmann, J., Pouliquen, V., & Sarr, B. (2024). *Discretion versus algorithms: Bureaucrats and tax equity in senegal* (Working Paper). Paris School of Economics. https://www.parisschoolofeconomics eu/app/uploads/2024/05/knebelmann-justine-widening-the-tax-net-when-information-is-scarce-the-role-of-bureaucrats-discretion.pdf

Bachas, P., Brockmeyer, A., Ferreira, A., & Sarr, B. (2025). *Algorithms and bureaucrats: Evidence from tax audit selection in senegal* (No. 11205). World Bank. http://hdl.handle.net/10986/43682

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., & Ho, D. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1479–1503. https://doi.org/10.1145/3531146.3533204

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations [_eprint: https://onlinelibrary.wi 0262.2004.00530.x]. *Econometrica*, *72*(4), 1221–1246. https://doi.org/10.1111/j.1468-0262.2004.00530.x

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, *166*(1), 157–165. https://doi.org/10.1016/j.jeconom.2011.06.013

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., & Wager, S. (2025). Evaluating treatment prioritization rules via rank-weighted average treatment effects [Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2024.2393466]. *Journal of the American Statistical Association*, *120*(549), 38–51. https://doi.org/10.1080/01621459.2024.2393466

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Chernozhukov, V., Lee, S., Rosen, A. M., & Sun, L. (2025, July 9). Policy learning with confidence. https://doi.org/10.48550/arXiv.2502.10653

Athey, S., & Wager, S. (2021). Policy learning with observational data [_eprint: https://onlinelibrary.wiley.com/d *Econometrica*, *89*(1), 133–161. https://doi.org/10.3982/ECTA15732

Pomeranz, D., & Vila-Belda, J. (2019). Taking state-capacity research to the field: Insights from collaborations with tax authorities. *Annual Review of Economics*, *11*(1), 755–781. https://doi.org/10.1146/annurev-economics-080218-030312

Antinyan, A., & Asatryan, Z. (2024). Nudging for tax compliance: A meta-analysis. *The Economic Journal*, ueae088. https://doi.org/10.1093/ej/ueae088

Brockmeyer, A., Estefan, A., Arras, K. R., & Serrato, J. C. S. (2021). Taxing property in developing countries: Theory and evidence from mexico. *NBER Working Paper 28637*. https://doi.org/10.3386/w28637

Castro, L., & Scartascini, C. (2015). Tax compliance and enforcement in the pampas evidence from a field experiment. *Journal of Economic Behavior & Organization*, *116*, 65–82. https://doi.org/10.1016/j.jebo.2015.04.002

Cruces, G., Tortarolo, D., & Vazquez-Bare, G. (2025). Design of partial population experiments with an application to spillovers in tax compliance. *The Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest_a_01552

Schächtele, S., Eguino, H., & Roman, S. (2023). Fiscal exchange and tax compliance: Evidence from a field experiment [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pam.22460]. *Journal of Policy Analysis and Management*, *42*(3), 796–814. https://doi.org/10.1002/pam.22460

Del Carpio, L. (2013). *Are the neighbors cheating? evidence from a social norm experiment on property taxes in peru* (Working Paper). Princeton University. Princeton, NJ.

Pfeifer, F. F., & Pacheco, T. S. (2020). *Increasing tax compliance with behavioral insights: Evidence from são paulo* (Working paper). FGV EAESP. São Paulo, Brazil.

Schächtele, S., Eguino, H., & Roman, S. (2022). Improving taxpayer registration through nudging? field experimental evidence from brazil. *World Development*, *154*, 105887. https://doi.org/10.1016/j.worlddev.2022.105887

Chirico, M., Inman, R., Loeffler, C., MacDonald, J., & Sieg, H. (2019). Deterring property tax delinquency in philadelphia: An experimental evaluation of nudge strategies [Publisher: The University of Chicago Press]. *National Tax Journal*, *72*(3), 479–506. https://doi.org/10.17310/ntj.2019.3.01

John, P., & Blume, T. (2018). How best to nudge taxpayers? the impact of message simplification and descriptive social norms on payment rates in a central london local authority. *Journal of Behavioral Public Administration*, *1*(1). https://doi.org/10.30636/jbpa.11.10

Okunogbe, O. (2021). *Becoming legible to the state: The role of detection and enforcement capacity in tax compliance* (Policy Research Working Paper No. 9852). World Bank. Washington, D.C.

Weigel, J. L. (2020). The participation dividend of taxation: How citizens in congo engage more with the state when it tries to tax them* [_eprint: https://academic.oup.com/qje/article-pdf/135/4/1849/33668569/qjaa019.pdf]. *The Quarterly Journal of Economics*, *135*(4), 1849–1903. https://doi.org/10.1093/qje/qjaa019

Collin, M., Di Maro, V., Evans, D. K., & Manang, F. (2025). Property tax compliance in tanzania: Can nudges help? [Publisher: The University of Chicago Press]. *Economic Development and Cultural Change*, *73*(4), 2063–2103. https://doi.org/10.1086/734186

Manwaring, P., & Regan, T. (2023). *Public disclosure and tax compliance: Evidence from uganda* (Discussion Paper No. 1937). Centre for Economic Performance. London, UK.

Ortega, D., & Sanguinetti, P. (2013). Deterrence and reciprocity effects on tax compliance: Experimental evidence from venezuela [Number: 253 Publisher: CAF Development Bank Of Latinamerica]. *Research Department working papers*. Retrieved July 12, 2025, from https://ideas.repec.org//p/dbl/dblwop/253.html

Ortega, D., & Scartascini, C. (2020). Don't blame the messenger. the delivery method of a message matters. *Journal of Economic Behavior & Organization*, *170*, 286–300. https://doi.org/10.1016/j.jebo.2019.12.008

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, 14–31. https://doi.org/10.1016/j.jpubeco.2017.02.003

Boning, W. C., Hendren, N., Sprung-Keyser, B., & Stuart, E. (2024). A welfare analysis of tax audits across the income distribution* [_eprint: https://academic.oup.com/qje/article-pdf/140/1/63/61227103/qjae037.pdf]. *The Quarterly Journal of Economics*, *140*(1), 63–112. https://doi.org/10.1093/qje/qjae037

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? evidence from a tax audit experiment in denmark [_eprint: https://onlinelibrary.wiley.com/doi/p
Econometrica, *79*(3), 651–692. https://doi.org/10.3982/ECTA9113

Advani, A., Elming, W., & Shaw, J. (2023). The dynamic effects of tax audits. *The Review of Economics and Statistics*, *105*(3), 545–561. https://doi.org/10.1162/rest_a_01101

Hebous, S., Jia, Z., Løyland, K., Thoresen, T. O., & Øvrum, A. (2023). Do audits improve future tax compliance in the absence of penalties? evidence from random audits in norway. *Journal of Economic Behavior & Organization*, *207*, 305–326. https://doi.org/10.1016/j.jebo.2023.01.001

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls†. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference [Publisher: ASA Website _eprint: https://doi.org/10.1198/jcgs.2010.08162]. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion) [Publisher: International Society for Bayesian Analysis]. *Bayesian Analysis*, *15*(3), 965–1056. https://doi.org/10.1214/19-BA1195

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms [ISSN: 2640-3498]. *Proceedings of the 34th International Conference on Machine Learning*, 3076–3085. Retrieved November 2, 2025, from https://proceedings.mlr.press/v70/shalit17a.html

Parikh, H., Rudin, C., & Volfovsky, A. (2022). MALTS: Matching after learning to stretch. *Journal of Machine Learning Research*, *23*(240), 1–42. Retrieved November 2, 2025, from http://jmlr.org/papers/v23/21-0053.html

Caron, A., Baio, G., & Manolopoulou, I. (2022). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 1115–1149. https://doi.org/10.1111/rssa.12824

Robinson, P. M. (1988). Root-n-consistent semiparametric regression [Publisher: [Wiley, Econometric Society]]. *Econometrica*, *56*(4), 931–954. https://doi.org/10.2307/1912705

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters [_eprint: https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf]. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Semenova, V., & Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions [_eprint: https://academic.oup.com/ectj/article-pdf/24/2/264/46748086/utaa027.pdf]. *The Econometrics Journal*, *24*(2), 264–289. https://doi.org/10.1093/ectj/utaa027

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application [Publisher: University of Pennsylvania Press]. *Observational Studies*, *5*(2), 37–51. Retrieved October 14, 2025, from https://muse.jhu.edu/pub/56/article/793356

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., & Wright, M. (2025, October 9). *Grf: Generalized random forests* (Version 2.5.0). Retrieved October 14, 2025, from https://cran.r-project.org/web/packages/grf/index.html

Chernozhukov, V., Demirer, M., Duflo, E., & Fernández-Val, I. (2025). Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india [_eprint: https://onlinelibrary.wiley.com/do *Econometrica*, *93*(4), 1121–1164. https://doi.org/10.3982/ECTA19303

Kitagawa, T., & Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA13288]. *Econometrica*, *86*(2), 591–616. https://doi.org/10.3982/ECTA13288

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, *1*(3), 323–338. https://doi.org/10.1016/0047-2727(72)90010-2

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature*, *57*(4), 904–954. https://doi.org/10.1257/jel.20181437

Wager, S. (2024). Causal inference: A statistical learning approach. Retrieved October 17, 2025, from https://web.stanford.edu/~swager/causal_inf_book.pdf

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander (Ed.), *Probability and statistics* (pp. 416–444). John Wiley.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed [Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.1 *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818

A The Property Tax in Kota Gorontalo - Pajak Bumi dan Bangunan (PBB)

The Property Tax (Pajak Bumi dan Bangunan, or PBB) is a tax levied on the ownership or use of land and/or buildings in Kota Gorontalo, Indonesia. Managed by the local government, this tax is regulated under local property taxation laws. It is the main source of revenue for the city, supporting public services and infrastructure.

A.1. PBB Calculation

The calculation of PBB is based on the taxable value of the land and/or buildings. The value of the property is assessed and compared with the Non-Taxable Value of Property to determine the taxable amount. The formula for calculating the PBB is as follows:

$$PBB_{itp} = (NJOP_{it} - NJOPTKP_{it}) \times Rate_{it} + Fine_{itp}$$

Where:

- PBB_{itp} is the property tax for property i of owner j in neighborhood n for fiscal year n and payment year p.
- $NJOP_{it}$ is the taxable sale value of the property (land and/or building).
- $NJOPTKP_{it}$ is the Non-Taxable Value of Property.
- $Rate_{it}$ is the applicable tax rate for the property.
- $Fine_{itp}$ is any fine or penalty for late payment.

A.2. Components of the Calculation

A.2.1. NJOP (Nilai Jual Objek Pajak)

NJOP refers to the taxable sale value of the land and/or building, as determined by the local government. The final NJOP is calculated by combining the assessed values of the land (NJT) and the building (NJB).

- **Land Value Assessment (NJT - Nilai Jual Tanah)**: The value of the land is calculated by multiplying the estimated value of the land per square meter by the total size of the land. The land class is assigned based on 27 categories, which are used to determine the value.
- **Building Value Assessment (NJB - Nilai Jual Bangunan)**: The value of the building is calculated by multiplying the estimated value of the building per square meter by the total area of the construction. The building class is assigned based on 38 categories with prices, and one category that sets prices = 1 if there are no buildings on the land.

A.2.2. NJOPTKP (Nilai Jual Objek Pajak Tidak Kena Pajak)

The Non-Taxable Value of Property (NJOPTKP) is the portion of a property's value that is exempt from taxation. If the property's value is below the NJOPTKP threshold, it will not be subject to PBB.

A.2.3. NJKP (Nilai Jual Kena Pajak)

NJKP represents the taxable base used for calculating the property tax. It is a percentage of the NJOP, which is determined based on local regulations. This value is subtracted by the NJOPTKP and multiplied by the tax rate. There are two tax rates: 0.1% if the property value is lower than 1000000000 or 0.2% if the property value is greater than 1000000000.

B The causal forest

We provide a more rigorous description of the causal forest estimator and our debiasing approach. Recall that we observe n i.i.d. samples $W_i = (Y_i, Z_i, X_i)$, where $Z_i \in 0, 1$ is a treatment indicator, $X_i \in \mathbb{R}^p$ is a feature vector, and Y_i is the outcome. We are interested in the conditional average treatment effect (CATE) function $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$. We assume the standard potential outcomes framework and a structural semi-parametric model:

$$Y_i = \mu(X_i) + \tau(X_i)\, Z_i + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X_i, Z_i] = 0, \tag{17}$$

where $\mu(x) := \mathbb{E}[Y_i \mid X_i = x, Z_i = 0]$ is the baseline outcome (the outcome one would expect without treatment). Define also the unconditional mean function $m(x) := \mathbb{E}[Y_i \mid X_i = x]$ and the propensity score $e(x) := \Pr(Z_i = 1 \mid X_i = x) = \mathbb{E}[Z_i \mid X_i = x]$. Robinson's transformation Robinson (1988) is applied to eliminate the nuisance $\mu(x)$ from (17). Taking the expectation of (17) conditional on $X_i = x$ and subtracting it from (17), we obtain the orthogonalized model:

$$Y_i - m(X_i) \;=\; \big(Z_i - e(X_i)\big)\, \tau(X_i) \;+\; \varepsilon_i. \tag{18}$$

This equation has the form of a weighted residual regression: the left-hand side is the residualized outcome and the right-hand side is the residualized treatment indicator multiplied by the target function $\tau(X_i)$. Importantly, $\mathbb{E}[Z_i - e(X_i) \mid X_i] = 0$ by construction, which ensures that the regressor $(Z_i - e(X_i))$ is orthogonal (uncorrelated) to the noise ε_i and any estimation errors in $e(\cdot)$.

Equation (18) suggests an objective for estimating $\tau(\cdot)$ known as the R-learner – Robinson's or Residual learner– Nie and Wager (2021). The population target is to find a function $\tau(x)$ that best fits the relationship $\mathbb{E}[Y - m(X) \mid X = x] = \mathbb{E}[Z - e(X) \mid X = x]\, \tau(x)$. In practice, we replace the unknown nuisances $m(\cdot)$ and $e(\cdot)$ with estimates $\hat{m}(\cdot)$ and $\hat{e}(\cdot)$. To avoid overfitting and achieve unbiasedness, we use cross-fitting Chernozhukov et al. (2018): we split the sample into K folds and for each observation i, we compute $\hat{m}^{(-i)}(X_i)$ and $\hat{e}^{(-i)}(X_i)$ using only the data in the other $K - 1$ folds (excluding observation i). The R-learner then solves:

$$\begin{aligned} &\hat{\tau}(.) = \operatorname{argmin}_{\tau} \left[\widehat{L}_n\{\tau(\cdot)\} + \Lambda_n\{\tau(\cdot)\}\right] \\ &\widehat{L}_n\{\tau(\cdot)\} = \frac{1}{n} \sum_{i=1}^{n} \left[\left\{Y_i - \hat{m}^{(-i)}\left(X_i\right)\right\} - \left\{Z_i - \hat{e}^{(-i)}\left(X_i\right)\right\} \tau\left(X_i\right)\right]^2. \end{aligned} \tag{19}$$

Where the term $\Lambda_n(\tau(\cdot))$ is interpreted as a regularizer on the complexity of the $\tau(\cdot)$ function.

This regularization could be explicit as in penalized regression, or implicit as with generalized random forests. This is a quadratic loss minimization whose solution yields the estimated CATE function $\hat{\tau}(x)$. Because of the orthogonal re-weighting, this estimator has desirable statistical properties: even if $\hat{m}$ and $\hat{e}$ are nonparametric estimates (with slower convergence rates), the orthogonal (or Neyman-orthogonal) moment condition implicit in (18) ensures that $\hat{\tau}(x)$ can achieve $\sqrt{n}$-consistency and asymptotic normality Chernozhukov et al. (2018) Nie and Wager (2021).

The causal forest estimator Wager and Athey (2018) implements the R-learner solution in a flexible, data-driven way. It is a special case of the generalized random forest (GRF) framework of Athey et al. (2019), designed specifically for treatment effect estimation. A causal forest consists of many randomized decision trees that partition the feature space and estimate a constant treatment effect within each leaf. The trees are grown in an honest manner Athey and Imbens (2016): half of the data (or a designated portion) is used to decide on the splits (structure of the tree), and the other half is used to estimate the leaf-wise treatment effects. Honesty guarantees that the same data used to estimate τ is not used to determine the partition, which prevents overfitting τ to idiosyncratic noise. Each tree yields a noisy estimate of $\tau(x)$ by looking at the difference in mean outcomes between treated and control units in the leaf containing x, after adjusting for nuisance estimates. The forest averages these estimates over many trees to stabilize the prediction. Formally, the causal forest can be viewed as producing a weighted nearest-neighbor estimator for $\tau(x)$:

$$\hat{\tau}(x) = \sum_{i=1}^{n} \alpha_i(x)\Gamma_i,$$

where the weights $\alpha_i(x)$ depend on how often training point i falls in the same leaves as target point x across all trees (and they satisfy $\alpha_i(x) \geq 0$, $\sum_i \alpha_i(x) = 1$), and Γ_i is the orthogonalized pseudo-outcome for observation i. The pseudo-outcome Γ_i is constructed to be an unbiased proxy for the individual treatment effect, using the orthogonal moment equation. Given estimated nuisance functions $(\hat{\mu}, \hat{e})$ from cross-fitting, one convenient form for Γ_i is:

$$\Gamma_i := \frac{Z_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left[1 - \hat{e}^{(-i)}(X_i)\right]}\Big(Y_i - \hat{\mu}^{(-i)}(X_i, Z_i)\Big) + \hat{\mu}^{(-i)}(X_i, 1) - \hat{\mu}^{(-i)}(X_i, 0).$$

Here $\hat{\mu}(x,z) = [Y_i \mid X_i = x, Z_i = z]$ is an estimate of the conditional outcome function (the regression of Y on X separately for each treatment arm). Intuitively, the term in parentheses $Y_i - \hat{\mu}^{(-i)}(X_i, Z_i)$ is the outcome residual for unit i (actual outcome minus predicted outcome given its treatment status), and it is scaled by $\frac{Z_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)(1-\hat{e}^{(-i)}(X_i))}$ which is an inverse propensity weight (positive for treated units, negative for control units) normalized by the variance of Z_i. This scaled residual represents an uplift or advantage that unit i experienced relative to the expected outcome, attributable to treatment assignment. Adding $\hat{\mu}^{(-i)}(X_i, 1) - \hat{\mu}^{(-i)}(X_i, 0)$ (the estimated conditional mean difference for someone with X_i) then shifts this residual to approximate the individual treatment effect. By construction $\mathbb{E}[\Gamma_i \mid X_i] \approx \tau(X_i)$, so Γ_i can be used as the outcome in a regression forest to predict $\tau(x)$. Wager and Athey (2018) prove that under regularity conditions, $\hat{\tau}(x)$ from a causal forest is a consistent estimator of $\tau(x)$ and that certain averages of $\hat{\tau}(x)$ (e.g. the average treatment effect) are asymptotically normal with an estimable variance.

C Debiasing Procedure with Bias-Corrected Outcomes

We now detail the bias-correction procedure introduced in the main text. The goal is to remove bias due to non-random treatment assignment by using pre-intervention data. Consider the year 2018 as a pre-treatment period (no program in effect). Let $X_i^{(17)}$ be the covariate history up to 2017. We assume that any difference in 2018 outcomes between treatment and control groups is due to selection bias. Formally, we write a reduced-form model for 2018 as:

$$Y_i^{(18)} = \big(\underbrace{\tau^{(18)}(X_i)}_{:=0} + B(X_i^{(17)})\big)Z_i + f(X_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid X_i, Z_i] = 0. \tag{20}$$

Here $B(X_i^{(17)})$ is the bias function and $f(X_i) = \mathbb{E}[Y_i^{(18)} \mid X_i]$ is the baseline outcome in 2018 (had the unit not been treated, analogous to $\mu(X_i)$ but for that year). Because the treatment has no effect in 2018 (program not yet started), the true CATE in 2018 is zero: $\tau^{(18)}(X) = 0$. Thus, in equation (20), the term $B(X_i^{(17)})Z_i$ represents the spurious effect of Z on Y—i.e. confounding bias. We can identify $B(x)$ by essentially estimating the "treatment effect" in the 2018 data, which should purely reflect bias. In particular, one can show that

$$B(X_i^{(17)}) = \mathbb{E}[Y_i^{(18)} \mid X_i^{(17)}, Z_i = 1] - \mathbb{E}[Y_i^{(18)} \mid X_i^{(17)}, Z_i = 0].$$

We use a causal forest on 2018 data to estimate this difference. Let $\hat{\mu}^{(-i)}(x, z)$ be the cross-fitted causal forest estimate of $\mathbb{E}[Y^{(18)} \mid X^{(17)} = x, Z = z]$. Then the estimated bias function is

$$\hat{B}(x) \;=\; \hat{\mu}^{(-i)}(x, 1) - \hat{\mu}^{(-i)}(x, 0), \tag{21}$$

for $x = X_i^{(17)}$. In other words, we predict each unit's 2018 outcome as if treated and as if untreated, and take the difference. This $\hat{B}(x)$ captures how much higher (or lower) the treated unit's outcome is expected to be relative to the control unit's, given the same covariates. We assume the structure of this selection bias $B(x)$ remains the same in subsequent years. This is a time stability assumption: although the levels of outcomes change after the intervention, the bias attributable to non-random selection (as a function of X) is similar to what it was just before the intervention. This assumption cannot be tested directly, but it is conceptually analogous to assuming parallel trends in difference-in-differences, except here on a conditional level. Using $\hat{B}(X_i^{(17)})$, we construct debiased outcomes for each post-intervention year $t \geq 2019$:

$$\tilde{Y}_i^{(t)} := Y_i^{(t)} - \hat{B}(X_i^{(17)})\, Z_i. \tag{22}$$

If our bias estimate is accurate, $\tilde{Y}_i^{(t)}$ should behave as if treatment was randomly assigned (conditional on covariates). Plugging $\tilde{Y}_i^{(t)}$ into the structural equation (17) for year t, we get a bias-adjusted model:

$$\tilde{Y}_i^{(t)} = f(X_i) + \tau^{(t)}(X_i)\, Z_i + \tilde{\varepsilon}_i, \qquad \mathbb{E}[\tilde{\varepsilon}_i \mid X_i, Z_i] = 0, \tag{23}$$

where $\tilde{\varepsilon}_i = \varepsilon_i - [B(X_i^{(17)}) - \hat{B}(X_i^{(17)})]\, Z_i$. Now, by construction, the treatment Z_i is mean-independent of the adjusted error term $\tilde{\varepsilon}_i$ given X_i. In other words, conditional unconfoundedness holds for $(\tilde{Y}^{(t)}, Z)$ given X. We can therefore apply the causal forest estimator in this adjusted setting without bias.

Given estimated nuisance functions $(\hat{\mu}, \hat{e}, \hat{B})$ from cross-fitting, one convenient form for Γ_i is:

$$\Gamma_i := \frac{Z_i - \hat{e}^{(-i)}(X_i)}{\hat{e}^{(-i)}(X_i)\left[1 - \hat{e}^{(-i)}(X_i)\right]} \left(Y_i^{(t)} - \hat{B}^{(-i)}(X_i^{(17)}) - \hat{\mu}^{(-i)}(X_i, Z_i)\right) + \hat{\mu}^{(-i)}(X_i, 1) - \hat{\mu}^{(-i)}(X_i, 0). \tag{24}$$

Lastly, let where the forest weights $\alpha_i(X_i)$ implicitly define a local neighborhood around each evaluation point x:

$$\hat{\tau}(X_i) = \sum_{i=1}^{n} \alpha_i(X_i)\, \Gamma_i,$$

It is worth noting a placebo test: if the bias-correction is successful, then in a year with no treatment effect, the causal forest should estimate no effect. Indeed, we verify that $\hat{\tau}^{(2019)}(x) \approx 0$ for all x (2019 being a year before the intervention took effect in our context). For years $t \geq 2021$ (after the program's onset), we find $\hat{\tau}^{(t)}(x)$ significantly above zero for many covariate profiles, consistent with a positive program impact emerging. Because we employ honest forests and cross-fitting at every stage (both in estimating $\hat{B}(x)$ and in estimating $\tau(x)$), these $\hat{\tau}^{(t)}(x)$ estimates are asymptotically unbiased for the true CATEs. Furthermore, the inference methods for causal forests Wager and Athey (2018) remain valid here, since the bias correction is itself based on a preliminary cross-fitted forest. Essentially, we treat $\hat{B}(x)$ as a first-stage estimate and incorporate it into the outcome; the orthogonalization technique described next ensures that any estimation error in $\hat{B}(x)$ does not compromise the consistency or coverage of $\hat{\tau}(x)$.

D Orthogonal Score Function with Bias Adjustment (AIPW)

To formally justify the use of machine learning estimates for nuisance functions (propensity, outcome regressions, bias function) while still performing valid inference on $\tau(x)$, we invoke the framework of orthogonal score functions Chernozhukov et al. (2018) and Neyman (1959) Semenova and Chernozhukov (2020). A score function $\psi(W;\eta)$, depending on data $W = (Y, Z, X)$ and nuisance parameters η, is Neyman-orthogonal if it is insensitive to small perturbations in η at the true parameter values. In other words, letting η_0 denote the true nuisance functions, orthogonality means:

$$\frac{\partial}{\partial t}\mathbb{E}\left[\psi\{\eta_0 + t(\eta - \eta_0)\}(W)\right]\Big|_{t=0} = 0, \quad \text{for all perturbations } \eta.$$

This property ensures that plugging in consistent estimates $\hat{\eta}$ for the nuisances will not spoil the first-order accuracy of the score. Thus, one can use flexible or high-dimensional models for η (machine learning) and still rely on ψ for inference on the target parameter. In our setting, the parameter of interest is the CATE function $\tau(x)$, and the nuisance components are $\eta = (\mu_B, e, B)$, defined as follows. We have already defined the bias-adjusted outcome $\tilde{Y} = Y - B(X)Z$. Correspondingly, define the bias-adjusted outcome regression

$$\mu_B(x,z) := \mathbb{E}[\tilde{Y} \mid X = x, Z = z] = \mathbb{E}[Y - B(X)Z \mid X = x, Z = z].$$

The target is the CATE functional

$$\tau_0(x) = \mu_{B,0}(x,1) - \mu_{B,0}(x,0) \tag{25}$$

And the propensity score is:

$$\psi_{AIPW}(W;\eta) := \left\{\mu_B(X,1) - \mu_B(X,0) - \tau(X)\right\} + \frac{Z - e(X)}{e(X)\{1 - e(X)\}}\left\{[Y - B(X)Z] - \mu_B(X,Z)\right\} \tag{26}$$

Let $t \in \mathbb{R}$ and consider small perturbations of the nuisance components:

$$\mu_B \mapsto \mu_B + t\,h_\mu, \qquad e \mapsto e + t\,h_e, \qquad B \mapsto B + t\,h_B,$$

where $h_\mu(X,Z)$, $h_e(X)$, and $h_B(X)$ are arbitrary direction functions. Define the population moment:

$$\Psi(t) := \mathbb{E}\big[\psi_{\text{AIPW}}\big(W;\tau,\mu_B + th_\mu, e + th_e, B + th_B\big)\big], \quad \dot{\Psi}(0) := \frac{\partial}{\partial t}\Psi(t)\bigg|_{t=0}.$$

Because $\mu_B(X,Z) = \mathbb{E}[Y - B(X)Z \mid X,Z]$, a perturbation h_B in B induces a perturbation in μ_B:

$$h_\mu^{\text{ind}}(X,Z) := \frac{\partial}{\partial t}\mu_{B+th_B}(X,Z)\bigg|_{t=0} = -\,Z\,h_B(X).$$

We decompose total μ_B perturbations as

$$h_\mu(X,Z) = h_\mu^{\text{ind}}(X,Z) + \bar{h}_\mu(X,Z) = -Z\,h_B(X) + \bar{h}_\mu(X,Z), \tag{27}$$

where $\bar{h}_\mu$ denotes the *free* (non-induced) component, i.e., variation in μ_B not mechanically linked to B.

Decompose the score as:

$$\psi_{\text{AIPW}} = A(X) + g(e(X),Z)\,R(W;\mu_B,B),$$

where

$$A(X) := \mu_B(X,1) - \mu_B(X,0) - \tau(X), \qquad g(e,Z) := \frac{Z-e}{e(1-e)}, \qquad R(W;\mu_B,B) := Y - B(X)Z - \mu_B(X,Z).$$

Perturbation with respect to μ_B

The derivative of $A(X)$ with respect to μ_B is immediate:

$$\frac{\partial}{\partial t}A(X)\bigg|_{t=0} = h_\mu(X,1) - h_\mu(X,0). \tag{28}$$

Next, differentiate the residual term (holding e and B fixed):

$$\frac{\partial}{\partial t}\bigg|_0 \big[g(e_0(X),Z)R(W;\mu_B + th_\mu, B)\big] = -\,g(e_0(X),Z)\,h_\mu(X,Z).$$

Taking conditional expectations given X, and using $\mathbb{E}[g(e_0,Z)\,Z \mid X] = 1$ and $\mathbb{E}[g(e_0,Z) \mid X] = 0$, we get

$$\mathbb{E}\big[-g(e_0,Z)\,h_\mu(X,Z) \mid X\big] = -\,[h_\mu(X,1) - h_\mu(X,0)]. \tag{29}$$

Hence, combining with (28),

$$\mathbb{E}\left[\left.\frac{\partial}{\partial t}\psi_{\mathrm{AIPW}}\right|_{t=0} \,\middle|\, X\right] = [h_\mu(X,1) - h_\mu(X,0)] - [h_\mu(X,1) - h_\mu(X,0)] = 0.$$

Thus the AIPW score is orthogonal with respect to μ_B perturbations.

Perturbation with respect to e

Now vary only the propensity score: $e_t = e_0 + t\,h_e$. The only dependence is through $g(e_t,Z)$:

$$\left.\frac{\partial}{\partial t}\right|_0 g(e_t(X),Z) = \left.\frac{\partial g}{\partial e}\right|_{e_0} h_e(X), \qquad \text{where} \quad \frac{\partial g}{\partial e} = \frac{\partial}{\partial e}\frac{Z-e}{e(1-e)}.$$

Hence,

$$\left.\frac{\partial}{\partial t}\right|_0 [g(e_t(X),Z)R(W;\mu_{B,0},B_0)] = \left.\frac{\partial g}{\partial e}\right|_{e_0} h_e(X)\,R(W;\mu_{B,0},B_0). \tag{30}$$

Because $\mathbb{E}[R(W;\mu_{B,0},B_0) \mid X] = 0$, it follows that

$$\mathbb{E}\left[\left.\frac{\partial}{\partial t}\psi_{\mathrm{AIPW}}\right|_{t=0} \,\middle|\, X\right] = 0. \tag{31}$$

Thus, the score is orthogonal with respect to $e(X)$ perturbations.

Perturbation with respect to B

Finally, let $B_t = B_0 + t\,h_B$ while keeping μ_B implicitly linked via its definition. Under a pure B–perturbation, we embed μ_B coherently in B via $\mu_{B_t}(X,Z) = \mathbb{E}[Y - (B_0 + t\,h_B)(X)Z \mid X,Z]$, so the μ_B drift decomposes as $h_\mu(X,Z) = h_\mu^{\mathrm{ind}}(X,Z) + \bar{h}_\mu(X,Z) = -Z\,h_B(X) + \bar{h}_\mu(X,Z)$ where $\bar{h}_\mu$ is the free (non-induced) component.

To avoid double counting, we (i) assign the induced part $h_\mu^{\mathrm{ind}} = -Z\,h_B$ to the residual, and (ii) let $A(X)$ depend only on the free component $\bar{h}_\mu$.

Hence the $A(X)$ contribution is

$$\frac{\partial}{\partial t} A(X)\bigg|_{t=0} = \bar{h}_\mu(X,1) - \bar{h}_\mu(X,0).$$

Under a pure B–perturbation we set $\bar{h}_\mu \equiv 0$, so this term is 0. Then the residual becomes

$$R(W;\mu_{B_t},B_t) = Y - (B_0 + t\,h_B)Z - \mu_{B_t}(X,Z).$$

Differentiating yields

$$\frac{\partial}{\partial t} R(W;\mu_{B_t},B_t)\bigg|_{t=0} = -\,Z\,h_B(X) - h_\mu^{\text{ind}}(X,Z),$$

and from (27) we have $h_\mu^{\text{ind}}(X,Z) = -Z\,h_B(X)$, so that

$$-\,Z\,h_B(X) - h_\mu^{\text{ind}}(X,Z) = -\,Z\,h_B(X) + Z\,h_B(X) = 0.$$

confirming orthogonality with respect to $B(X)$ as well.

E Asymptotic Normality - Temporary

This section is temporary. It outlines the main steps for establishing the asymptotic normality of our debiased estimator, following the framework of Chernozhukov et al. (2018) and Wager (2024). We first summarize the identifying assumptions, then present the oracle version of our debiased estimator. Lastly, we discuss the impact of first-stage machine-learning estimation and the role of conversion rates of the nuisance estimators.

F Best Linear Projection of Heterogeneous Treatment Effects

An intuitive summary of the estimated Conditional Average Treatment Effects (CATEs) is obtained through the *Best Linear Projection* (BLP). The BLP provides a doubly robust and interpretable linear approximation of the conditional treatment effect function:

$$\tau(X_i) = \beta_0 + A_i'\beta + u_i, \tag{32}$$

where A_i is a vector of pre-specified covariates (e.g., those with highest variable importance in the causal forest or any other set) and u_i is a mean-zero residual. The coefficients β capture how average treatment effects vary linearly with selected observable characteristics.

Simply regressing the forest-based estimates $\hat{\tau}_i$ on A_i using ordinary least squares (OLS) is inconsistent, since $\hat{\tau}_i$ is estimated with error and correlated with first-stage residuals Chernozhukov, Demirer, et al. (2025). To address this problem, Semenova and Chernozhukov (2020) propose a procedure that constructs an orthogonal estimating equation ensuring robustness to regularization bias in high-dimensional first-stage estimates.

Orthogonal signal construction: Let $\psi(\eta)$ denote a signal that depends on nuisance parameters η, such as the outcome regression and propensity score. A score is said to be *Neyman-orthogonal* **neyman1959<empty citation>** if small perturbations in η do not affect the conditional expectation of the score at the truth:

$$\frac{\partial}{\partial t}\mathbb{E}\big[\,\psi(\eta_0 + t(\eta - \eta_0)) \mid W = w\big]\bigg|_{t=0} = 0, \quad \forall\, w, \eta. \tag{33}$$

This orthogonality ensures that plug-in estimates $(\hat{\eta})$ remain insensitive to first-stage estimation errors, allowing the use of machine learning methods for η while retaining valid inference for the target parameter $g(x) = \mathbb{E}[Y(\eta_0) \mid X = x]$.

G Rank-Weighted Average Treatment Effect - RATE

This appendix formalizes the Rank-Weighted Average Treatment Effect (RATE) used to evaluate treatment prioritization rules in settings with heterogeneous effects Yadlowsky et al. (2025). RATE provides a scalar measure of how effectively a score concentrates treatment on units with larger causal payoffs, aggregating performance across all feasible program sizes. The framework is model-agnostic: it applies to any prioritization score, including estimated conditional average treatment effects (CATE), predicted risks, or composite indices.

Setup and notation

Let $Y_i(1)$ and $Y_i(0)$ denote potential outcomes for unit i, and let $\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]$ be the CATE. A prioritization rule maps observed covariates X_i to a score $S(X_i) \in \mathbb{R}$, with larger scores intended to indicate larger payoffs from treatment. For a coverage rate $q \in (0,1]$, define the ranking threshold $t_q = F_S^{-1}(1-q)$, where F_S is the marginal distribution of $S(X_i)$, and the induced policy

$$\pi_q(X_i) = \mathbb{1}\{S(X_i) \geq t_q\}.$$

In many applications it is useful to evaluate targeting on different margins by introducing nonnegative weights Γ_i. Two margins are central in this paper: deterrence, which sets $\Gamma_i \equiv 1$, and revenue, which weights by a monetary scale such as tax liability or expected payment, $\Gamma_i = W_i$.

Targeting Operator Characteristic (TOC)

The Targeting Operator Characteristic (TOC) curve measures how the average weighted treatment effect among the top-ranked q fraction compares to the population mean:

$$\mathrm{TOC}(q;\Gamma,S) = \mathbb{E}\left[\Gamma_i\,\tau(X_i) \mid S(X_i) \geq t_q\right] - \mathbb{E}\left[\Gamma_i\,\tau(X_i)\right]. \tag{34}$$

If $\mathrm{TOC}(q;\Gamma,S) > 0$ for small q, the score is successfully concentrating treatment on units with above-average effects for the evaluation margin encoded by Γ_i. The TOC is invariant to strictly monotone transformations of S, depends only on the rank ordering, and is conceptually analogous to the ROC curve in classification.

RATE as area under the TOC

RATE aggregates performance across all coverage rates as the area under the TOC curve:

$$\mathrm{RATE}(\Gamma,S) = \int_0^1 \mathrm{TOC}(q;\Gamma,S)\,dq. \tag{35}$$

RATE equals zero when effects are homogeneous or when the score is independent of $\tau(X_i)$ on the evaluation margin. Positive values indicate that the score ranks units by larger payoffs; negative values indicate perverse ranking. Alternative summaries reweight the TOC. A prominent example is the Qini coefficient **radcliffe_2007<empty citation>**,

$$\mathrm{QINI}(\Gamma, S) = \int_0^1 q\,\mathrm{TOC}(q; \Gamma, S)\,dq,$$

which emphasizes broader, diffuse gains rather than extreme upper-tail gains.

Estimation with sample splitting and doubly robust scores

In practice $\tau(X_i)$ is unknown. We estimate TOC and RATE using out-of-sample rankings and doubly robust scores. Split the data into training and test sets. On the training set, estimate a CATE function $\hat{\tau}^{\mathrm{train}}(x)$ (e.g., a causal forest) and any nuisance functions needed for doubly robust estimation on the test set: the propensity score $e(x) = \Pr(W = 1 \mid X = x)$ and outcome regressions $m_w(x) = \mathbb{E}[Y \mid W = w, X = x]$, $w \in \{0, 1\}$. Construct for each test observation the AIPW pseudo-outcome

$$\phi_i \;=\; m_1(X_i) - m_0(X_i) \;+\; \frac{W_i}{e(X_i)}\big\{Y_i - m_1(X_i)\big\} \;-\; \frac{1 - W_i}{1 - e(X_i)}\big\{Y_i - m_0(X_i)\big\}. \tag{36}$$

This pseudo-outcome is unbiased for $\tau(X_i)$ if either $e(\cdot)$ or $(m_0(\cdot), m_1(\cdot))$ is correctly specified, and it admits cross-fitting for high-dimensional learners. Ranking on the test set is performed using $\hat{\tau}^{\mathrm{train}}(X_i^{\mathrm{test}})$ to avoid overfitting.

For a given q, let $\mathscr{I}_q = \{i \in \mathrm{test} : \hat{\tau}^{\mathrm{train}}(X_i) \geq \hat{t}_q\}$ be the index set of top-ranked test observations, with $\hat{t}_q$ the empirical $(1-q)$ quantile of $\hat{\tau}^{\mathrm{train}}(X)$ on the test set. The sample analogue of (34) is

$$\widehat{\mathrm{TOC}}(q; \Gamma, S) = \frac{\sum_{i \in \mathscr{I}_q} \Gamma_i \phi_i}{\sum_{i \in \mathscr{I}_q} \Gamma_i} - \frac{\sum_{i \in \mathrm{test}} \Gamma_i \phi_i}{\sum_{i \in \mathrm{test}} \Gamma_i}. \tag{37}$$

The RATE is then estimated by numerical integration over a grid $\mathscr{Q} \subset (0, 1]$,

$$\widehat{\mathrm{RATE}}(\Gamma, S) = \sum_{q \in \mathscr{Q}} w(q)\,\widehat{\mathrm{TOC}}(q; \Gamma, S), \tag{38}$$

with nonnegative weights $w(q)$ that approximate the unit integral (for instance, equal weights on an equally spaced grid). The same construction yields $\widehat{\mathrm{QINI}}$ by replacing $w(q)$ with weights proportional to q.

Inference and implementation details

Under standard regularity conditions, sample splitting, and cross-fitting, $\widehat{\text{TOC}}(q;\Gamma,S)$ is asymptotically linear for each fixed q, and $\widehat{\text{RATE}}(\Gamma,S)$ is $\sqrt{n}$-consistent and asymptotically normal Yadlowsky et al. (2025). Pointwise standard errors can be obtained from influence-function estimates or a nonparametric bootstrap on the test set that respects the ranking induced by $\hat{\tau}^{\text{train}}(X)$. Uniform bands for the TOC curve can be constructed using a multiplier bootstrap, and propagated to RATE via the numerical integration in (38).

Several practical considerations improve finite-sample performance. Ties in $\hat{\tau}^{\text{train}}(X)$ can be handled via mid-ranks; extreme q values can be trimmed to limit sensitivity to very small treated sets; monotone transformations of S leave the TOC and RATE unchanged; and all results can be reported for multiple evaluation margins by choosing Γ_i to reflect deterrence or revenue.

Interpretation for policy analysis

RATE provides a direct mapping from estimated heterogeneity to implementable targeting gains. A steep TOC for small q and a large positive RATE indicate that a small share of top-ranked units accounts for a disproportionate share of the potential payoff, making targeted deployment of scarce enforcement capacity attractive. A flat TOC and a RATE near zero suggest limited benefits from prioritization relative to random assignment. Because the evaluation margin is encoded by Γ_i, the same score can be assessed for both compliance (deterrence) and monetary outcomes (revenue) without re-estimating the ranking rule.

In our application to property-tax enforcement, the TOC and RATE quantify how much additional compliance or revenue could be achieved by ranking properties according to predicted treatment effects, compared with uniform assignment. These statistics connect the heterogeneous effects uncovered by the causal forest to the operational question of whom to visit first, and they do so with transparent uncertainty quantification compatible with cross-fitting and modern machine learning.

H Concentration Curves and the Concentration Index

A concentration curve provides a graphical representation of how a continuous outcome variable—such as income, tax payments, or health expenditure—is distributed across the population when individuals are ranked according to another variable of interest, such as a policy priority score or socioeconomic status. The associated concentration index (CI) summarizes this relationship in a single scalar that measures the degree of association between the outcome and the ranking variable.

The Lorenz curve is a special case of the concentration curve, corresponding to the situation in which individuals are ranked by their own outcome variable. The Lorenz curve always lies below or on the 45° line ($L = P$). The closer the curve is to this diagonal, the more equal the distribution of the outcome. Perfect equality implies $L(P) = P$, while perfect inequality—where a single individual holds all income—results in a curve that coincides with the axes and jumps to $(1, 1)$. The Gini coefficient is the concentration index (CI) for the area between the Lorenz curve and the 45° line.

The value of CI for the general concentration curve lies within the theoretical bounds $[-1, 1]$, where the sign indicates the direction of concentration and the magnitude reflects its strength. A value of CI = 1 represents perfect positive concentration: all of the outcome (e.g., income or payments) is held by the individual or group with the highest rank in the ordering variable s_i. In this limiting case, the concentration curve follows the horizontal and right-hand axes before jumping vertically to $(1, 1)$, indicating complete inequality in favor of those with the highest s_i (for instance, all income accruing to the richest individual). Conversely, CI = -1 corresponds to perfect negative concentration: all of the outcome is held by the individual or group with the lowest rank in s_i. The curve then follows the left-hand and upper axes before dropping vertically to $(1, 1)$, indicating complete inequality in favor of those with the lowest s_i (for example, all benefits concentrated among the poorest).

In practice, empirical estimates of CI rarely approach these theoretical extremes, since real-world distributions typically display partial dispersion across ranks.In our policy context, however, it is useful to identify conceptual benchmarks that define the feasible range within which policy can operate. At one extreme lies a policy that concentrates all attention or resources on the richest individuals—the Most Regressive Policy—while at the other lies a policy that fully prioritizes the poorest individuals—the Most Progressive Policy. These two hypothetical allocations serve as natural bounds for evaluating where an observed targeting rule falls along the equity–efficiency spectrum. To formalize this comparison, we introduce the Policy Concentration Index (PCI).

To formalize this comparison, we introduce the Policy Concentration Index (PCI), which benchmarks any empirical prioritization rule against two conceptual extremes: the most progressive and most regressive possible allocations, corresponding respectively to PCI = -1 and PCI = 1. The policy space is thus bounded by the area between these two limiting curves.

We normalize each observed concentration index relative to these bounds.

$$\text{PCI} = \frac{\text{CI} - \text{CI}_{\text{min}}}{\text{CI}_{\text{max}} - \text{CI}_{\text{min}}},$$

where $\text{CI}_{\text{min}} = -1$ and $\text{CI}_{\text{max}} = 1$. Under this normalization, PCI = −1 corresponds to full prioritization of the poorest (maximally progressive) allocation, while PCI = 1 represents full prioritization of the richest (maximally regressive) allocation. A positive PCI indicates regressive targeting—favoring higher-income or higher-priority taxpayers—while a negative CI indicates progressive targeting—favoring lower-income or lower-priority taxpayers.

This standardized measure allows direct comparison of alternative targeting rules, facilitating an interpretation of how closely each policy aligns with either progressive or regressive targeting.

H.1. Definition and Construction of Concentration Curves

Let (y_i, s_i, w_i) denote, respectively, the outcome of interest (e.g., income), the ranking variable (e.g., priority score), and the sampling weight for each individual $i = 1, \ldots, n$, with $w_i > 0$. Let $W = \sum_{i=1}^{n} w_i$ denote total population weight, and let the weighted mean of the outcome be

$$\mu = \frac{1}{W} \sum_{i=1}^{n} w_i \, y_i.$$

Individuals are ordered by ascending s_i (or descending, depending on the context). Define cumulative population and outcome shares up to rank k as

$$P_k = \frac{1}{W} \sum_{i=1}^{k} w_i, \qquad L_k = \frac{1}{W\mu} \sum_{i=1}^{k} w_i \, y_i. \tag{39}$$

The concentration curve connects the sequence of points $\{(P_k, L_k)\}_{k=0}^{n}$ with linear segments, where $P_0 = L_0 = 0$ and $P_n = L_n = 1$. The horizontal axis (P) measures the cumulative share of the population ranked by s_i, while the vertical axis (L) measures the cumulative share of the outcome held by those individuals.

If all individuals have identical outcomes, the curve coincides with the 45° line $(L = P)$. A curve lying below the diagonal indicates that higher-ranked individuals hold a disproportionately large share of the outcome (a regressive pattern), whereas a curve lying above it indicates that outcomes are concentrated among lower-ranked individuals (a progressive pattern).

H.2. The Concentration Index

The concentration index (CI) equals twice the area between the concentration curve and the 45° line. Let A denote the area under the concentration curve. Then:

$$\mathrm{CI} = 1 - 2A. \tag{40}$$

In discrete form, using the trapezoidal rule:

$$A = \sum_{k=1}^{n} \frac{(L_k + L_{k-1})}{2}\,(P_k - P_{k-1}), \qquad \text{so that} \qquad \mathrm{CI} = 1 - \sum_{k=1}^{n} (L_k + L_{k-1})(P_k - P_{k-1}). \tag{41}$$

An equivalent and often more convenient formulation expresses the CI as a weighted covariance between the outcome and individuals' fractional ranks in the s_i distribution. Define the weighted fractional rank:

$$R_i = \frac{1}{W}\left(\sum_{j=1}^{i-1} w_j + \tfrac{1}{2} w_i\right), \tag{42}$$

which lies in $(0,1]$ for observations sorted by s_i. Then:

$$\mathrm{CI} = \frac{2}{\mu}\mathrm{Cov}_w(y_i, R_i) = \frac{2}{W\mu}\sum_{i=1}^{n} w_i\,(y_i - \mu)\,(R_i - \tfrac{1}{2}). \tag{43}$$

Equation (43) highlights that CI measures how strongly outcomes covary with individuals' ranks in the prioritization distribution: a positive value indicates that outcomes rise with rank (concentration among higher-ranked individuals), while a negative value indicates the opposite.

H.3. The Policy Concentration Index (PCI)

The standard concentration index measures inequality in outcomes relative to a ranking variable but does not reflect the underlying policy objective that motivates that ranking. In practice, targeting rules often embody competing normative goals—for instance, maximizing revenue (efficiency) versus promoting redistribution (equity). To compare alternative prioritization rules within a unified framework, we define the Policy Concentration Index (PCI), a normalized version of the CI that situates each rule between two conceptual extremes: the most progressive and most regressive allocations.

Let $\mathrm{CI}(s_i)$ denote the concentration index corresponding to prioritization rule s_i. We define two benchmark allocations:

- **Most Progressive Policy (MPP):** policy effort or resources are concentrated entirely among the lowest-ranked individuals. Numerically, this case corresponds to ranking individuals inversely by income, yielding $\mathrm{CI}_{MPP} = -\mathrm{GINI}_{w_i}$.

- **Most Regressive Policy (MRP):** policy effort is concentrated entirely among the highest-ranked individuals, yielding $\mathrm{CI}_{MRP} = \mathrm{GINI}_{w_i}$.

The Policy Concentration Index rescales each observed CI relative to these bounds:

$$\mathrm{PCI} = \frac{2\,\mathrm{CI}(s_i)}{\mathrm{CI}_{MPP} - \mathrm{CI}_{MRP}} = \frac{\mathrm{CI}(s_i)}{\mathrm{GINI}_{w_i}}. \tag{44}$$

Under this normalization, PCI = 1 corresponds to maximally regressive targeting (full prioritization of the richest), PCI = −1 to maximally progressive targeting (full prioritization of the poorest), and PCI = 0 to a neutral allocation where outcomes are evenly distributed across ranks.

Figure **??** illustrates these concepts using Lorenz-style concentration curves. The dashed 45° line represents perfect equality. The solid blue curve labeled Most Progressive (Extreme) lies above this line, showing concentration among the poorest, while the solid red curve labeled Most Regressive (Extreme) lies below it, showing concentration among the richest. The shaded blue and red areas visualize the degree of progressive and regressive concentration, respectively. Two additional dotted curves (Policy A and Policy B) represent empirically plausible allocations that fall within these bounds, providing visual diagnostics of each policy's position along the equity–efficiency continuum.

H.4. Estimation and Inference

We estimate the concentration index using the weighted covariance representation in Equation (43). Sampling uncertainty is assessed through nonparametric bootstrap resampling. For each bootstrap replicate $b = 1, \ldots, B$, a sample of individuals (or clusters) is drawn with replacement, and the concentration index is recomputed:

$$\widehat{\mathrm{CI}}^{(b)} = \frac{2}{\hat{\mu}^{(b)}} \mathrm{Cov}_{w^{(b)}}(y_i^{(b)}, R_i^{(b)}), \qquad b = 1, \ldots, B. \tag{45}$$

Let $\widehat{\mathrm{CI}}$ denote the estimate from the original sample. Approximate 95% confidence intervals are obtained from the empirical 2.5th and 97.5th percentiles of $\{\widehat{\mathrm{CI}}^{(b)}\}_{b=1}^{B}$.

Interpretation and Visualization

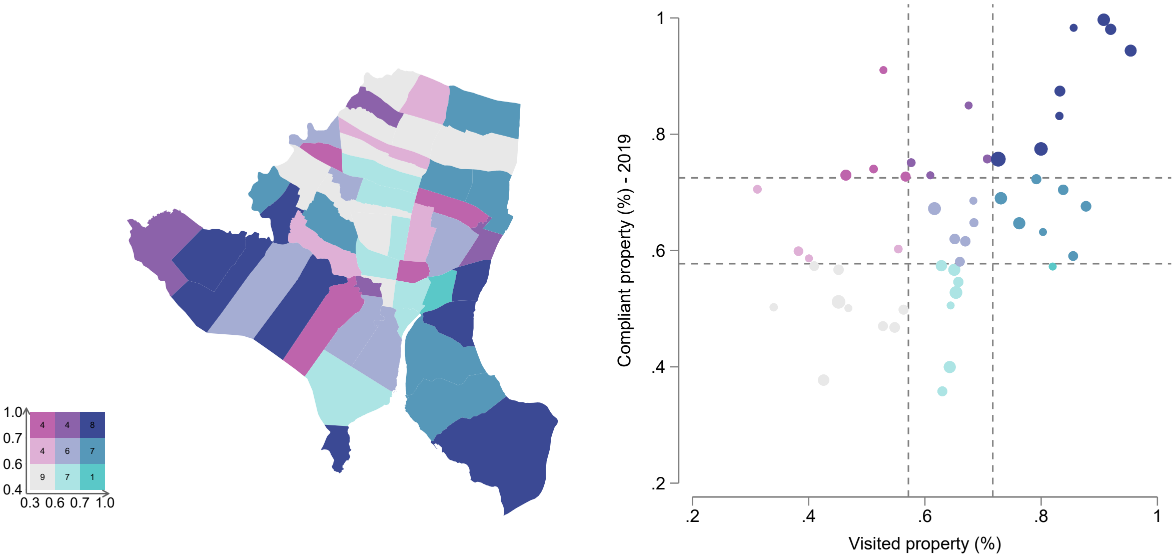
The PCI provides a policy-oriented, unit-free metric of targeting inequality that enables comparison across rules, outcomes, and policy goals. A higher PCI indicates regressive targeting—where enforcement or benefits are concentrated among high-income or high-priority individuals—whereas a lower PCI signals progressive targeting toward lower-income or low-priority groups.

Graphically, the PCI can be represented through normalized concentration curves bounded by the two conceptual extremes. The Robin Hood curve (maximally progressive) lies above the 45° line, while the Negative Robin Hood curve (maximally regressive) lies below it. Observed prioritization rules trace curves within these bounds, and their relative position indicates the degree of progressivity or regressivity.

In the context of tax enforcement or benefit allocation, these curves provide a transparent visualization of how administrative effort or policy attention is distributed across the population. Comparing CIs or PCIs across multiple targeting rules therefore offers a concise, interpretable measure of the equity–efficiency trade-offs that underlie alternative policy designs.

I Replication of Section 3 Figures Using the Balanced Sample

Figure 14: Pre-Treatment Compliance and Treatment Intensity (Balanced Sample)



Note: This figure replicates Figure 5 using only properties observed in all pre- and post-treatment periods (balanced sample). The left panel maps the share of properties that paid before the 2019 deadline against the share visited during the intervention across 50 neighborhoods. The right panel presents the corresponding scatterplot, with point size proportional to the number of properties. Vertical and horizontal dashed lines denote the 33rd and 67th percentiles along each axis.

J Long Tables

Table 8: Variables, Definitions, and Balance by Treatment Status

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
IDs & Location						
1	Property ID	`ID_PROPT` ID	Unique identifier for the property (parcel-level tax object).			
2	Owner ID	`ID_OWNER` ID	Unique identifier for the registered owner.			
3	Subdistrict (Kelurahan)	`l_subdistrict` Categorical (string)	Administrative subdistrict (*kelurahan*) of the property.			
4	Village/Neighborhood	`l_name_village` Categorical (string)	Village or neighborhood name for finer location detail.			
5	Treatment	`Z_treatment` Binary	Assumes value 1 if the property has been visited			
Property Characteristics (prefix `p_`)						
6	Log tax base (2021)	`p_ln_tax_base` Continuous (log)	Log of the property's tax base (FY21).	18.325 (0.006)	18.733 (0.009)	0.408***
7	Δ log tax base (21–15)	`p_delta_ln_tax_base` Continuous (log change)	Change in log tax base, 2021–2015.	0.745 (0.003)	0.760 (0.004)	0.015***
8	Log tax assessment (2120)	`p_ln_tax_assessment` Continuous (log)	Log assessed liability (2021).	11.176 (0.009)	11.659 (0.014)	0.483***
9	Δ log tax assessment (21–15)	`p_delta_ln_tax_assessment` Continuous (log change)	Change in log assessed liability, 2021–2015.	1.322 (0.012)	1.154 (0.012)	-0.167***
10	Tax rate always high (15–21)	`p_tax_rate_always_high` Binary	=1 if rate high in all years 2015–2021.	0.003 (0.000)	0.011 (0.001)	0.007***
11	Tax rate new high by 2021	`p_tax_rate_new_high` Binary	=1 if moved to high bracket by 2021.	0.018 (0.001)	0.050 (0.002)	0.032***
12	Log land area (m^2, 2021)	`p_ln_land_area` Continuous (log)	Log land area (2021).	5.636 (0.006)	5.901 (0.009)	0.265***

Continued on next page

Table 8 (continued)

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
13	Δ log land area (21–15)	`p_delta_ln_land_area` Continuous (log change)	Change in log land area, 2021–2015.	-0.020 (0.002)	-0.024 (0.002)	-0.004
14	Log land NJOP value/m^2 (2021)	`p_ln_njop_land_value` Continuous (log)	Log land NJOP per m^2 (2021).	12.163 (0.005)	12.410 (0.007)	0.247***
15	Δ log land NJOP/m^2 (21–15)	`p_delta_ln_njop_land_value` Continuous (log change)	Change in log land NJOP/m^2, 2021–2015.	0.771 (0.002)	0.783 (0.002)	0.012***
16	New construction by 2021	`p_new_contruction` Binary	=1 if construction newly present by 2021.	0.082 (0.002)	0.068 (0.002)	-0.013***
17	Construction existed in 2015	`p_old_contruction` Binary	=1 if construction present in 2015.	0.736 (0.003)	0.610 (0.004)	-0.126***
18	Log building area (m^2, 2021)	`p_ln_bldg_area` Continuous (log)	Log building area (2021).	3.563 (0.011)	3.132 (0.018)	-0.431***
19	Δ log building area (21–15)	`p_delta_ln_bldg_area` Continuous (log change)	Change in log building area, 2021–2015.	0.546 (0.008)	0.505 (0.010)	-0.041***
20	Log building NJOP/m^2 (2021)	`p_ln_njop_bldg_value` Continuous (log)	Log building NJOP per m^2 (2021).	10.510 (0.031)	8.810 (0.049)	-1.700***
21	Δ log building NJOP/m^2 (21–15)	`p_delta_ln_njop_bldg_value` Continuous (log change)	Change in log building NJOP/m^2, 2021–2015.	1.183 (0.022)	0.964 (0.027)	-0.218***
Owner Characteristics (prefix `ow_`)						
22	Multiple properties (2021)	`ow_multiple` Binary	=1 if owner holds >1 property (2021).	0.043 (0.001)	0.076 (0.002)	0.034***
23	Lives in Gorontalo (2021)	`ow_lives_gorontalo` Binary	=1 if owner's residence in Gorontalo.	0.991 (0.001)	0.972 (0.001)	-0.019***
24	Owner address matches property (2021)	`ow_same_adress` Binary	=1 if owner and property addresses match.	0.384 (0.003)	0.330 (0.004)	-0.053***

Continued on next page

Table 8 (continued)

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
Historical Payments (prefix h_) — Year 2015						
25	Paid before deadline (2015)	h_Y_bd2015 Binary	Indicator for timely payment in 2015.	0.722 (0.003)	0.638 (0.004)	-0.084***
26	Any payment (2015)	h_Y_at2015 Binary	Indicator for any payment in 2015.	0.915 (0.002)	0.925 (0.002)	0.010***
27	Paid after deadline (2015)	h_Y_ad2015 Binary	Indicator for late payment in 2015.	0.193 (0.002)	0.287 (0.004)	0.095***
28	ln(days early, 2015)	h_ln_day_bd2015 Continuous	Log days early (if early).	3.526 (0.015)	3.028 (0.020)	-0.498***
29	ln(days late, 2015)	h_ln_day_ad2015 Continuous	Log days late (if late).	1.187 (0.016)	1.790 (0.023)	0.603***
Historical Payments — 2016						
30	h_Y_bd2016	h_Y_bd2016 Binary	Paid before deadline (2016).	0.756 (0.003)	0.677 (0.004)	-0.080***
31	h_Y_at2016	h_Y_at2016 Binary	Any payment (2016).	0.927 (0.002)	0.929 (0.002)	0.002
32	h_Y_ad2016	h_Y_ad2016 Binary	Paid after deadline (2016).	0.171 (0.002)	0.252 (0.003)	0.082***
33	h_ln_day_bd2016	h_ln_day_bd2016 Continuous	ln(days early, 2016).	3.751 (0.014)	3.256 (0.019)	-0.495***
34	h_ln_day_ad2016	h_ln_day_ad2016 Continuous	ln(days late, 2016).	1.031 (0.015)	1.561 (0.022)	0.530***
Historical Payments — 2017						
35	h_Y_bd2017	h_Y_bd2017 Binary	Paid before deadline (2017).	0.743 (0.003)	0.624 (0.004)	-0.119***
36	h_Y_at2017	h_Y_at2017 Binary	Any payment (2017).	0.968 (0.001)	0.960 (0.002)	-0.007***
37	h_Y_ad2017	h_Y_ad2017 Binary	Paid after deadline (2017).	0.225 (0.003)	0.337 (0.004)	0.112***

Continued on next page

Table 8 (continued)

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
38	h_ln_day_bd2017	h_ln_day_bd2017 Continuous	ln(days early, 2017).	3.124 (0.013)	2.537 (0.018)	-0.587***
39	h_ln_day_ad2017	h_ln_day_ad2017 Continuous	ln(days late, 2017).	1.182 (0.015)	1.899 (0.023)	0.717***
Historical Payments — 2018						
40	h_Y_bd2018	h_Y_bd2018 Binary	Paid before deadline (2018).	0.736 (0.003)	0.603 (0.004)	-0.133***
41	h_Y_at2018	h_Y_at2018 Binary	Any payment (2018).	0.960 (0.001)	0.934 (0.002)	-0.026***
42	h_Y_ad2018	h_Y_ad2018 Binary	Paid after deadline (2018).	0.225 (0.003)	0.331 (0.004)	0.106***
43	h_ln_day_bd2018	h_ln_day_bd2018 Continuous	ln(days early, 2018).	3.301 (0.014)	2.580 (0.018)	-0.721***
44	h_ln_day_ad2018	h_ln_day_ad2018 Continuous	ln(days late, 2018).	1.231 (0.015)	1.871 (0.023)	0.640***
Historical Payments — 2019						
45	h_Y_bd2019	h_Y_bd2019 Binary	Paid before deadline (2019).	0.697 (0.003)	0.563 (0.004)	-0.134***
46	h_Y_at2019	h_Y_at2019 Binary	Any payment (2019).	0.949 (0.001)	0.902 (0.002)	-0.046***
47	h_Y_ad2019	h_Y_ad2019 Binary	Paid after deadline (2019).	0.252 (0.003)	0.339 (0.004)	0.087***
48	h_ln_day_bd2019	h_ln_day_bd2019 Continuous	ln(days early, 2019).	2.720 (0.013)	2.139 (0.017)	-0.581***
49	h_ln_day_ad2019	h_ln_day_ad2019 Continuous	ln(days late, 2019).	1.327 (0.015)	1.910 (0.023)	0.583***
Historical Payments — Aggregates						
50	Years before deadline (15–19)	h_Y_bd_15_19 Count	# years paid early, 2015–2019.	3.654 (0.009)	3.104 (0.013)	-0.550***

Continued on next page

Table 8 (continued)

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
51	Years after deadline (15–19)	`h_Y_ad_15_19` Count	# years paid late, 2015–2019.	1.065 (0.008)	1.547 (0.012)	0.482***
52	Years with any payment (15–19)	`h_Y_at_15_19` Count	# years any payment, 2015–2019.	4.719 (0.005)	4.651 (0.007)	-0.068***
53	Mean ln(days early, 15–19)	`h_mean_ln_day_bd_15_19` Continuous	Avg. log days early.	3.284 (0.009)	2.708 (0.012)	-0.576***
54	Mean ln(days late, 15–19)	`h_mean_ln_day_ad_15_19` Continuous	Avg. log days late.	1.192 (0.010)	1.806 (0.016)	0.615***
55	Years before deadline (15–18)	`h_Y_bd_15_18` Count	# years paid early, 2015–2018.	2.957 (0.007)	2.541 (0.011)	-0.416***
56	Years after deadline (15–18)	`h_Y_ad_15_18` Count	# years paid late, 2015–2018.	0.813 (0.007)	1.208 (0.010)	0.395***
57	Years with any payment (15–18)	`h_Y_at_15_18` Count	# years any payment, 2015–2018.	3.770 (0.004)	3.749 (0.006)	-0.021***
58	Mean ln(days early, 15–18)	`h_mean_ln_day_bd_15_18` Continuous	Avg. log days early (15–18).	3.425 (0.009)	2.850 (0.013)	-0.575***
59	Mean ln(days late, 15–18)	`h_mean_ln_day_ad_15_18` Continuous	Avg. log days late (15–18).	3.425 (0.009)	2.850 (0.013)	-0.575***
Outcomes (prefix `Y_`) — 2019–2024						
60	Paid before deadline (2019)	`Y_Y_bd2019` Binary	Indicator for timely payment.	0.697 (0.003)	0.563 (0.004)	-0.134***
61	Any payment (2019)	`Y_Y_at2019` Binary	Indicator for any payment.	0.949 (0.001)	0.902 (0.002)	-0.046***
62	Paid after deadline (2019)	`Y_Y_ad2019` Binary	Indicator for late payment.	0.252 (0.003)	0.339 (0.004)	0.087***

Continued on next page

Table 8 (continued)

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
63	ln(days early, 2019)	Y_ln_day_bd2019 Continuous	Log days early.	2.720 (0.013)	2.139 (0.017)	-0.581***
64	ln(days late, 2019)	Y_ln_day_ad2019 Continuous	Log days late.	1.327 (0.015)	1.910 (0.023)	0.583***
65	Paid before deadline (2020)	Y_Y_bd2020 Binary	Timely payment (2020).	0.697 (0.003)	0.516 (0.004)	-0.181***
66	Any payment (2020)	Y_Y_at2020 Binary	Any payment (2020).	0.941 (0.001)	0.864 (0.003)	-0.076***
67	Paid after deadline (2020)	Y_Y_ad2020 Binary	Late payment (2020).	0.244 (0.003)	0.348 (0.004)	0.105***
68	ln(days early, 2020)	Y_ln_day_bd2020 Continuous	Log days early (2020).	2.522 (0.012)	1.778 (0.016)	-0.744***
69	ln(days late, 2020)	Y_ln_day_ad2020 Continuous	Log days late (2020).	1.230 (0.014)	1.917 (0.022)	0.687***
70	Paid before deadline (2021)	Y_Y_bd2021 Binary	Timely payment (2021).	0.693 (0.003)	0.533 (0.004)	-0.160***
71	Any payment (2021)	Y_Y_at2021 Binary	Any payment (2021).	0.895 (0.002)	0.823 (0.003)	-0.073***
72	Paid after deadline (2021)	Y_Y_ad2021 Binary	Late payment (2021).	0.202 (0.002)	0.290 (0.004)	0.088***
73	ln(days early, 2021)	Y_ln_day_bd2021 Continuous	Log days early (2021).	2.799 (0.013)	1.988 (0.017)	-0.811***
74	ln(days late, 2021)	Y_ln_day_ad2021 Continuous	Log days late (2021).	1.045 (0.013)	1.565 (0.021)	0.520***
75	Paid before deadline (2022)	Y_Y_bd2022 Binary	Timely payment (2022).	0.726 (0.003)	0.557 (0.004)	-0.168***
76	Any payment (2022)	Y_Y_at2022 Binary	Any payment (2022).	0.881 (0.002)	0.780 (0.003)	-0.100***

Continued on next page

Table 8 (continued)

#	Variable (Paper)	Variable (Dataset)/Type	Description	Mean (Visited)	Mean (Unvisited)	Test
77	Paid after deadline (2022)	`Y_Y_ad2022` Binary	Late payment (2022).	0.155 (0.002)	0.223 (0.003)	0.068***
78	ln(days early, 2022)	`Y_ln_day_bd2022` Continuous	Log days early (2022).	2.900 (0.012)	2.120 (0.017)	-0.780***
79	ln(days late, 2022)	`Y_ln_day_ad2022` Continuous	Log days late (2022).	0.748 (0.011)	1.123 (0.018)	0.375***
80	Paid before deadline (2023)	`Y_Y_bd2023` Binary	Timely payment (2023).	0.742 (0.003)	0.605 (0.004)	-0.137***
81	Any payment (2023)	`Y_Y_at2023` Binary	Any payment (2023).	0.865 (0.002)	0.758 (0.003)	-0.106***
82	Paid after deadline (2023)	`Y_Y_ad2023` Binary	Late payment (2023).	0.122 (0.002)	0.153 (0.003)	0.030***
83	ln(days early, 2023)	`Y_ln_day_bd2023` Continuous	Log days early (2023).	3.357 (0.014)	2.588 (0.019)	-0.769***
84	ln(days late, 2023)	`Y_ln_day_ad2023` Continuous	Log days late (2023).	0.575 (0.010)	0.748 (0.015)	0.173***
85	Paid before deadline (2024)	`Y_Y_bd2024` Binary	Timely payment (2024).	0.729 (0.003)	0.602 (0.004)	-0.126***
86	Any payment (2024)	`Y_Y_at2024` Binary	Any payment (2024).	0.773 (0.003)	0.651 (0.004)	-0.122***
87	Paid after deadline (2024)	`Y_Y_ad2024` Binary	Late payment (2024).	0.045 (0.001)	0.049 (0.002)	0.004*
88	ln(days early, 2024)	`Y_ln_day_bd2024` Continuous	Log days early (2024).	3.019 (0.013)	2.324 (0.017)	-0.695***
89	ln(days late, 2024)	`Y_ln_day_ad2024` Continuous	Log days late (2024).	0.146 (0.004)	0.156 (0.006)	0.010

Notes: “Visited”=Treated (2); “Unvisited”=Untreated (1). Entries in Mean (Visited/Unvisited) show mean with SE in parentheses beneath.

K Robustness Checks and Alternative Specifications

This appendix replicates all empirical results from the applied sections of the paper under two alternative robustness specifications. These exercises are designed to assess the stability of the main findings reported in Section 5 and to verify that the estimated treatment effects are not driven by modeling choices, sample composition, or overfitting introduced by the debiasing procedure.

First, we re-estimate all models using causal forests *without* the orthogonalization (debiased) step described in Section **??**. This specification provides a direct benchmark for evaluating whether the double-robust correction materially affects the magnitude or significance of the results.

Second, although all covariates used in the main specification are pre-treatment, one might still be concerned that data collected between 2018 (when the debiasing procedure was implemented) and 2021 (when the intervention began) could indirectly influence the estimates. To address this possibility, we further restrict the covariate set to characteristics observed strictly before the intervention period, namely between 2015 and 2018.

Across both robustness exercises, the estimated treatment effects remain highly consistent in magnitude, direction, and statistical significance. This stability reinforces the conclusion that the observed post-intervention improvements in tax compliance reflect genuine behavioral responses to the door-to-door program rather than artifacts of model specification, sample design, or estimation procedure. Equivalent tables and figures corresponding to these robustness checks are presented below.

Table 9: Treatment Effect Estimates by Year and Sample (AIPW, 2021–2024)

	Full Data (95% CI)				Train (AIPW Estimates)				Test (AIPW Estimates)			
	2021	**2022**	**2023**	**2024**	**2021**	**2022**	**2023**	**2024**	**2021**	**2022**	**2023**	**2024**
Panel A: Average Treatment Effects (AIPW Estimates)												
ATE	[0.071, 0.085]	[0.064, 0.080]	[0.054, 0.074]	[0.041, 0.061]	0.076^{***}	0.075^{***}	0.059^{***}	0.052^{***}	0.077^{***}	0.072^{***}	0.060^{***}	0.051^{***}
	–	–	–	–	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
ATT	[0.070, 0.090]	[0.063, 0.083]	[0.058, 0.078]	[0.042, 0.062]	0.078^{***}	0.077^{***}	0.063^{***}	0.052^{***}	0.081^{***}	0.083^{***}	0.061^{***}	0.057^{***}
	–	–	–	–	(0.006)	(0.006)	(0.006)	(0.006)	(0.009)	(0.009)	(0.009)	(0.009)
ATU	[0.064, 0.084]	[0.059, 0.079]	[0.046, 0.066]	[0.038, 0.058]	0.072^{***}	0.071^{***}	0.052^{***}	0.050^{***}	0.069^{***}	0.062^{***}	0.041^{***}	0.061^{***}
	–	–	–	–	(0.006)	(0.006)	(0.006)	(0.006)	(0.024)	(0.025)	(0.024)	(0.025)
Outcome Mean:												
Average	0.628	0.657	0.685	0.676	0.627	0.655	0.684	0.672	0.632	0.665	0.690	0.686
Treated	0.684	0.716	0.733	0.719	0.682	0.713	0.729	0.715	0.692	0.728	0.744	0.732
Untreated	0.533	0.557	0.605	0.602	0.534	0.556	0.608	0.598	0.532	0.559	0.601	0.609
Panel B: Heterogeneous Effects by Predicted $\hat{\tau}$ (High vs. Low)												
High (τ_{High})	[0.117, 0.145]	[0.104, 0.132]	[0.094, 0.122]	[0.067, 0.095]	0.136^{***}	0.118^{***}	0.100^{***}	0.081^{***}	0.102^{***}	0.109^{***}	0.089^{***}	0.080^{***}
	–	–	–	–	(0.008)	(0.009)	(0.008)	(0.008)	(0.013)	(0.013)	(0.013)	(0.013)
Low (τ_{Low})	[0.013, 0.037]	[0.013, 0.037]	[0.007, 0.031]	[0.008, 0.032]	0.016^{***}	0.033^{***}	0.018^{***}	0.023^{***}	0.043^{***}	0.041^{***}	0.033^{***}	0.004^{***}
	–	–	–	–	(0.008)	(0.008)	(0.008)	(0.008)	(0.016)	(0.016)	(0.016)	(0.015)
Difference (High–Low)	[0.088, 0.123]	[0.075, 0.111]	[0.071, 0.106]	[0.043, 0.079]	0.119^{***}	0.085^{***}	0.082^{***}	0.058^{***}	0.059^{***}	0.069^{***}	0.056^{***}	0.075^{***}
	–	–	–	–	(0.012)	(0.012)	(0.012)	(0.012)	(0.020)	(0.021)	(0.020)	(0.020)
Outcome Mean by $\hat{\tau}$ Group:												
High	0.618	0.629	0.662	0.660	0.638	0.644	0.674	0.641	0.636	0.663	0.692	0.685
Low	0.638	0.686	0.709	0.692	0.616	0.666	0.695	0.702	0.631	0.663	0.691	0.678

Notes: The "Full Data" columns show 95% confidence intervals in brackets, without point estimates. Train and Test columns display point estimates with clustered standard errors in parentheses and significance levels indicated as: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. "High" and "Low" denote observations above/below the median predicted $\hat{\tau}$ within each sample-year. Year 2020 omitted due to pandemic-related disruptions.

Table 10: Treatment Effect Estimates Before and After Intervention - Non-debiased Causal Forest

	Before Intervention		**After Intervention**			
	2018	2019	2021	2022	2023	2024
Panel A: Average Treatment Effects (AIPW Estimates)						
Average (ATE)	0.013**	0.015***	0.089***	0.820***	0.074***	0.061***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Treated (ATT)	0.009**	0.013***	0.090***	0.820***	0.078***	0.061***
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Untreated (ATU)	0.015***	0.015***	0.086***	0.081***	0.068***	0.060***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Outcome Mean:						
Average	0.686	0.647	0.634	0.663	0.691	0.682
Treated	0.736	0.697	0.693	0.726	0.742	0.729
Untreated	0.603	0.563	0.533	0.557	0.605	0.602
Panel B: Heterogeneous Effects by Predicted $\hat{\tau}$ (High vs. Low)						
High (τ_{High})	0.032***	0.027***	0.145***	0.129***	0.118***	0.099***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Low (τ_{Low})	-0.006	0.003	0.033***	0.035***	0.031***	0.023***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Difference (High–Low)	0.038***	0.024**	0.112***	0.094***	0.088***	0.076***
	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Outcome Mean:						
High	0.674	0.634	0.617	0.629	0.664	0.664
Low	0.699	0.661	0.651	0.697	0.718	0.700

Notes: All treatment effects are estimated using out-of-bag forest-based Augmented Inverse Probability Weighting (AIPW; Athey et al. (2019) and Robins et al. (1994)). Panel A reports average treatment effects (ATE), average treatment effects on the treated (ATT), and on the untreated (ATU). Panel B reports estimates for observations above ("High") and below ("Low") the median predicted treatment effect $\hat{\tau}$ in each year. Standard errors clustered at the administrative-unit level are shown in parentheses. "Before Intervention" covers the pre-rollout period (2018–2019), and "After Intervention" refers to the full rollout (2021–2024). Year 2020 is excluded because of COVID-related disruptions (payment holidays and exceptional billing rules). Significance levels: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 11: Treatment Effect Estimates Before and After Intervention - Non-debiased Causal Forest

	Before Intervention		**After Intervention**			
	2018	2019	2021	2022	2023	2024
Panel A: Average Treatment Effects (AIPW Estimates)						
Average (ATE)	0.013^{**}	0.015^{***}	0.089^{***}	0.820^{***}	0.074^{***}	0.061^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Treated (ATT)	0.009^{**}	0.013^{***}	0.090^{***}	0.820^{***}	0.078^{***}	0.061^{***}
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Untreated (ATU)	0.015^{***}	0.015^{***}	0.086^{***}	0.081^{***}	0.068^{***}	0.060^{***}
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Outcome Mean:						
Average	0.686	0.647	0.634	0.663	0.691	0.682
Treated	0.736	0.697	0.693	0.726	0.742	0.729
Untreated	0.603	0.563	0.533	0.557	0.605	0.602
Panel B: Heterogeneous Effects by Predicted $\hat{\tau}$ (High vs. Low)						
High (τ_{High})	0.032^{***}	0.027^{***}	0.145^{***}	0.129^{***}	0.118^{***}	0.099^{***}
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Low (τ_{Low})	-0.006	0.003	0.033^{***}	0.035^{***}	0.031^{***}	0.023^{***}
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Difference (High–Low)	0.038^{***}	0.024^{**}	0.112^{***}	0.094^{***}	0.088^{***}	0.076^{***}
	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Outcome Mean:						
High	0.674	0.634	0.617	0.629	0.664	0.664
Low	0.699	0.661	0.651	0.697	0.718	0.700

Notes: All treatment effects are estimated using out-of-bag forest-based Augmented Inverse Probability Weighting (AIPW; Athey et al. (2019) and Robins et al. (1994)). Panel A reports average treatment effects (ATE), average treatment effects on the treated (ATT), and on the untreated (ATU). Panel B reports estimates for observations above ("High") and below ("Low") the median predicted treatment effect $\hat{\tau}$ in each year. Standard errors clustered at the administrative-unit level are shown in parentheses. "Before Intervention" covers the pre-rollout period (2018–2019), and "After Intervention" refers to the full rollout (2021–2024). Year 2020 is excluded because of COVID-related disruptions (payment holidays and exceptional billing rules). Significance levels: $^{*}\,p < 0.10$, $^{**}\,p < 0.05$, $^{***}\,p < 0.01$.

Table 12: Best Linear Projection Results by Variable Set (2019 vs. 2021) - Base 2018

	2019					**2021**				
Variable	(i)	(ii)	(iii)	(iv)	all	(i)	(ii)	(iii)	(iv)	all
Intercept	0.057	0.054	0.052	0.006	0.069	0.622***	0.412***	0.167***	0.122***	0.522***
	(0.075)	(0.080)	(0.037)	(0.010)	(0.091)	(0.078)	(0.086)	(0.041)	(0.010)	(0.098)
A. Historical Payment Behavior (2015–2017)										
Paid 1 bill on time				0.025	0.026				-0.033*	-0.043**
				(0.017)	(0.017)				(0.018)	(0.018)
Paid 2 bills on time				0.037**	0.037**				-0.008	-0.016
				(0.015)	(0.015)				(0.015)	(0.015)
Paid 4 bills on time				0.011	0.011				-0.040***	-0.038***
				(0.011)	(0.011)				(0.012)	(0.012)
B. Owner Characteristics (2021)										
Lives in Gorontalo			-0.030		-0.030			-0.064		-0.073*
			(0.038)		(0.038)			(0.042)		(0.042)
Multiple Properties			-0.018		-0.017			-0.019		-0.016
			(0.019)		(0.020)			(0.020)		(0.021)
Same Address as Property			-0.007		-0.006			-0.013		-0.008
			(0.009)		(0.009)			(0.010)		(0.010)
C. Property Characteristics										
ln(Building Area)		-0.004			-0.004		-0.026***			-0.026***
		(0.008)			(0.008)		(0.008)			(0.008)
ln(Land Area)		-0.001			0.000		-0.018***			-0.018***
		(0.005)			(0.005)		(0.006)			(0.006)
ln(Building Value)		0.011*			0.011*		0.008			0.008
		(0.007)			(0.007)		(0.007)			(0.007)
ln(Land Value)		-0.003			-0.003		-0.014**			-0.015**
		(0.006)			(0.006)		(0.006)			(0.006)
ln(Tax Base, 2021)	-0.002					-0.028***				
	(0.004)					(0.004)				
Dummy: New Construction		-0.113			-0.119		-0.020			-0.020
		(0.088)			(0.088)		(0.089)			(0.089)
Dummy: Old Construction		-0.120*			-0.123*		-0.043			-0.040
		(0.086)			(0.086)		(0.087)			(0.087)

Notes: Coefficients on top, robust standard errors in parentheses below.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Omitted category for payment history: "Paid 3 bills on time."

Table 13: Best Linear Projection Results by Variable Set (2019 vs. 2021) - Base CF

	2019					2021				
Variable	(i)	(ii)	(iii)	(iv)	all	(i)	(ii)	(iii)	(iv)	all
Intercept	0.088	0.091	0.050	0.010	0.127	0.617***	0.436***	0.145***	0.093***	0.506***
	(0.071)	(0.076)	(0.034)	(0.011)	(0.085)	(0.074)	(0.081)	(0.038)	(0.012)	(0.093)
A. Historical Payment Behavior (2015–2018 / 2015–2019)										
Reference category: Paid 3 bills on time										
Paid 1 bill on time				0.015	0.016				-0.015	-0.020
				(0.018)	(0.018)				(0.022)	(0.022)
Paid 2 bills on time				0.025	0.026				-0.005	-0.006
				(0.017)	(0.017)				(0.019)	(0.019)
Paid 4 bills on time				0.008	0.008				0.016	0.016
				(0.014)	(0.014)				(0.017)	(0.017)
Paid 5 bills on time				-0.000	-0.000				0.035**	0.038**
				(0.012)	(0.013)				(0.016)	(0.016)
Paid 6 bills on time									-0.041***	-0.035**
									(0.014)	(0.014)
B. Owner Characteristics (2021)										
Lives in Gorontalo			-0.033		-0.033			-0.054		-0.063
			(0.034)		(0.034)			(0.039)		(0.039)
Multiple Properties			-0.016		-0.012			-0.009		-0.005
			(0.018)		(0.018)			(0.019)		(0.020)
Same Address as Property			-0.002		-0.002			-0.009		-0.004
			(0.009)		(0.009)			(0.009)		(0.009)
C. Property Characteristics										
ln(Building Area)		-0.001			-0.000		-0.023***			-0.022***
		(0.008)			(0.008)		(0.008)			(0.008)
ln(Land Area)		-0.005			-0.004		-0.017***			-0.016***
		(0.005)			(0.005)		(0.005)			(0.005)
ln(Building Value)		0.011*			0.011*		0.004			0.004
		(0.006)			(0.006)		(0.007)			(0.007)
ln(Land Value)		-0.004			-0.005		-0.017***			-0.018***
		(0.006)			(0.006)		(0.006)			(0.006)
ln(Tax Base, 2021)	-0.004					-0.029***				
	(0.004)					(0.004)				
Dummy: New Construction		-0.122			-0.125		0.025			0.019
		(0.082)			(0.082)		(0.085)			(0.085)
Dummy: Old Construction		-0.134*			-0.136*		0.001			0.001
		(0.080)			(0.080)		(0.083)			(0.083)

Notes: Coefficients on top, robust standard errors in parentheses below.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Reference category for payment history: "Paid 3 bills on time."

Table 14: Best Linear Projection Results by Variable Set (2022–2024)

	2022					2023						
Variable	(i)	(ii)	(iii)	(iv)	all	(i)	(ii)	(iii)	(iv)	all	(i)	(ii)
Intercept	0.556***	0.457***	0.110***	0.080***	0.512***	0.440***	0.331***	0.161***	0.083***	0.456***	0.463***	0.404***
	(0.075)	(0.081)	(0.039)	(0.012)	(0.093)	(0.075)	(0.083)	(0.039)	(0.013)	(0.094)	(0.075)	(0.081)
A. Historical Payment Behavior (2015–2019)												
Paid 1 bill on time				-0.005	-0.003				-0.060**	-0.059**		
				(0.023)	(0.023)				(0.024)	(0.024)		
Paid 2 bills on time				-0.002	-0.002				-0.035*	-0.035*		
				(0.019)	(0.019)				(0.020)	(0.020)		
Paid 4 bills on time				0.003	-0.000				-0.001	-0.003		
				(0.017)	(0.017)				(0.017)	(0.017)		
Paid 5 bills on time				0.022*	0.023*				-0.005	-0.006		
				(0.016)	(0.016)				(0.016)	(0.016)		
Paid 6 bills on time				-0.040***	-0.037***				-0.037***	-0.036***		
				(0.014)	(0.015)				(0.015)	(0.015)		
B. Owner Characteristics (2021)												
Lives in Gorontalo			-0.041		-0.048			-0.096**		-0.104**		
			(0.039)		(0.039)			(0.039)		(0.039)		
Multiple Properties			-0.053***		-0.043**			-0.038**		-0.028*		
			(0.018)		(0.019)			(0.019)		(0.019)		
Same Address as Property			0.014*		0.016**			-0.004		-0.004		
			(0.009)		(0.009)			(0.009)		(0.010)		
C. Property Characteristics												
ln(Building Area)		0.002			0.003		-0.011*			-0.011*		-0.002
		(0.008)			(0.008)		(0.008)			(0.008)		(0.008)
ln(Land Area)		-0.030***			-0.029***		-0.013**			-0.012**		-0.020***
		(0.005)			(0.005)		(0.005)			(0.005)		(0.005)
ln(Building Value)		0.001			0.002		0.003			0.003		0.005
		(0.008)			(0.008)		(0.008)			(0.008)		(0.007)
ln(Land Value)		-0.015**			-0.016**		-0.016**			-0.016**		-0.018***
		(0.006)			(0.006)		(0.006)			(0.006)		(0.006)
ln(Tax Base, 2021)	-0.026***					-0.020***					-0.022***	
	(0.004)					(0.004)					(0.004)	
Dummy: New Construction		-0.072			-0.090		0.019			0.017		-0.090
		(0.094)			(0.094)		(0.094)			(0.094)		(0.092)
Dummy: Old Construction		-0.051			-0.065		0.001			0.002		-0.079
		(0.092)			(0.092)		(0.091)			(0.092)		(0.090)

Notes: Coefficients on top, robust standard errors in parentheses below.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Omitted category for payment history: “Paid 3 bills on time.”

Table 15: Treatment Effect Estimates (After Intervention, 2021–2024)

	2021			2022			2023			2024		
	$\hat{\tau}$	$\hat{\tau}_{\text{Train}}$	$\hat{\tau}_{\text{Test}}$	$\hat{\tau}$	$\hat{\tau}_{\text{Train}}$	$\hat{\tau}_{\text{Test}}$	$\hat{\tau}$	$\hat{\tau}_{\text{Train}}$	$\hat{\tau}_{\text{Test}}$	$\hat{\tau}$	$\hat{\tau}_{\text{Train}}$	$\hat{\tau}_{\text{Test}}$
Panel A: Average Treatment Effects (AIPW Estimates)												
Average (ATE)	0.089^{***}	0.088^{***}	0.087^{***}	0.820^{***}	0.087^{***}	0.088^{***}	0.074^{***}	0.071^{***}	0.071^{***}	0.061^{***}	0.064^{***}	0.065^{***}
	(0.004)	(0.006)	(0.006)	(0.004)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)
Treated (ATT)	0.090^{***}	0.088^{***}	0.089^{***}	0.820^{***}	0.088^{***}	0.089^{***}	0.078^{***}	0.074^{***}	0.073^{***}	0.061^{***}	0.065^{***}	0.067^{***}
	(0.005)	(0.006)	(0.006)	(0.005)	(0.007)	(0.006)	(0.005)	(0.007)	(0.006)	(0.005)	(0.007)	(0.006)
Untreated (ATU)	0.086^{***}	0.087^{***}	0.087^{***}	0.081^{***}	0.084^{***}	0.087^{***}	0.068^{***}	0.065^{***}	0.066^{***}	0.060^{***}	0.062^{***}	0.062^{***}
	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)
Outcome Mean:												
Average	0.634	0.634	0.632	0.663	0.663	0.665	0.691	0.691	0.690	0.682	0.682	0.686
Treated	0.693	0.693	0.692	0.726	0.726	0.728	0.742	0.742	0.744	0.729	0.729	0.732
Untreated	0.533	0.533	0.532	0.557	0.557	0.559	0.605	0.605	0.601	0.602	0.602	0.609
Panel B: Heterogeneous Effects by Predicted $\hat{\tau}$ (High vs. Low)												
High (τ_{High})	0.145^{***}	0.133^{***}	0.151^{***}	0.129^{***}	0.123^{***}	0.151^{***}	0.118^{***}	0.107^{***}	0.131^{***}	0.099^{***}	0.093^{***}	0.120^{***}
	(0.007)	(0.008)	(0.007)	(0.007)	(0.009)	(0.008)	(0.007)	(0.009)	(0.007)	(0.007)	(0.009)	(0.007)
Low (τ_{Low})	0.033^{***}	0.043^{***}	0.025^{***}	0.035^{***}	0.051^{***}	0.025^{***}	0.031^{***}	0.035^{***}	0.010^{***}	0.023^{***}	0.036^{***}	0.010^{***}
	(0.006)	(0.008)	(0.007)	(0.006)	(0.008)	(0.007)	(0.006)	(0.008)	(0.006)	(0.006)	(0.008)	(0.007)
Difference (High–Low)	0.112^{***}	0.089^{***}	0.127^{***}	0.094^{***}	0.072^{***}	0.126^{***}	0.088^{***}	0.072^{***}	0.121^{***}	0.076^{***}	0.058^{***}	0.111^{***}
	(0.009)	(0.012)	(0.010)	(0.009)	(0.012)	(0.010)	(0.009)	(0.012)	(0.010)	(0.009)	(0.012)	(0.010)
95% CI [low, high]	[–, –]	[0.067, 0.112]	[0.107, 0.146]	[–, –]	[0.049, 0.095]	[0.106, 0.146]	[–, –]	[0.049, 0.095]	[0.102, 0.140]	[–, –]	[0.034, 0.081]	[0.091, 0.130]
Outcome Mean by $\hat{\tau}$ Group:												
High	0.617	0.632	0.634	0.629	0.655	0.662	0.664	0.683	0.692	0.664	0.681	0.683
Low	0.651	0.635	0.633	0.697	0.671	0.664	0.718	0.699	0.691	0.700	0.683	0.680

Notes: Train/Test columns report estimates from the corresponding CATE Forest samples you provided; the $\hat{\tau}$ column preserves your overall figures. Standard errors in parentheses. The 95% CI row under "Difference (High–Low)" uses your `ci_low` and `ci_high` values (Train/Test only). Year 2020 is excluded due to pandemic-related disruptions. Significance: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.