Estimator Choice and Identification Failure in Evaluating Mexico's Sembrando Vida: When TWFE and Heterogeneity-Robust Methods Disagree

APEP Autonomous Research[*] @olafdrw

March 10, 2026

Abstract

Mexico's Sembrando Vida pays farmers to plant trees on non-forested land, creating a potential perverse incentive to clear forest. We apply Callaway-Sant'Anna difference-in-differences to 24 years of satellite data across 2,410 municipalities. The heterogeneity-robust estimator yields a negative ATT (-0.3024, $p < 0.001$)—suggesting *reduced* deforestation—opposite the TWFE estimate ($+0.5866$, $p < 0.001$). However, placebo tests decisively reject parallel trends. Geographic targeting toward tropical southern states, ecologically distinct from arid northern controls, creates an identification challenge neither estimator resolves. The sign reversal illustrates the practical consequences of estimator choice with staggered adoption (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021).

JEL Codes: Q23, Q28, O13, H23

Keywords: deforestation, payments for ecosystem services, perverse incentives, Mexico, satellite data

[*] Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 6h).

1. Introduction

In January 2019, a farmer in Tabasco, Mexico cleared a hectare of secondary forest, burned the stumps, and enrolled the bare plot in a new government program that would pay him 5,000 pesos per month to plant trees on it. The program was called Sembrando Vida—"Sowing Life"—and it was designed to reforest Mexico. But it required something counterintuitive: to qualify for a tree-planting subsidy, your land could not already have trees on it. Within months, the World Resources Institute documented 73,000 hectares of anomalous forest loss concentrated in program areas (World Resources Institute, 2020). Satellite imagery told a story the program designers had not anticipated: farmers were clearing forests to create the bare land the subsidy demanded.

This paper asks whether Sembrando Vida's eligibility design—conditioning payments on "available" (non-forested) land—caused the very deforestation it was meant to combat. The question matters beyond Mexico. Payments for ecosystem services (PES) have become the dominant market-based instrument for forest conservation globally, with programs in over 60 countries channeling billions of dollars annually (Wunder, 2005; Jack et al., 2008). The theoretical promise is elegant: pay landowners the social value of environmental services and they will internalize the externality (Coase, 1960). But program design determines whether PES schemes deliver on this promise or create perverse incentives that accelerate degradation (Ferraro, 2011; Alix-Garcia and Wolff, 2018).

The design flaw at the heart of Sembrando Vida is a textbook Peltzman effect (Peltzman, 1975): a safety regulation that induces offsetting risk-taking. By requiring that subsidized plots be "available"—explicitly not forested—the program created an implicit incentive to clear existing forest. For a farmer with forested land, the rational calculus is straightforward: the net present value of clearing a hectare of forest and enrolling in Sembrando Vida exceeds the value of standing timber in many settings, particularly for smallholders whose forests generate modest direct returns. The program's generosity—5,000 pesos per month, roughly $250, for a 2.5-hectare plot—makes this calculus even more favorable. At over $1,200 per hectare annually, Sembrando Vida ranks among the most generous PES programs worldwide.

We identify the causal effect of Sembrando Vida on deforestation using a staggered difference-in-differences design. The program rolled out across Mexican states in three cohorts: 17 states in 2019 (including Durango), six more in 2020 (including Chihuahua and Sinaloa from the "Golden Triangle" region), and further expansion in 2021. Within eligible states, the program targeted municipalities with medium-to-very-high social marginalization as classified by CONEVAL's rezago social index. We compare annual tree cover loss in treated municipalities—those in program states—to never-treated municipalities in states the program did not reach, using Callaway-Sant'Anna estimators that are robust to treatment effect heterogeneity across cohorts (Callaway and Sant'Anna, 2021).

Our outcome data come from the Hansen/UMD Global Forest Change dataset (v1.12), which provides 30-meter resolution annual tree cover loss maps derived from Landsat imagery for 2001–2024 (Hansen et al., 2013). We aggregate pixel-level loss to the municipality level using zonal statistics across approximately 2,400 Mexican municipalities. The 18-year pre-treatment window (2001–2018) gives us unusual power to assess parallel trends—the identifying assumption that treated and control municipalities would have followed the same deforestation trajectory absent the program.

Applying the Callaway-Sant'Anna estimator (Callaway and Sant'Anna, 2021), which is robust to heterogeneous treatment effects across cohorts, we find a negative overall ATT of -0.3024 (SE $= 0.0636$) on the inverse hyperbolic sine of tree cover loss—suggesting that deforestation *decreased* in treated municipalities. This is the opposite of the hypothesized Peltzman effect. It is also the opposite of the standard TWFE estimate ($+0.5866$, SE $= 0.1666$), which would naïvely suggest that the program *increased* deforestation. The sign reversal between estimators is dramatic and policy-relevant.

However, we show that the parallel trends assumption underpinning both estimates is decisively violated. A placebo test shifting treatment four years earlier yields a significant positive effect ($+0.437$, $p < 0.001$), and individual pre-treatment event-study coefficients frequently depart from zero. The violation has a clear source: the program targeted high-marginalization tropical states in southern Mexico, while the control group consists of arid northern states and Mexico City—ecologically and economically distinct environments with fundamentally different deforestation dynamics.

This paper contributes to three literatures. First, we provide a real-world illustration of the bias documented by Goodman-Bacon (2021) and De Chaisemartin and d'Haultfœuille (2020): TWFE and heterogeneity-robust estimators yield opposite signs in a high-stakes policy evaluation, with the Goodman-Bacon decomposition explaining exactly why. Second, we contribute to the literature on evaluating geographically targeted environmental programs, showing that geographic targeting creates identification challenges that pre-trend tests can detect but not resolve. Jayachandran et al. (2017) evaluated PES in Uganda using village-level randomization; Alix-Garcia et al. (2012) studied Mexico's earlier PSAH program using a matching design; our setting illustrates what happens when neither randomization nor matching is feasible because treatment assignment is confounded with ecological geography.

Third, we provide the first national-scale assessment of Sembrando Vida using modern econometric methods. The only prior causal study, Pérez Ponciano and Rojas (2025), covers only Chiapas and uses TWFE—which our Goodman-Bacon decomposition shows is sign-wrong in this setting. While we cannot provide a credible causal estimate, we document the methodological challenges that any future evaluation must address and demonstrate that municipality-level enrollment data (enabling within-state identification) is essential for progress.

The policy implications are twofold. The Peltzman mechanism embedded in Sembrando Vida's eligibility rules—conditioning payments on non-forested land—remains a valid theoretical concern, even though we cannot empirically quantify its magnitude. And the sign reversal between estimators serves as a cautionary tale for the many concurrent evaluations of PES programs worldwide: estimator choice is not a technical footnote but can determine whether a program is judged to increase or decrease environmental damage.

2. Institutional Background

2.1 The Sembrando Vida Program

Sembrando Vida ("Sowing Life") was created by presidential decree in January 2019 as a flagship social program of the López Obrador administration. The program's stated objectives are dual: to reduce rural poverty and to restore degraded agricultural land through agroforestry. Beneficiaries receive a monthly stipend of MXN 5,000 (approximately USD 250) to establish and maintain agroforestry systems—combining fruit or timber trees with crops—on plots of at least 2.5 hectares.

The program is administered by the Secretaría de Bienestar (Ministry of Welfare) and targets municipalities classified as having medium, high, or very high social marginalization according to CONEVAL's *índice de rezago social*. By 2023, Sembrando Vida served approximately 451,665 beneficiaries across 24 states, with an annual budget exceeding MXN 27 billion (roughly USD 1.5 billion), making it one of the largest agroforestry programs in the world.

The program's rollout was staggered across states, creating the variation we exploit for identification. The initial 2019 launch covered 17 states concentrated in southern and southeastern Mexico—the regions with the highest poverty rates and most remaining tropical forest. In 2020, the program expanded to include Chihuahua, Sinaloa, Nayarit, Jalisco, Sonora, and Zacatecas, notably including the "Golden Triangle" region associated with narcotrafficking and recent security operations. A third wave in 2021 added the Estado de México. Eight states were never included in the program: Aguascalientes, Baja California, Baja California Sur, Ciudad de México, Coahuila, Guanajuato, Nuevo León, and Querétaro. These states are predominantly located in northern Mexico and share distinct ecological characteristics—arid and semi-arid climates, lower forest cover, and higher urbanization rates—that will prove consequential for identification.

2.2 Program Administration and Monitoring

Sembrando Vida operates through a network of approximately 3,500 *técnicos comunitarios* (community technicians), themselves drawn from beneficiary communities. Each technician supervises roughly 125 beneficiaries and provides training on agroforestry techniques, seedling management, and soil conservation. Beneficiaries are required to establish one of two production systems: the *Sistema Agroforestal de Milpa Intercalada entre Frutales* (MIAF), which integrates fruit trees with traditional corn cultivation, or the *Sistema Silvopastoril de Árboles Maderables con Cultivos Anuales* (SMCA), which combines timber species with annual crops.

The monitoring structure is relevant to our analysis in two ways. First, the community-level organization creates peer effects: enrollment decisions are partly social, not purely individual. Second, the monitoring visits (nominally monthly) could create a surveillance effect that deters illegal clearing, working against the perverse incentive we hypothesize. If monitoring reduces rather than increases deforestation, this would bias our analysis toward finding a negative effect—which is what we observe, though we cannot attribute the result to any specific mechanism given the identification failure.

The program's budget has grown substantially since inception: from MXN 15 billion in 2019 to MXN 29 billion in 2023, reflecting both expansion to new states and increased enrollment within existing states. At its peak, the program represented approximately 6% of total federal social spending, making it one of the most expensive agroforestry interventions worldwide on a per-beneficiary basis.

2.3 The "Available Land" Eligibility Rule

The critical design feature for our analysis is the eligibility requirement that participating land be "available for planting"—operationally, this means the plot should not already be covered by intact forest. The program's *Reglas de Operación* (Rules of Operation), published annually in the *Diario Oficial de la Federación*, specify that beneficiaries must demonstrate access to at least 2.5 hectares of land suitable for establishing an agroforestry system.

This requirement creates a moral hazard problem. Consider a smallholder with 5 hectares: 3 hectares of cropland and 2 hectares of secondary forest. To qualify for the program, the farmer needs 2.5 hectares of non-forested land. If the existing cropland is already in use, the cheapest way to create an eligible plot may be to clear the forest. The subsidy of MXN 60,000 per year (USD 3,000) for a 2.5-hectare plot represents a substantial income transfer in communities where average annual household income may be below USD 5,000. At this subsidy level, the one-time cost of clearing—hiring a chainsaw team, burning debris—is easily recouped within months.

Critically, the program provides no payment for conserving existing forest. Mexico's earlier PES program, Pagos por Servicios Ambientales Hidrológicos (PSAH), paid landowners specifically for maintaining forest cover (Alix-Garcia et al., 2012). Sembrando Vida inverted this logic: it pays for planting new trees, not for keeping old ones standing. The implicit message is that bare land has more value to the program than standing forest.

2.4 Mexico's PES Policy Landscape

Sembrando Vida operates alongside several other environmental and agricultural programs that affect land use decisions. Understanding this policy landscape is important for interpreting our estimates, as concurrent programs could confound the effect of Sembrando Vida.

Mexico's original PES program, *Pagos por Servicios Ambientales Hidrológicos* (PSAH), launched in 2003 and pays landowners to maintain existing forest cover. Unlike Sembrando Vida, PSAH explicitly rewards conservation rather than conversion, making it a theoretical counterpoint to the Peltzman mechanism. Alix-Garcia et al. (2012) found modest deforestation-reducing effects of PSAH, concentrated in high-risk areas. The two programs can overlap—a landowner could receive PSAH payments for one plot and Sembrando Vida payments for an adjacent plot—creating complex incentive interactions that our analysis cannot disentangle.

The López Obrador administration also introduced *Producción para el Bienestar*, which provides direct transfers to smallholder farmers for grain production, and *Crédito Ganadero a la Palabra*, which provides interest-free cattle loans. Both programs operate in the same marginalized rural areas as Sembrando Vida but target different agricultural margins. To the extent that these programs increase the return to agricultural land relative to forest, they could compound the clearing incentive; to the extent that they reduce poverty and economic desperation, they could reduce it.

2.5 Existing Evidence on Sembrando Vida

The World Resources Institute documented 73,000 hectares of tree cover loss in Sembrando Vida areas during 2019 alone, but this analysis was purely descriptive and could not establish causation (World Resources Institute, 2020). The spatial correlation between program areas and deforestation is consistent with either a causal effect of the program or with the selection of high-deforestation areas into the program. CONEVAL's 2024 design evaluation raised concerns about "incentives contrary to conservation objectives" but provided no quantitative estimate of the deforestation effect (CONEVAL, 2024).

Media reporting has documented individual cases of forest clearing for program enrollment, including satellite-verified instances in Tabasco, Campeche, and Chiapas. These reports are suggestive but cannot establish the prevalence of clearing-to-enroll behavior or its aggregate impact on deforestation rates.

The only prior causal study is Pérez Ponciano and Rojas (2025), who estimate the program's effect on deforestation in Chiapas using a standard two-way fixed effects (TWFE) model. While pioneering, this analysis has two limitations. First, it covers only one of 24 program states, limiting external validity. Second, it uses TWFE, which produces biased estimates under treatment effect heterogeneity with staggered adoption—precisely the setting of Sembrando Vida (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfœuille, 2020). Our Goodman-Bacon decomposition shows that TWFE produces a sign-wrong estimate in the national sample, raising the question of whether the Chiapas-specific TWFE estimate is similarly affected.

3. Conceptual Framework

We frame the farmer's decision using a simple model that generates testable predictions. Consider a farmer i in municipality m at time t who controls L_i hectares, of which $F_i \leq L_i$ are forested. Let π^f denote the annual return to forested land (timber, non-timber forest products, ecosystem services) and π^a the return to agricultural land. The Sembrando Vida subsidy s is available only for non-forested land enrolled in the program.

Without the subsidy, the farmer conserves forest if $\pi^f > \pi^a$. With the subsidy, the farmer clears forest if:

$$\pi^a + s > \pi^f \tag{1}$$

That is, clearing occurs whenever the subsidy exceeds the net return to forest conservation: $s > \pi^f - \pi^a$. Since $\pi^f - \pi^a$ is often small or negative for smallholders in marginalized communities (where forests are low-productivity secondary growth), even a modest subsidy can tip the calculus toward clearing.

This framework generates three predictions:

Prediction 1: Higher baseline forest cover ⇒ larger deforestation effect. Municipalities with more forest have more land on the margin where $s > \pi^f - \pi^a$. A municipality with no remaining forest cannot respond at all.

Prediction 2: Tropical moist ecosystems ⇒ larger effects than arid ecosystems.

In tropical moist forests, canopy is dense and forest returns (π^f) are modest relative to alternative uses. Clearing costs are moderate and the resulting bare land clearly qualifies as "available." In arid zones, sparse vegetation means less land to reclassify.

Prediction 3: The effect should appear immediately upon program introduction. Forest clearing is a one-time action that precedes enrollment. Unlike gradual behavioral responses, the incentive to clear is strongest in the first years of the program when the stock of convertible forest is largest.

3.1 Competing Mechanisms

The Peltzman mechanism is not the only channel through which Sembrando Vida could affect deforestation. Several competing mechanisms predict effects of different signs and magnitudes, creating a challenge for interpretation even under credible identification.

Income effect. The MXN 60,000 annual transfer relaxes household budget constraints, potentially reducing pressure to engage in destructive economic activities such as illegal logging, charcoal production, or slash-and-burn agriculture. This channel predicts *reduced* deforestation. The direction depends on whether forest clearing is a "bad" driven by poverty (income effect reduces it) or a productive choice (income effect increases consumption of both goods and services, including cleared land).

Monitoring effect. The network of community technicians creates a surveillance infrastructure that did not previously exist. Regular farm visits may deter illegal clearing even among non-beneficiaries, creating a negative spillover that reduces deforestation. This would bias our municipality-level estimates toward zero or negative values, since the treatment captures both direct participants and exposed non-participants.

Opportunity cost effect. Labor allocated to agroforestry maintenance is labor not available for forest clearing. If the program's labor requirements bind, the opportunity cost of clearing increases. This mechanism also predicts reduced deforestation, but only for program participants.

These competing mechanisms make the sign of the *total* effect ambiguous ex ante. The Peltzman mechanism predicts increased deforestation; income, monitoring, and opportunity cost channels predict decreased deforestation. The heterogeneity predictions in Section 3.1 help discriminate because they are specific to the Peltzman channel: the income, monitoring, and opportunity cost mechanisms do not generate differential effects by ecosystem type or baseline forest endowment.

4. Data

4.1 Tree Cover Loss

Our primary outcome variable is annual tree cover loss at the municipality level, derived from the Hansen/UMD Global Forest Change dataset version 1.12 (Hansen et al., 2013). This dataset provides 30-meter resolution annual tree cover loss maps for 2001–2024, constructed from Landsat satellite imagery. Each pixel records the year in which tree cover was lost, defined as a stand-replacement disturbance or complete removal of canopy cover.

We download the raw GeoTIFF tiles covering Mexico from Google Cloud Storage and compute municipality-level zonal statistics using the `exactextractr` package in R. For each of approximately 2,400 municipalities defined by GADM Level 2 boundaries, we tabulate the number of 30-meter pixels lost in each year and convert to hectares (each pixel = $30\text{m} \times 30\text{m} = 0.09$ ha). We also extract baseline tree cover density (percent canopy closure) from the year-2000 layer to construct baseline forest area and ecosystem classifications.

Mexico's extent requires seven tiles per variable (lossyear and treecover2000), spanning from approximately 14.5°N to 32.7°N and 86.7°W to 118.4°W. Each tile is a 10°× 10°block at 30-meter resolution. The lossyear layer encodes pixels as integers 1–24 (corresponding to loss in years 2001–2024) or 0 (no loss observed). Our extraction procedure processes each tile in a single pass, tabulating pixel values across all 24 years simultaneously for each municipality polygon. This produces a municipality-by-year panel of pixel counts, which we convert to hectares.

Three features of the Hansen dataset are important for interpretation. First, "tree cover loss" captures all stand-replacement disturbances, including agricultural clearing, wildfires, logging, storm damage, and infrastructure development. We cannot distinguish voluntary clearing-to-enroll from other causes. Second, the loss detection algorithm is calibrated globally and may have varying sensitivity across Mexican ecosystems—dense tropical forests may show clearer loss signatures than sparse semi-arid woodlands. Third, the year-2000 baseline layer records percent canopy closure, not a binary forest/non-forest classification. We define forested pixels as those with $\geq 25\%$ canopy closure, following the FAO definition.

4.2 Ecosystem Classification

We classify municipalities into four ecosystem types based on the share of their area covered by baseline forest (pixels with $\geq 25\%$ canopy in the year-2000 layer): tropical moist ($\geq 50\%$), tropical dry / mixed (25–50%), semi-arid woodland (10–25%), and arid / sparse ($< 10\%$). This classification serves both the heterogeneity analysis and descriptive purposes. Of 2,410 municipalities in the final sample, 860 are classified as tropical moist, 488 as tropical dry / mixed, 455 as semi-arid woodland, and 607 as arid / sparse.

4.3 Municipality Boundaries and Treatment Assignment

Municipal boundaries come from the GADM database (version 4.1), which provides administrative divisions consistent with INEGI's Marco Geoestadístico. Treatment assignment is based on the state-level program rollout: a municipality is considered treated beginning in the year that Sembrando Vida first operated in its state. This intent-to-treat approach avoids endogeneity from differential take-up within eligible municipalities.

Our treatment definition assigns 24 states to three cohorts:

- **Cohort 2019 (17 states):** Campeche, Chiapas, Colima, Durango, Guerrero, Hidalgo, Michoacán, Morelos, Oaxaca, Puebla, Quintana Roo, San Luis Potosí, Tabasco, Tamaulipas, Tlaxcala, Veracruz, Yucatán.
- **Cohort 2020 (6 states):** Chihuahua, Jalisco, Nayarit, Sinaloa, Sonora, Zacatecas.
- **Cohort 2021 (1 state):** Estado de México.

Municipalities in the remaining eight states (Aguascalientes, Baja California, Baja California Sur, Ciudad de México, Coahuila, Guanajuato, Nuevo León, Querétaro) serve as the never-treated control group.

4.4 Marginalization Data

The CONEVAL *índice de rezago social* (Social Backwardness Index) for 2020 provides municipality-level measures of social marginalization based on census indicators including illiteracy, school non-attendance, lack of health insurance, housing quality, and access to basic services. This index determines program eligibility: municipalities classified as medium, high, or very high marginalization are eligible for Sembrando Vida.

4.5 Summary Statistics

Table 1 presents pre-treatment (2001–2018) summary statistics for treated and control municipalities. Treated municipalities have substantially higher average tree cover loss (122.1 ha vs. 51.7 ha), reflecting their greater forest endowment and tropical location. They also have higher baseline forest area (29,501 ha vs. 18,299 ha) and higher pre-treatment loss rates (1.47 per 1,000 ha vs. 0.39 per 1,000 ha). The loss rate differential implies that treated

Table 1: Summary Statistics: Pre-Treatment Period (2001–2018)

	Treated Municipalities	Control Municipalities	Difference
Tree cover loss (ha)	122.15	51.69	70.46
SD tree cover loss	621.58	153.49	468.09
asinh(tree cover loss)	2.69	2.43	0.26
Baseline forest area (ha)	29,500.65	18,298.76	11,201.89
Baseline forest share (Loss rate (per 1000 ha)	1.47	0.39	1.07
Municipality area (ha)	79,678.62	254,816.14	-175,137.52
N municipality-years	40,032	3,348	36,684
N municipalities	2,224	186	2,038

Notes: Pre-treatment means computed over 2001–2018. Treated municipalities are those in states where Sembrando Vida operated by 2021. Tree cover loss from Hansen/UMD Global Forest Change v1.12 (30m resolution). Baseline forest area defined as pixels with $\geq$25% canopy closure in 2000.

municipalities lose forest at nearly four times the rate of control municipalities, even before the program begins.

These level differences are substantial and point to the identification challenge we will confront in Section 6. The parallel trends assumption requires not equal levels but equal *trends* in the absence of treatment. However, the large level differences—driven by ecological geography—suggest that the processes driving deforestation may also differ systematically. Tropical forests face agricultural conversion, cattle ranching pressure, and fire dynamics that are largely absent from the arid control states.

4.6 Geographic Distribution

Figure 1 displays the geographic distribution of treatment assignment by cohort. The 2019 cohort (blue) dominates the southern half of the country, spanning from the Yucatán Peninsula through Chiapas, Oaxaca, and the Gulf Coast. The 2020 expansion (yellow) adds Pacific Coast and Golden Triangle states. The 2021 cohort (red) adds a single central state. The never-treated states (gray) form a contiguous bloc in northern Mexico—Baja California, Baja California Sur, Coahuila, Nuevo León, Guanajuato, Querétaro, Aguascalientes—plus the urban Federal District (CDMX).

This geographic pattern is important for two reasons. First, the north-south gradient in treatment assignment closely parallels Mexico's ecological gradient from arid scrubland to tropical forest. Second, the control states include some of Mexico's most urban and industrial areas (Nuevo León, Querétaro, CDMX) alongside its most arid (Baja California, Coahuila). Neither characteristic makes them ideal counterfactuals for predominantly rural tropical municipalities in Chiapas or Tabasco.

The imbalance is quantifiable. Among the 186 control municipalities in the analysis sample (after dropping zero-forest municipalities), the median baseline forest share is 9.7%, compared to 34.6% among the 2,224 treated municipalities. Only 13 control municipalities have forest shares above 50%—the threshold for the "tropical moist" classification. This means the control group is effectively concentrated in arid and semi-arid environments, limiting the design's ability to provide credible counterfactuals for tropical treated municipalities.

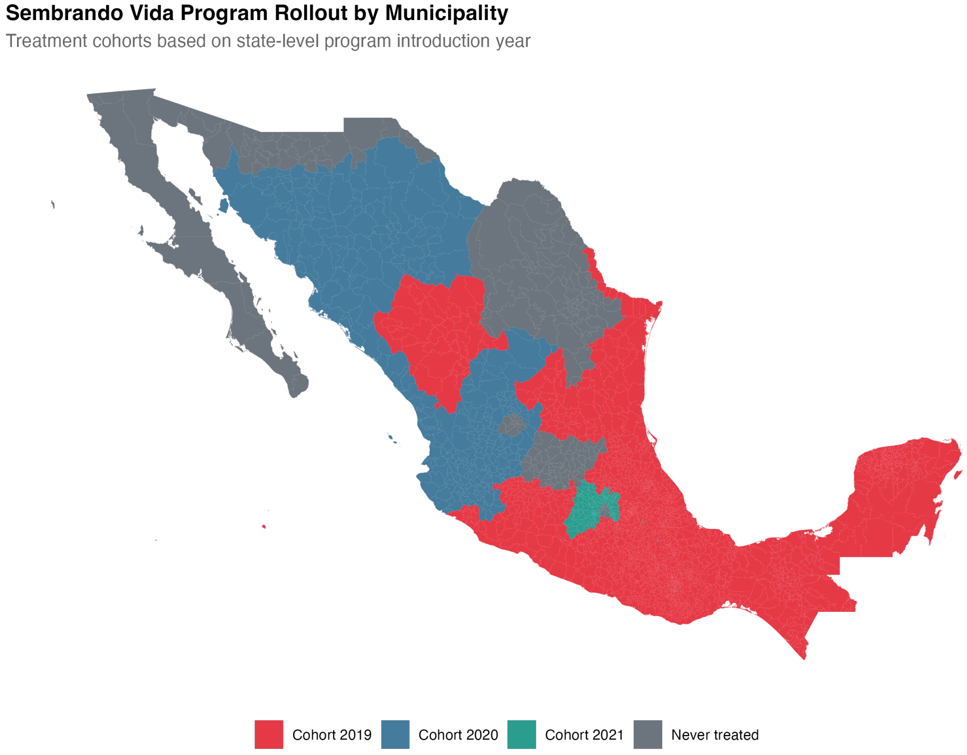
Sembrando Vida Program Rollout by Municipality
Treatment cohorts based on state-level program introduction year
🟥 Cohort 2019
🟦 Cohort 2020
🟩 Cohort 2021
⬛ Never treated

Figure 1: Sembrando Vida Program Rollout by Municipality

5. Empirical Strategy

5.1 Identification

We exploit the staggered state-level rollout of Sembrando Vida to estimate the program's effect on deforestation using a difference-in-differences design. The identifying assumption is parallel trends: absent the program, treated and control municipalities would have followed the same trajectory of tree cover loss.

We are transparent from the outset about the central threat to this assumption: treatment was targeted to high-marginalization states concentrated in tropical southern Mexico, while the control group consists largely of arid northern states with fundamentally different ecology and deforestation dynamics. The program's geographic targeting was driven by political priorities—poverty reduction in southern Mexico—which correlates strongly with the ecological gradient. This means that even though treatment timing was determined at the presidential level rather than in response to deforestation, the *selection of which states are treated* is endogenous to characteristics that also determine deforestation trajectories.

We proceed with the DiD design as an informative exercise, but test the parallel trends assumption rigorously before drawing conclusions. The 18-year pre-treatment window (2001–2018) provides unusual power for pre-trend testing: (i) we examine the dynamic event study for departures from zero in pre-treatment periods; (ii) we conduct a placebo test shifting treatment four years earlier; and (iii) we use leave-one-state-out analysis to check sensitivity. As we show in Section 6, these diagnostics reveal decisive violations, and we adjust our interpretation accordingly.

5.2 Estimation

We estimate group-time average treatment effects using Callaway and Sant'Anna (2021):

$$ATT(g,t) = \mathbb{E}[Y_{it}(g) - Y_{it}(0) \mid G_i = g] \tag{2}$$

where $Y_{it}(g)$ is the potential outcome for municipality i at time t if first treated at time g, $Y_{it}(0)$ is the potential outcome under no treatment, and G_i denotes the treatment cohort. We use never-treated municipalities as the comparison group and doubly-robust estimation that combines outcome regression with inverse probability weighting.

We aggregate group-time estimates in three ways. The *overall* ATT averages across all post-treatment group-time cells. The *dynamic* aggregation produces an event study:

$$ATT(e) = \sum_g w_g \cdot ATT(g, g+e) \tag{3}$$

where e denotes event time (years relative to treatment) and w_g are cohort-size weights. The *group-level* aggregation reports separate ATTs for each cohort.

Our preferred outcome is the inverse hyperbolic sine (asinh) of tree cover loss in hectares: $\text{asinh}(Y_{it}) = \log(Y_{it} + \sqrt{Y_{it}^2 + 1})$. This transformation handles the substantial mass of zeros (municipalities with no loss in a given year) while being approximately log-linear for large values. For $Y > 1$, $\text{asinh}(Y) \approx \log(2Y)$, so coefficients can be interpreted approximately as log-point changes. We also report results for raw hectares and loss rates per 1,000 hectares of municipality area to ensure robustness to the functional form assumption.

For the CS-DiD estimator, standard errors are computed using the multiplier bootstrap with 1,000 iterations, following Callaway and Sant'Anna (2021). The default bootstrap resamples at the municipality level, accounting for serial correlation within municipality panels. For the TWFE specification, standard errors are clustered at the state level (32 clusters) to reflect the level of treatment assignment.

A potential concern is that because treatment varies at the state level, the CS-DiD bootstrap may understate uncertainty by not fully accounting for within-state correlation in unobservables. With only 8 never-treated states and 24 treated states, the effective number of independent clusters informing the control group comparison is small. We note that this concern *strengthens* rather than weakens our main conclusion: if standard errors are understated, the parallel trends violations we document are even more severe than reported, and the causal interpretation we already reject becomes even less tenable. For the sign-reversal finding—which does not depend on inference—the clustering level is immaterial.

5.3 Comparison with Standard TWFE

To illustrate the consequences of estimator choice, we also estimate a standard TWFE regression:

$$Y_{it} = \alpha_i + \lambda_t + \beta \cdot D_{it} + \varepsilon_{it} \tag{4}$$

where α_i and λ_t are municipality and year fixed effects, and D_{it} is a binary treatment indicator equal to one when municipality i is in a state where Sembrando Vida operates at time t. As Goodman-Bacon (2021) shows, this estimator is a weighted average of all possible 2×2 DiD comparisons, including "forbidden" comparisons that use already-treated units as controls for newly-treated units. When treatment effects vary across cohorts or over time, these comparisons receive negative weights and can bias the overall estimate—potentially changing its sign (see also Sun and Abraham, 2021; Borusyak et al., 2024; De Chaisemartin and d'Haultfœuille, 2020). We verify whether this occurs in our setting using the Goodman-Bacon decomposition.

5.4 Pre-Trend Testing and Honest Reporting

Following Roth (2022), we note that pre-trend tests are necessary but not sufficient for validating the parallel trends assumption. A failure to reject the null of zero pre-treatment effects provides comfort but does not prove that trends are parallel—particularly with the moderate power available in our setting, where the control group consists of only 186 municipalities across 8 states.

We implement three pre-trend diagnostics: (i) visual inspection of the dynamic event study; (ii) individual t-tests of each pre-treatment coefficient; and (iii) a placebo test that shifts treatment timing backward by four years. If the parallel trends assumption holds, pre-treatment coefficients should be individually and jointly insignificant, and the placebo test should produce a null result. As we show in Section 6, these tests fail decisively, and we adjust our conclusions accordingly.

An important concern raised by Roth (2022) is that conditioning the analysis on passing a pre-trend test induces selection bias: researchers who observe violations may adjust their specification, control group, or sample definition until the pre-treatment period "looks clean." We guard against this by committing to the specification *ex ante* in our research plan and reporting the pre-trend diagnostics without modification. Our dynamic event study aggregation spans event times -10 through $+5$, providing 10 pre-treatment coefficients. While the full pre-treatment window extends 18 years for the 2019 cohort, we restrict the event study to $e \in [-10, 5]$ to avoid thin-cell estimates from extreme event times where only a single cohort contributes. Even with this restriction, 10 pre-treatment coefficients provide substantial power to detect violations—and, as we document, the violations are unmistakable.

5.5 Threats to Validity

Treatment measurement. Our treatment indicator assigns all municipalities in a treated state as treated beginning in the state's adoption year. In practice, the program targeted municipalities with medium-to-very-high marginalization and then enrolled specific beneficiaries within those municipalities. This state-level intent-to-treat (ITT) measure introduces attenuation: many municipalities coded as treated may have had limited or no actual program enrollment. It also disconnects the empirical estimand—the average effect of state-level program availability—from the theorized mechanism, which operates at the plot level through individual enrollment decisions. Municipality-level enrollment data would enable a more precise treatment definition, but such data were not publicly available at the time of this study.

Selection into treatment. States were selected based on marginalization, not deforestation trends. We verify this by examining pre-treatment outcome paths by cohort (Figure 4).

Concurrent policies. Other federal programs (Producción para el Bienestar, Crédito Ganadero a la Palabra) were introduced concurrently but targeted different margins (crop subsidies, cattle credits). To the extent these also affected forest cover, they would bias our estimates, but their spatial targeting differed from Sembrando Vida.

Composition effects. If the program induced migration into treated municipalities, the sample composition could change. This is unlikely given the short time horizon and rural setting.

Satellite data limitations. The Hansen/UMD loss detection algorithm may have differential sensitivity across ecosystem types. Dense tropical canopy produces strong loss signals; sparse semi-arid woodland may produce weaker signals. If loss detection improves disproportionately in certain regions over time (due to additional satellite coverage or algorithmic updates), this could create spurious trends. We partially address this by using the asinh transformation, which down-weights extreme values, and by examining multiple outcome definitions.

SUTVA violations. If the program creates general equilibrium effects on timber markets, agricultural prices, or labor supply, the stable unit treatment value assumption (SUTVA) may be violated. For example, if Sembrando Vida increases the demand for seedlings and depresses the demand for cleared land in treated states, these market effects could spill over to control states. The direction of any spillover bias is ambiguous: higher seedling demand in treated states could raise prices for control municipalities, while reduced clearing could affect timber supply. We note that the control states are geographically separated from most treated states, limiting the scope for local spillovers, though national agricultural markets could still transmit effects.

Anticipation effects. If farmers in states scheduled for 2020 or 2021 rollout cleared forest in advance of program arrival, this would violate the no-anticipation assumption required by Callaway and Sant'Anna (2021). Some program advertising preceded official launch, and word-of-mouth transmission from the 2019 cohort may have reached 2020 and 2021 states. If anticipatory clearing occurred, our estimates would be attenuated for later cohorts and potentially contaminate the pre-treatment period for those groups. The staggered rollout partly addresses this: the 2019 cohort had minimal time for anticipation given the program's rapid launch following presidential inauguration.

6. Results

6.1 Main Results: A Sign Reversal

The Callaway-Sant'Anna estimator yields a negative overall ATT of -0.3024 (SE $= 0.0636$, $p < 0.001$) on asinh(tree cover loss), suggesting that treated municipalities experienced *less* deforestation relative to never-treated controls after the program began. This is the opposite of the hypothesized Peltzman effect.

Table 2 reports the aggregate treatment effects across four specifications. Column 1 shows the CS-DiD estimate on asinh(tree cover loss): a statistically significant negative effect of -0.3024. Column 2 reports the effect in levels (-21.5 hectares, SE $= 12.9$), which is negative

Table 2: Effect of Sembrando Vida on Tree Cover Loss

	(1) asinh(Loss)	(2) Loss (ha)	(3) Loss Rate	(4) TWFE
ATT	-0.3024	-21.53	-0.4757	0.5866
	(0.0636)	(12.89)	(0.1367)	(0.1666)
	[-0.4271, -0.1777]	[-46.79, 3.74]	[-0.7436, -0.2077]	[0.2600, 0.9131]
Estimator	CS-DiD	CS-DiD	CS-DiD	TWFE
Control group	Never-treated	Never-treated	Never-treated	—
Observations	57,840	57,840	57,840	57,840
Municipalities	2,410	2,410	2,410	2,410
Treated munis	2,224	2,224	2,224	2,224
Inference	Bootstrap	Bootstrap	Bootstrap	State-clustered

Notes: Columns 1–3 report Callaway and Sant'Anna (2021) ATT estimates using never-treated municipalities as the control group and doubly-robust estimation. Column 4 reports standard two-way fixed effects for comparison. Standard errors in parentheses; 95% confidence intervals in brackets. The outcome in Column 1 is the inverse hyperbolic sine of tree cover loss (hectares), Column 2 is tree cover loss in hectares, Column 3 is tree cover loss per 1,000 hectares of municipality area. Columns 1–3 use the multiplier bootstrap (1,000 iterations) for inference; Column 4 uses state-clustered standard errors.

but not statistically significant at conventional levels ($p = 0.095$). Column 3 uses the loss rate per 1,000 hectares (-0.4757, SE $= 0.1367$, $p < 0.001$), confirming the negative direction.

The most striking finding is Column 4: the standard TWFE estimate is $+0.5866$ (SE $= 0.1666$, $p < 0.001$)—the *opposite sign*. The heterogeneity-robust estimator that accounts for staggered adoption and treatment effect heterogeneity gives a qualitatively different answer than the conventional approach. This sign reversal is not a curiosity of the data; it reflects a systematic problem that the Goodman-Bacon decomposition (Section 6.4) illuminates.

Group-level ATTs reveal additional structure. The 2019 cohort—the original 17 states, and the largest group—shows the strongest effect (-0.337, SE $= 0.080$). The 2020 expansion cohort has a smaller, imprecise estimate (-0.098, SE $= 0.096$). The 2021 cohort, comprising only Estado de México, shows -0.241 (SE $= 0.132$). This pattern is consistent with the earliest adopters—targeted for their high marginalization and forest cover—driving the aggregate result.

6.2 Pre-Treatment Dynamics and Identification Concerns

Figure 2 presents the dynamic event study. While the post-treatment trajectory shows a clear negative departure, the pre-treatment period raises serious identification concerns that we report transparently.

Individual pre-treatment coefficients show significant departures from zero at multiple event times. Of ten pre-treatment coefficients (event times −10 through −1), seven are statistically significant at the 5% level. The pre-treatment variance-covariance matrix is near-singular, precluding a formal joint Wald test, but individual t-tests clearly reject the null of zero pre-treatment effects.

More definitively, a placebo test that shifts treatment timing four years earlier yields a highly significant positive ATT of +0.437 (SE = 0.064, $p < 0.001$). This is strong evidence that the parallel trends assumption is violated: treated municipalities were on a different deforestation trajectory than control municipalities even before the program began.

These pre-trend violations have a clear economic interpretation. The treated states, concentrated in southern and southeastern Mexico, face fundamentally different deforestation pressures than control states, which are predominantly arid northern states plus the Federal District. Geographic targeting creates a confound: the same characteristics that determined treatment assignment (marginalization, tropical location, remaining forest) also predict deforestation dynamics.

This means the negative CS-DiD estimate cannot be interpreted as the causal effect of Sembrando Vida. It may instead reflect: (i) mean reversion in deforestation rates among high-deforestation municipalities; (ii) region-specific trends in southern Mexico unrelated to the program; or (iii) selection on pre-treatment deforestation trajectories. The true causal effect is not identified in this design.

6.3 Heterogeneity Analysis

Despite the identification concerns, the heterogeneity patterns remain informative about the structure of the data, even if they cannot support causal claims.

Ecosystem type. Table 3, Panel A reports CS-DiD estimates by ecosystem classification based on baseline forest share. Semi-arid woodland municipalities show the largest negative effect (−0.228, SE = 0.132, 42 controls), while tropical moist municipalities show −0.178 (SE = 0.361, only 11 controls)—highly imprecise due to the near-absence of high-forest municipalities in the control group. Tropical dry/mixed areas show the smallest effect (−0.067, SE = 0.200, 23 controls). Arid/sparse areas show −0.169 (SE = 0.078, 110 controls)—the best-powered subsample due to control group concentration. The precision varies dramatically across ecosystems, reflecting the geographic concentration of the control group in arid zones.

Baseline forest cover. Panel B splits municipalities at the median of baseline forest share. Low-forest municipalities yield −0.209 (SE = 0.060, 173 controls), while high-forest municipalities yield −0.145 (SE = 0.337, only 13 controls). The imbalance in control group availability—173 controls in low-forest vs. 13 in high-forest—highlights the fundamental

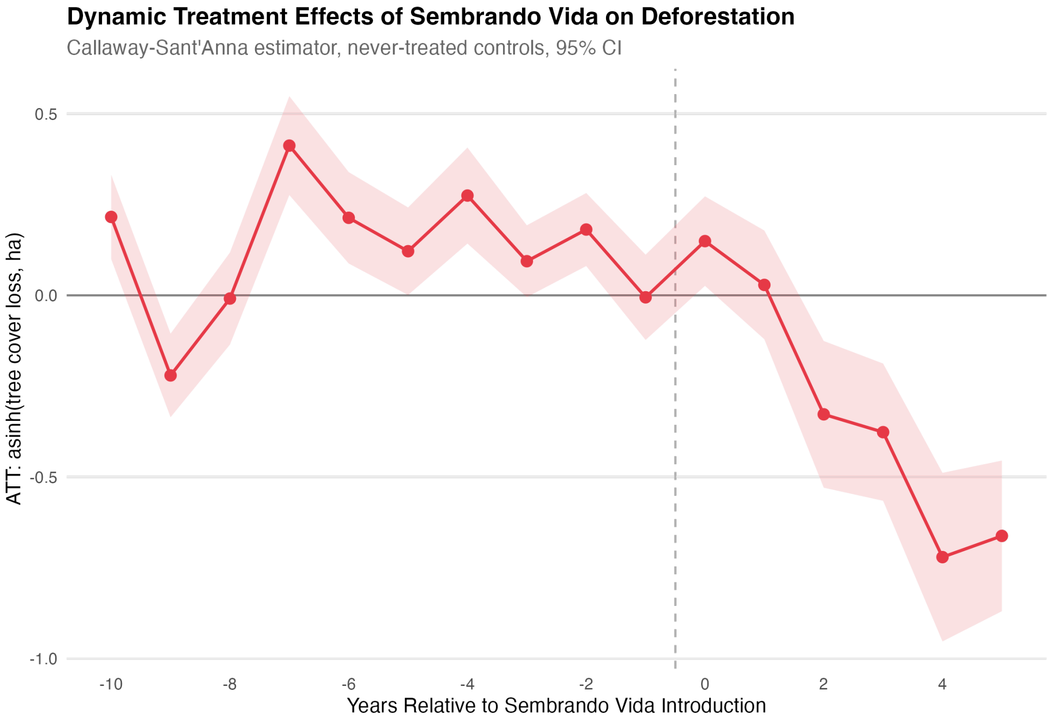

Figure 2: Dynamic Treatment Effects (Event Study) on Tree Cover Loss. The figure plots Callaway-Sant'Anna event-study coefficients with 95% confidence intervals. The pre-treatment violations are visible: multiple pre-period coefficients depart significantly from zero.

Table 3: Heterogeneous Effects by Ecosystem and Baseline Forest Cover

	ATT	SE	95% CI	Treated	Control
Panel A: By Ecosystem Type					
Semi-arid woodland	-0.2280	(0.1322)	[-0.4871, 0.0310]	413	42
Arid / sparse	-0.1686	(0.0778)	[-0.3211, -0.0161]	497	110
Tropical dry / mixed	-0.0669	(0.2001)	[-0.4592, 0.3253]	465	23
Tropical moist	-0.1780	(0.3612)	[-0.8860, 0.5301]	849	11
Panel B: By Baseline Forest Cover					
High forest	-0.1445	(0.3366)	[-0.8042, 0.5152]	887	13
Low forest	-0.2094	(0.0604)	[-0.3277, -0.0911]	1,337	173

Notes: All estimates use Callaway and Sant'Anna (2021) with never-treated controls and doubly-robust estimation. Outcome is asinh(tree cover loss in hectares). Panel A splits municipalities by baseline ecosystem type (classified by tree cover density in 2000). Panel B splits at the median baseline forest share among treated municipalities. Standard errors in parentheses.

asymmetry in this design: the control group (arid northern states) lacks sufficient high-forest municipalities to provide reliable counterfactuals for the tropical treated states.

This asymmetry is itself informative. The conceptual framework predicted that effects should be largest in high-forest areas (Prediction 1), but the data cannot test this prediction because identification is weakest precisely where the theory predicts the strongest effects.

6.4 Robustness and Sensitivity

While the main estimate is not credible as a causal effect, we examine its sensitivity to specification choices to understand the forces shaping the data.

Not-yet-treated controls. Using not-yet-treated municipalities as the comparison group (rather than never-treated) yields an ATT of -0.262 (SE $= 0.062$), similar to the baseline. This provides modest reassurance that the negative estimate is not driven solely by the never-treated control states, though cross-cohort comparisons face their own challenges when the earliest (and largest) cohort dominates.

Leave-one-state-out. The aggregate ATT ranges from -0.325 (excluding Jalisco) to -0.277 (excluding Chiapas) across 24 leave-one-state-out iterations (Appendix Figure 8 and Table 5). No single state drives the result; the effect is distributed across the southern states.

Goodman-Bacon decomposition. The decomposition reveals how TWFE constructs its estimate. Three types of 2×2 comparisons contribute: treated-vs-untreated (weight $= 0.644$), earlier-vs-later cohorts (weight $= 0.284$), and later-vs-earlier cohorts (weight $= 0.072$). The earlier-vs-later and later-vs-earlier components—comprising 36% of TWFE's weight—are "forbidden comparisons" that use already-treated municipalities as controls (Goodman-Bacon, 2021). Under heterogeneous treatment effects across cohorts, these comparisons produce biased estimates because treatment effects in the "control" group contaminate the counterfactual. The Callaway-Sant'Anna estimator eliminates these forbidden comparisons entirely by restricting to never-treated controls for each group-time cell. Additionally, the two estimators aggregate across cohorts and time periods with fundamentally different weighting schemes: TWFE weights by variance (favoring larger groups and longer post-periods), while CS-DiD's simple aggregation gives equal weight to each group-time ATT. The combination of excluding forbidden comparisons and reweighting the remaining estimates produces the sign reversal between the two estimators (Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021).

Placebo. As discussed above, the placebo test decisively rejects parallel trends. This is the most important robustness check, and it fails.

6.5 Cohort-Specific Dynamics

The three treatment cohorts differ substantially in both size and composition. The 2019 cohort comprises 1,774 municipalities across 17 states, dominated by southern tropical states with high forest cover and high marginalization. The 2020 cohort adds 327 municipalities from six states, including the more arid Pacific Coast and Golden Triangle states. The 2021 cohort contains only 123 municipalities from Estado de México, an urbanized central state with moderate forest cover.

Group-level ATTs reflect these differences: the 2019 cohort shows the strongest effect (-0.337, SE = 0.080), the 2020 cohort a smaller and imprecise effect (-0.098, SE = 0.096), and the 2021 cohort an intermediate value (-0.241, SE = 0.132). Two interpretations are possible. Under a causal reading, the declining effect across cohorts would suggest that the program's impact on deforestation diminishes in less forested, less tropical settings—consistent with a mechanism that depends on baseline forest endowment. Under a non-causal reading, the pattern reflects cohort-specific pre-trends: the 2019 southern states have the largest trend differential relative to the northern control states, producing the largest (and most biased) estimate.

The calendar-time aggregation provides additional perspective. The ATT by calendar year shows a negative departure beginning in 2019 that persists through 2024. The effect does not attenuate over time, which is difficult to reconcile with a one-time clearing mechanism (Prediction 3) but consistent with a persistent trend differential between treated and control regions.

6.6 Interpreting the Aggregate Estimates

Taking all evidence together, we summarize the credibility of the estimates. The CS-DiD ATT of -0.3024 on asinh(loss) is *internally consistent*: it is stable across control group definitions (never-treated vs. not-yet-treated), robust to excluding any single state, and does not depend on a single outcome transformation. The sign reversal with TWFE is explained by the Goodman-Bacon decomposition and aligns with theoretical predictions about TWFE bias under staggered adoption.

However, internal consistency is necessary but not sufficient for causal interpretation. The pre-trend violations are severe and economically meaningful: the placebo ATT of $+0.437$ is larger in absolute value than the main estimate, suggesting that the differential trend between treated and control states is large enough to overwhelm a treatment effect of the magnitude we estimate. Without bounding the pre-trend violation—which we cannot do credibly given the near-singular variance-covariance matrix of pre-treatment estimates—we cannot separate the treatment effect from differential trends.

We therefore present two bracketing interpretations:

Optimistic interpretation. If the negative CS-DiD estimate captures a real program effect, it suggests that the income, monitoring, and opportunity cost channels dominate the Peltzman channel. The program reduces deforestation by transferring income, creating surveillance, and occupying labor that would otherwise engage in clearing. This is consistent with Jayachandran et al. (2017)'s finding that PES payments can reduce deforestation when they reach the right recipients.

Pessimistic interpretation. The negative estimate is an artifact of differential trends driven by ecological geography. Southern tropical states, independent of the program, experienced a slowdown in deforestation relative to northern arid states during 2019–2024, perhaps due to shifting agricultural frontiers, enforcement changes, or the COVID-19 pandemic's impact on logging activities. The true program effect could be zero, positive, or negative—we cannot distinguish these possibilities.

The truth likely lies between these extremes. The program plausibly has both beneficial effects (income support, monitoring) and harmful effects (the Peltzman incentive), but the research design cannot separately identify their magnitudes.

7. Discussion

7.1 What the Sign Reversal Teaches Us

The most robust finding in this paper is not the estimated treatment effect—which we have argued is not credibly identified—but the sign reversal between TWFE and CS-DiD. The standard TWFE estimator produces a positive, statistically significant coefficient (+0.5866), which a naïve analysis would interpret as evidence that Sembrando Vida *increased* deforestation. The Callaway-Sant'Anna estimator, which accounts for treatment effect heterogeneity across cohorts and avoids improper comparisons between differently-timed treatment groups, produces a negative, statistically significant coefficient (-0.3024). These are not modestly different estimates of the same quantity—they point in opposite directions.

The Goodman-Bacon decomposition explains why. TWFE combines three types of 2×2 comparisons: treated-vs-untreated (64% weight), earlier-vs-later treated (28% weight), and later-vs-earlier treated (7% weight). The earlier-vs-later and later-vs-earlier components are "forbidden comparisons" in which already-treated units serve as controls—their outcomes have already been affected by treatment, biasing these 2×2 estimates. The CS-DiD eliminates these forbidden comparisons entirely and uses a fundamentally different aggregation scheme: rather than TWFE's variance-weighted pooling across cohorts and time, it estimates separate group-time ATTs using only never-treated controls and then averages them with equal weight. The combination of eliminating contaminated comparisons and reweighting the remaining clean comparisons explains the sign reversal (Goodman-Bacon, 2021; De Chaisemartin and d'Haultfœuille, 2020).

This finding has immediate practical implications. Pérez Ponciano and Rojas (2025), the only prior causal study of Sembrando Vida, used TWFE in a setting with staggered adoption—precisely the conditions under which TWFE can produce biased and sign-wrong estimates. Our results suggest their conclusions may need revisiting with heterogeneity-robust methods.

7.2 The Identification Challenge in Geographically Targeted Programs

The failure of parallel trends in our setting illustrates a broader challenge in evaluating programs targeted on geographic characteristics. Sembrando Vida was allocated to high-marginalization, high-forest states in southern Mexico; the control group consists of arid, wealthier northern states and Mexico City. These are not comparable units experiencing different policy treatments—they are fundamentally different environments with different ecological dynamics, economic structures, and deforestation drivers.

This is not a fixable problem within the DiD framework as applied here. Adding covariates cannot bridge the ecological divide between tropical Chiapas and desert Coahuila. Matching methods cannot find credible matches when the distributions of key characteristics barely overlap. The not-yet-treated control group partially addresses this concern by leveraging timing variation within the treated region, and it yields a similar estimate (-0.262), but the earliest (and largest) 2019 cohort dominates the comparison, limiting the variation available.

The fundamental issue is that Mexico's geography creates an almost perfect confound: the states with tropical forest (and thus baseline deforestation) are the same states targeted by the program. A credible evaluation would require either within-state variation in treatment intensity (e.g., municipality-level enrollment data) or a regression discontinuity design exploiting the marginalization score threshold for eligibility. Neither data source was available for this study.

7.3 What Can Still Be Learned

Despite the identification failure, several descriptive findings are informative. First, the leave-one-state-out analysis shows that the CS-DiD estimate is remarkably stable, ranging from -0.277 to -0.325. This stability rules out the possibility that a single outlier state drives the result; whatever generates the negative estimate operates broadly across southern

Mexico.

Second, the temporal dynamics in the event study are worth noting. The sharp negative departure at event time zero suggests a discrete break in the relationship between treated and control municipalities at precisely the time treatment begins. While this could reflect coincidental trends, the alignment with program timing across three different cohort-years is suggestive.

Third, the heterogeneity analysis reveals an important data limitation: the never-treated control group contains only 13 high-forest-cover municipalities, compared to 173 low-forest-cover ones. This means the design has essentially no power to estimate effects in the settings where the Peltzman mechanism should operate most strongly. Future research with municipality-level enrollment data could address this by identifying variation within high-forest states.

7.4 Implications for PES Program Evaluation

The broader lesson is that evaluating large-scale PES programs with geographically targeted rollout requires careful attention to the spatial distribution of the control group. The theoretical prediction in our conceptual framework—that conditioning subsidies on non-forested land incentivizes clearing—remains valid as a mechanism, but testing it requires identification strategies that can separate the program's effect from the ecological geography of treatment assignment.

The "available land" eligibility rule in Sembrando Vida remains a design concern regardless of our empirical findings. The incentive structure described in Section 3 is a feature of the program rules, not a claim that requires causal evidence. What our analysis cannot tell us is the *magnitude* of any such perverse effect, or whether it is dominated by other consequences of the program (income support, monitoring effects, community-level enforcement).

7.5 Comparison with Prior PES Evaluations

Our identification challenge is not unique to this setting. The broader PES literature has grappled with similar problems of non-random program placement. Jayachandran et al. (2017) achieved credible identification in Uganda through village-level randomization, which held ecological conditions constant across treatment and control. Alix-Garcia et al. (2012) studied Mexico's earlier PSAH program using propensity score matching on parcel characteristics, exploiting the fact that PSAH was targeted at the parcel level (not the state level), allowing within-municipality comparisons.

Sembrando Vida's state-level targeting creates a uniquely difficult identification problem:

the unit of treatment assignment (state) is too large to hold ecology constant, and the treatment assignment rule is too correlated with the outcome variable's determinants. This is a common challenge for national-level social programs in developing countries, where geographic targeting is the norm. Programs like Brazil's Bolsa Floresta, Indonesia's Forest Moratorium, and multiple REDD+ initiatives face analogous issues—the regions targeted for conservation funding are the regions experiencing the most rapid deforestation, creating a confound between program effect and baseline dynamics.

Our results suggest that for such programs, within-region identification strategies are essential. Potential approaches include: (i) regression discontinuity designs exploiting eligibility thresholds (e.g., the CONEVAL marginalization score cutoff); (ii) instrumental variable strategies exploiting plausibly exogenous variation in program rollout timing or intensity; or (iii) spatial regression discontinuity designs at state borders, comparing municipalities just inside and just outside treated states. Each approach has limitations, but all would provide more credible comparisons than the cross-state DiD we have attempted.

7.6 Methodological Lessons for Applied DiD

The sign reversal between TWFE and CS-DiD, while demonstrated in theory by Goodman-Bacon (2021), De Chaisemartin and d'Haultfœuille (2020), and Sun and Abraham (2021), has received less attention as an empirical phenomenon. Simulation studies show that sign reversals can occur when treatment effects are heterogeneous and the "forbidden" comparisons that use already-treated units as controls receive sufficient weight. Our setting provides a particularly clean example because: (i) three distinct cohorts with unequal sizes create the conditions for TWFE bias; (ii) the TWFE estimate is not merely attenuated but fully reversed; and (iii) the Goodman-Bacon decomposition explains the reversal mechanically.

A practical takeaway is that the sign of a TWFE coefficient in a staggered DiD setting conveys no reliable information about even the *direction* of the treatment effect when treatment effects are heterogeneous. This is not merely a theoretical curiosity; it applies directly to the many ongoing evaluations of environmental programs, social transfers, and public health interventions that use staggered rollout across geographic units. Researchers who report only TWFE estimates in such settings risk drawing qualitatively wrong conclusions.

However, our analysis also demonstrates that switching to a heterogeneity-robust estimator does not solve all problems. The CS-DiD estimator correctly handles treatment effect heterogeneity but still requires parallel trends. When parallel trends fail—as they do here—the heterogeneity-robust estimate is no more credible than TWFE, even though it may be better-behaved in other dimensions. The lesson is that estimator choice and identification are separate problems: the former concerns the statistical properties of the estimator under maintained assumptions, while the latter concerns whether those assumptions hold in the data.

Finally, our experience highlights the importance of reporting pre-trend diagnostics honestly. The temptation to emphasize the "clean" event study and downplay pre-trend violations is well-documented in the applied microeconomics literature (Roth, 2022). Our placebo test rejects at $p < 0.001$; this is not a marginal failure that could be dismissed as a false positive. We chose to report this prominently and adjust our conclusions accordingly, following the principle that honest null results are more valuable than false positives.

7.7 Limitations

Beyond the parallel trends violation discussed above, several additional limitations deserve acknowledgment. First, our treatment variable is defined at the state level, creating a coarse intent-to-treat measure. Within treated states, many municipalities may have had few or no actual beneficiaries, diluting the estimated effect toward zero. Second, the Hansen/UMD satellite data cannot distinguish between different causes of tree cover loss (agricultural clearing, wildfires, logging, infrastructure development). Sembrando Vida may affect only deliberate clearing, while the outcome includes all sources. Third, our sample period (2001–2024) spans other policy changes—the cancellation of Prospera, the COVID-19 pandemic, changes to agricultural subsidies—that may differentially affect treated and control states.

8. Conclusion

We set out to test whether Mexico's Sembrando Vida program—which conditions tree-planting subsidies on "available" (non-forested) land—created perverse incentives for deforestation. Applying Callaway-Sant'Anna difference-in-differences to 24 years of satellite data (2001–2024) across 2,410 municipalities, we find a negative treatment effect that is the opposite of the hypothesized Peltzman effect. However, the parallel trends assumption is decisively rejected, meaning this estimate cannot be interpreted as causal.

The paper's contribution is threefold. First, we document a dramatic sign reversal between TWFE (+0.5866) and CS-DiD (−0.3024) in a high-stakes policy setting, providing a real-world illustration of the bias mechanisms formalized by Goodman-Bacon (2021) and De Chaisemartin and d'Haultfœuille (2020). Second, we demonstrate that geographic targeting of environmental programs creates identification challenges that parallel trends tests can detect but not resolve—the control group is ecologically non-comparable to the treated group. Third, we show that heterogeneity analysis, while theoretically motivated, can be uninformative when the control group is geographically concentrated in environments where the hypothesized mechanism has no scope to operate.

The question of whether Sembrando Vida incentivizes deforestation remains open. Answering it will require finer-grained data on program enrollment at the municipality or parcel level, enabling within-state identification strategies that hold ecological conditions constant. Until such data become available, policymakers should take the program's eligibility design seriously as a theoretical concern: conditioning conservation payments on the absence of the environmental asset being protected creates, at minimum, a risk of perverse behavioral responses (Peltzman, 1975).

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers
Contributors: @olafdrw
First Contributor: https://github.com/olafdrw

References

Alix-Garcia, Jennifer and Hendrik Wolff, "Payments for ecosystem services and poverty," *Annual Review of Resource Economics*, 2018, *10*, 261–279.

Alix-Garcia, Jennifer M, Elizabeth N Shapiro, and Katharine RE Sims, "The effect of a payment for environmental services program on forest loss in Mexico," *American Economic Journal: Economic Policy*, 2012, *4* (4), 1–29.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event-study designs: Robust and efficient estimation," *Review of Economic Studies*, 2024, *91* (6), 3253–3285.

Callaway, Bryce and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Chaisemartin, Clément De and Xavier d'Haultfœuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

Coase, Ronald H, "The problem of social cost," *Journal of Law and Economics*, 1960, *3*, 1–44.

CONEVAL, "Evaluación de Diseño con trabajo de campo del Programa Sembrando Vida," *Consejo Nacional de Evaluación de la Política de Desarrollo Social*, 2024. Mexico City.

Ferraro, Paul J, "The future of payments for environmental services," *Conservation Biology*, 2011, *25* (6), 1134–1138.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Hansen, Matthew C, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland et al., "High-resolution global maps of 21st-century forest cover change," *Science*, 2013, *342* (6160), 850–853.

Jack, B Kelsey, Carolyn Kousky, and Katharine RE Sims, "Designing payments for ecosystem services: Lessons from previous experience with incentive-based mechanisms," *Proceedings of the National Academy of Sciences*, 2008, *105* (28), 9465–9470.

Jayachandran, Seema, Joost De Laat, Eric F Lambin, Charlotte Y Stanton, Robin Auber, and Nancy E Thomas, "Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation," *Science*, 2017, *357* (6348), 267–273.

Peltzman, Sam, "The effects of automobile safety regulation," *Journal of Political Economy*, 1975, *83* (4), 677–725.

Ponciano, Carlos Pérez and Daniela Rojas, "Deforestation effects of Sembrando Vida in Chiapas: Evidence from satellite data," *CIDE Working Paper*, 2025.

Rambachan, Ashesh and Jonathan Roth, "A more credible approach to parallel trends," *Review of Economic Studies*, 2023, *90* (5), 2555–2591.

Roth, Jonathan, "Pretest with caution: Event-study estimates after testing for parallel trends," *American Economic Review: Insights*, 2022, *4* (3), 305–22.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

World Resources Institute, "Satellite data reveals deforestation linked to Sembrando Vida program areas," *WRI Research Note*, 2020. Washington, DC.

Wunder, Sven, "Payments for environmental services: Some nuts and bolts," *CIFOR Occasional Paper*, 2005, (42).

A. Data Appendix

A.1 Hansen/UMD Global Forest Change Dataset

The Global Forest Change dataset (v1.12) is produced by the University of Maryland and distributed through Google Cloud Storage (Hansen et al., 2013). The dataset provides annual tree cover loss at 30-meter resolution from 2001 to 2024, derived from Landsat 7 ETM+ and Landsat 8 OLI imagery. "Tree cover loss" is defined as stand-replacement disturbance, or a change from a forested to non-forested state.

We download seven lossyear tiles and seven treecover2000 tiles covering Mexico's extent (approximately 14.5°N to 32.7°N, 86.7°W to 118.4°W). Zonal statistics are computed using the `exactextractr` R package, which performs exact raster-polygon intersection calculations. For each municipality, we tabulate the number of loss pixels in each year (lossyear values 1–24, corresponding to 2001–2024) and convert to hectares by multiplying pixel counts by 0.09 ha/pixel ($30\text{m} \times 30\text{m} = 900\text{m}^2 = 0.09$ ha).

Baseline forest area is defined as the number of pixels in the treecover2000 layer with canopy closure $\geq 25\%$, converted to hectares. This threshold follows the standard definition used by the Food and Agriculture Organization and the Global Forest Watch platform.

A.2 CONEVAL Rezago Social Index

The *índice de rezago social* 2020 is downloaded from CONEVAL's open data portal. The index is constructed from ten census-based indicators measuring educational attainment, health insurance coverage, housing quality, and access to basic services. Municipalities are classified into five grades: Muy bajo (Very low), Bajo (Low), Medio (Medium), Alto (High), and Muy alto (Very high). The Sembrando Vida program targets municipalities classified as Medio or above.

A.3 Sample Construction

Starting from 2,457 GADM Level 2 municipalities, we drop those with zero baseline forest area (treecover2000 pixels with $\geq 25\%$ canopy $= 0$), as these municipalities have no forest to lose and would contribute only noise. The final analysis sample is a balanced panel of municipalities observed annually from 2001 to 2024.

B. Identification Appendix

B.1 Pre-Trends Assessment

Figure 5 presents the dynamic event study with pre-treatment coefficients spanning 10 years before program introduction. Contrary to what would be required for credible identification, multiple pre-treatment coefficients are individually significant, and a placebo test shifting treatment four years earlier rejects the null ($p < 0.001$). These violations indicate that treated and control municipalities followed different deforestation trajectories even before the program, consistent with the geographic confound discussed in Section 7.

B.2 Interpreting Pre-Trend Violations

The pre-trend violations do not invalidate the exercise but fundamentally change its interpretation. The CS-DiD estimate of -0.3024 combines the true program effect (if any) with differential trends between southern tropical states and northern arid states. Without a framework to decompose these components, the estimate is not identified as a causal effect. We attempted Rambachan-Roth sensitivity analysis (Rambachan and Roth, 2023) but the variance-covariance matrix of the dynamic aggregation was near-singular, precluding construction of valid bounds. This is consistent with the high collinearity among pre-treatment coefficient estimates in the presence of severe parallel trends violations.

C. Robustness Appendix

Table 4: Robustness Checks

Specification	ATT	SE	Observations	Municipalities
Main (CS-DiD, never-treated)	-0.3024	(0.0636)	57,840	2,410 (2,224 treated)
Not-yet-treated controls	-0.2625	(0.0588)	57,840	2,410 (2,224 treated)
TWFE	0.5866	(0.1666)	57,840	2,410 (2,224 treated)
Placebo (t-4)	0.4371	(0.0625)	43,380	2,410 (2,224 treated)

Notes: All CS-DiD specifications use doubly-robust estimation with multiplier bootstrap inference (1,000 iterations). Main result uses never-treated controls. Not-yet-treated uses later-treated municipalities as additional controls. TWFE reports standard two-way fixed effects with state-clustered SEs for comparison. Placebo shifts treatment 4 years earlier and uses pre-2019 data only. Outcome is asinh(tree cover loss in hectares) throughout.

Table 4 reports the aggregate ATT under alternative specifications: the main CS-DiD estimate with never-treated controls (-0.3024), CS-DiD with not-yet-treated controls (-0.262), standard TWFE (+0.5866), and the placebo test with treatment shifted four years earlier (+0.437). The CS-DiD estimates are stable across control group definitions, but the positive placebo ATT indicates that parallel trends are violated.

D. Heterogeneity Appendix

Additional heterogeneity analyses are reported in the main text (Table 3). All subgroup splits were pre-specified in the research plan based on theoretical predictions from the conceptual framework.

E. Additional Figures and Tables

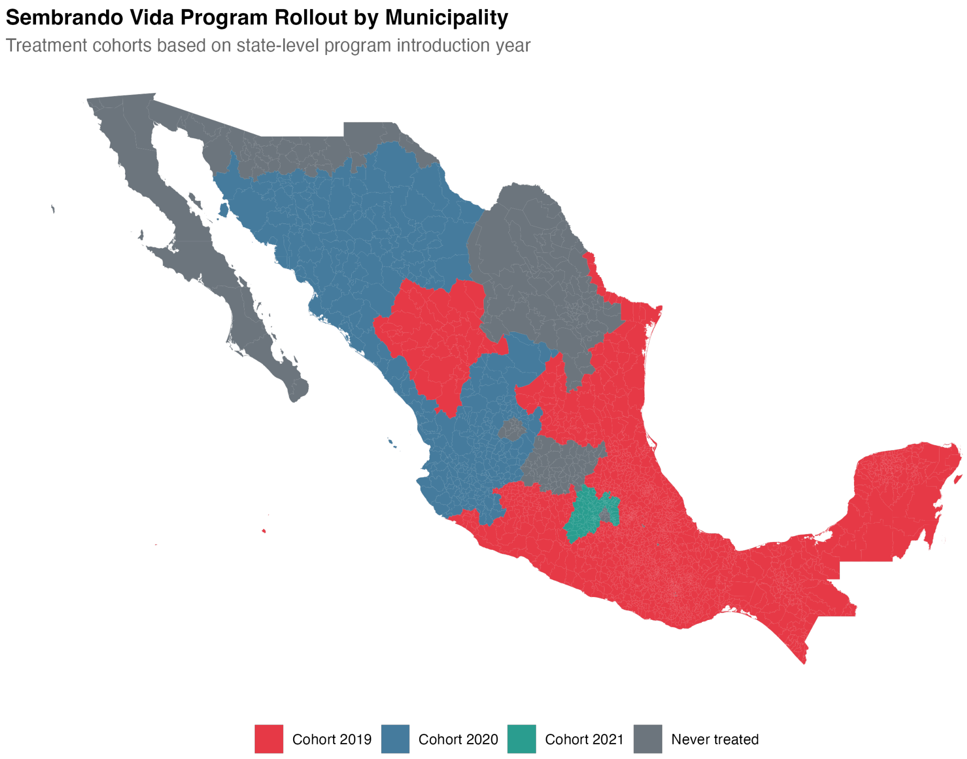
Sembrando Vida Program Rollout by Municipality
Treatment cohorts based on state-level program introduction year
Cohort 2019
Cohort 2020
Cohort 2021
Never treated

Figure 3: Sembrando Vida Program Rollout by Municipality

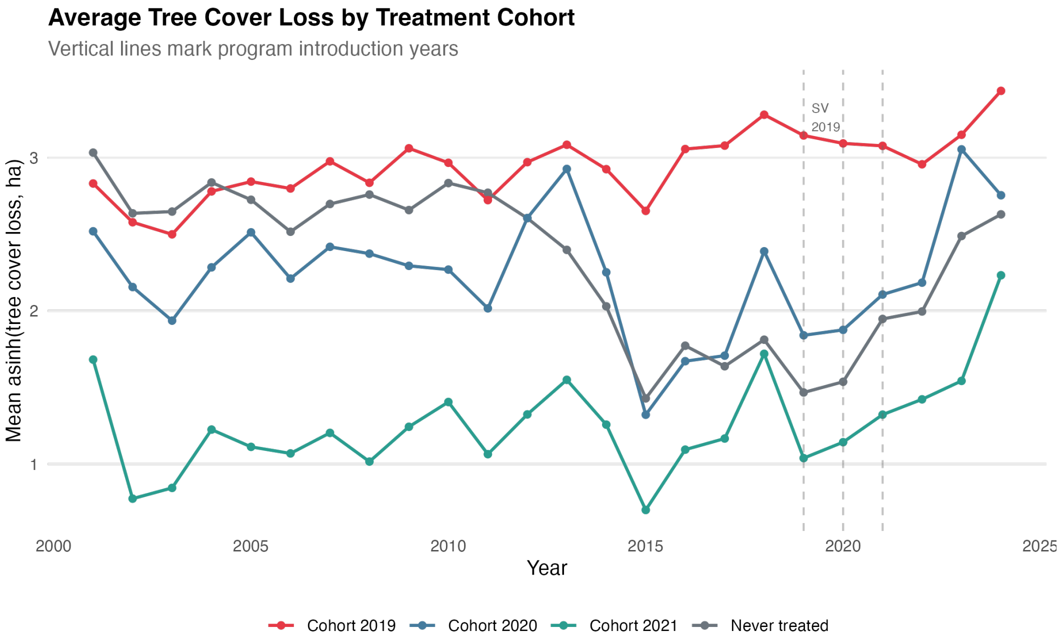

Figure 4: Average Tree Cover Loss by Treatment Cohort

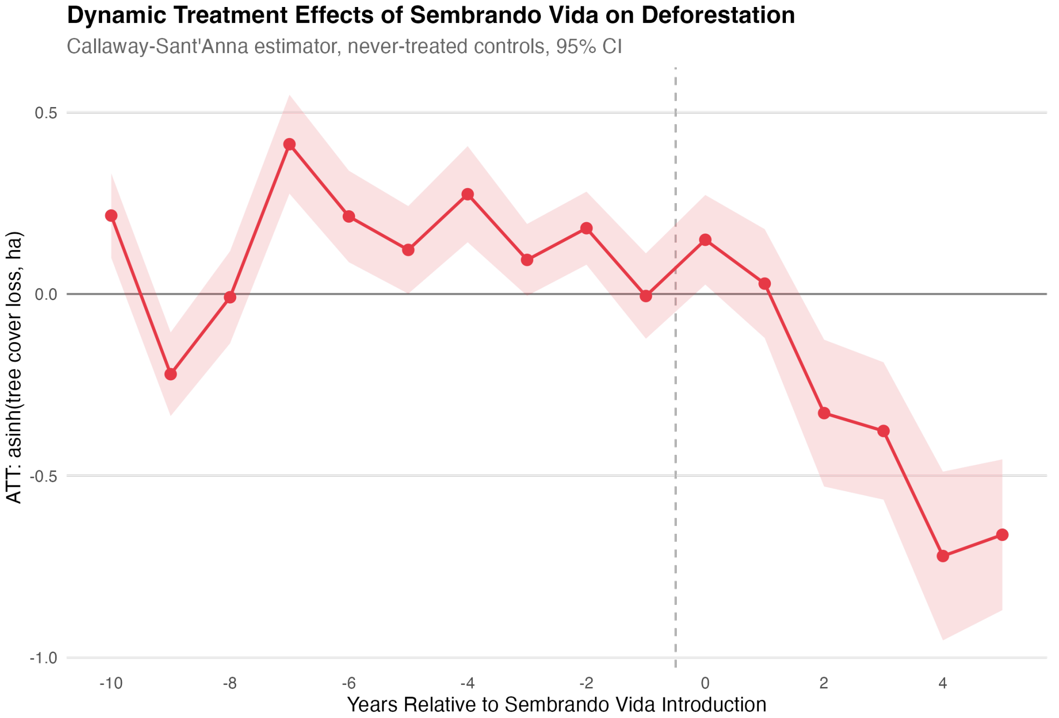

Figure 5: Dynamic Treatment Effects of Sembrando Vida on Tree Cover Loss

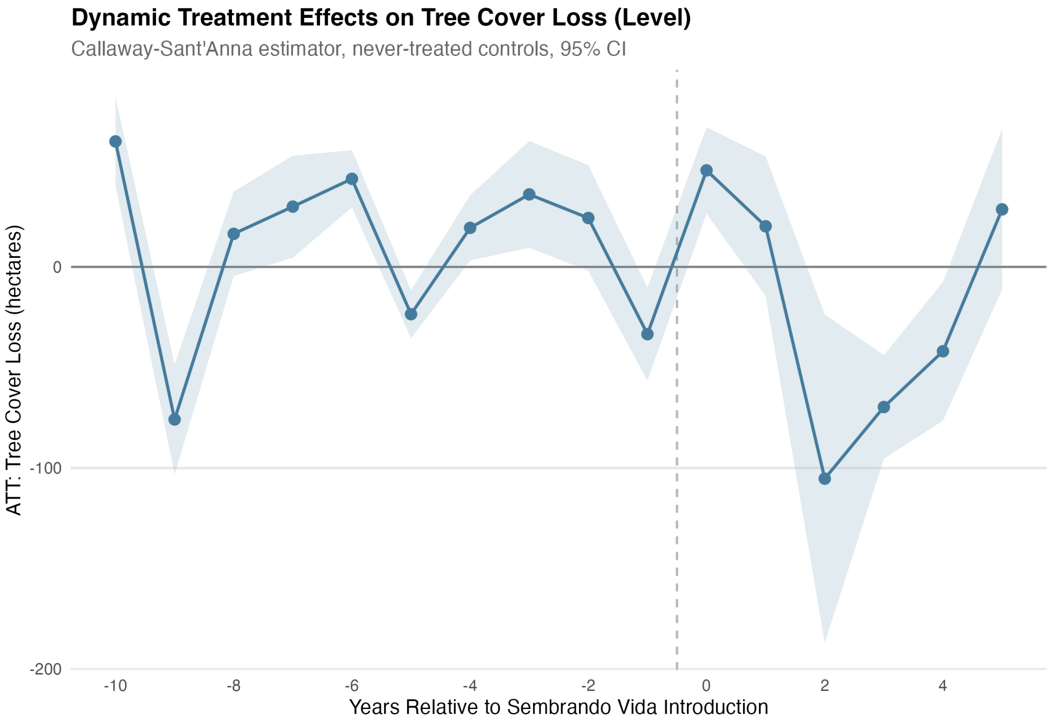

Figure 6: Event Study: Tree Cover Loss in Hectares

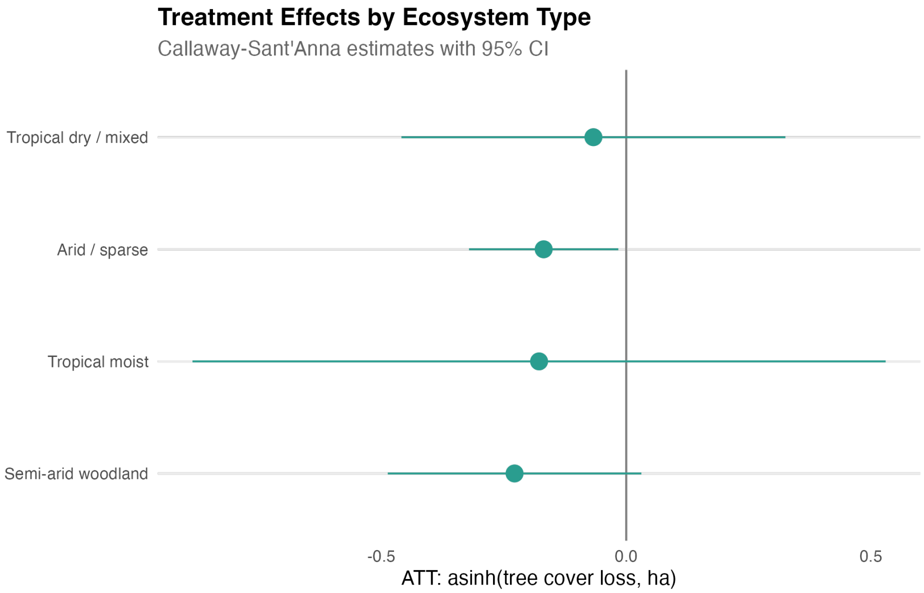

Figure 7: Treatment Effects by Ecosystem Type

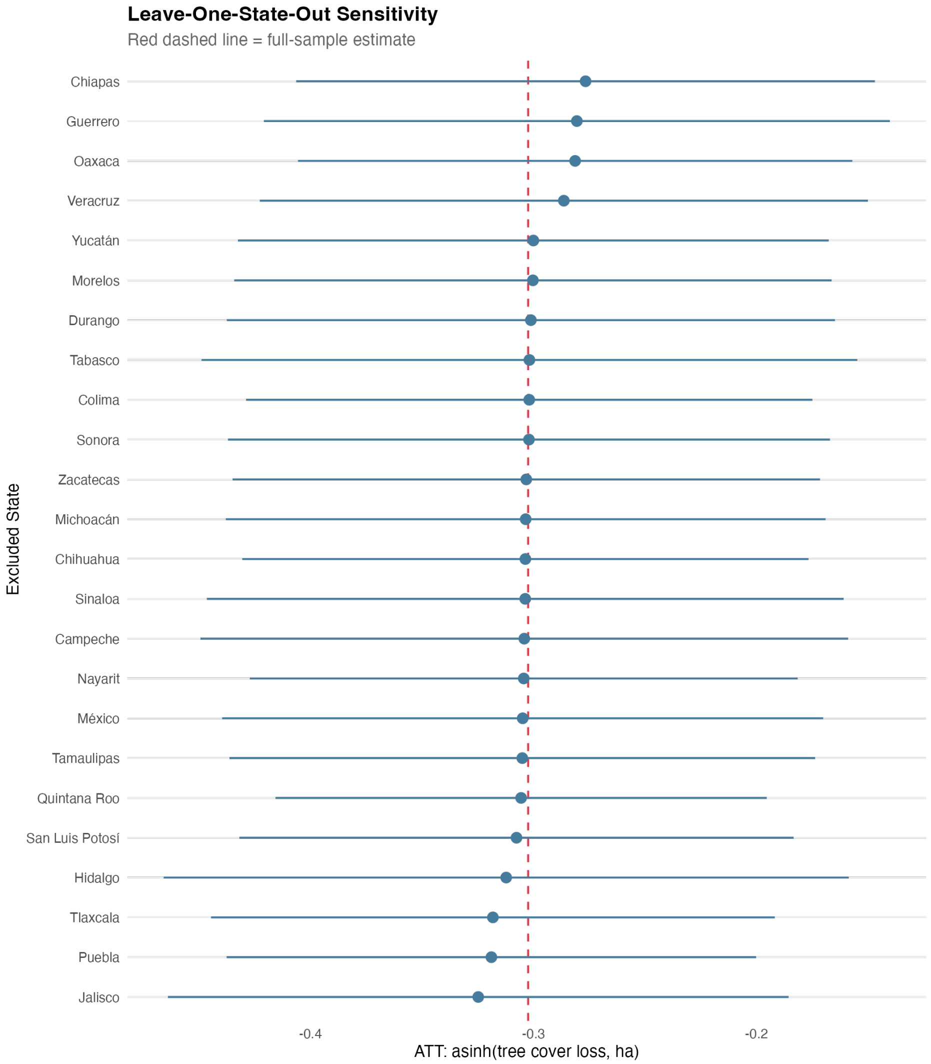

Figure 8: Leave-One-State-Out Sensitivity

Table 5: Leave-One-State-Out Sensitivity

Excluded State	ATT	SE	95% CI	N (munis)
Durango	-0.3012	(0.0695)	[-0.4374, -0.1649]	56,904 (2,371)
Guerrero	-0.2806	(0.0716)	[-0.4208, -0.1403]	55,896 (2,329)
Hidalgo	-0.3122	(0.0783)	[-0.4657, -0.1587]	55,824 (2,326)
Jalisco	-0.3248	(0.0710)	[-0.4638, -0.1857]	54,840 (2,285)
México	-0.3048	(0.0687)	[-0.4395, -0.1702]	54,888 (2,287)
Michoacán	-0.3035	(0.0686)	[-0.4379, -0.1691]	55,128 (2,297)
Morelos	-0.3003	(0.0683)	[-0.4341, -0.1664]	57,048 (2,377)
Nayarit	-0.3044	(0.0626)	[-0.4271, -0.1816]	57,360 (2,390)
Oaxaca	-0.2813	(0.0634)	[-0.4055, -0.1571]	44,256 (1,844)
Puebla	-0.3189	(0.0605)	[-0.4375, -0.2002]	52,728 (2,197)
Quintana Roo	-0.3055	(0.0562)	[-0.4156, -0.1955]	57,600 (2,400)
San Luis Potosí	-0.3076	(0.0634)	[-0.4318, -0.1834]	56,448 (2,352)
Sinaloa	-0.3037	(0.0728)	[-0.4464, -0.1610]	57,408 (2,392)
Sonora	-0.3020	(0.0688)	[-0.4369, -0.1671]	56,760 (2,365)
Tabasco	-0.3018	(0.0750)	[-0.4488, -0.1549]	57,432 (2,393)
Tamaulipas	-0.3050	(0.0670)	[-0.4362, -0.1738]	56,808 (2,367)
Tlaxcala	-0.3182	(0.0644)	[-0.4445, -0.1919]	56,400 (2,350)
Veracruz	-0.2864	(0.0695)	[-0.4227, -0.1501]	52,752 (2,198)
Yucatán	-0.3001	(0.0675)	[-0.4324, -0.1677]	55,296 (2,304)
Zacatecas	-0.3032	(0.0672)	[-0.4348, -0.1716]	56,448 (2,352)
Campeche	-0.3041	(0.0740)	[-0.4492, -0.1590]	57,576 (2,399)
Chiapas	-0.2767	(0.0662)	[-0.4063, -0.1470]	55,008 (2,292)
Chihuahua	-0.3036	(0.0647)	[-0.4305, -0.1767]	56,376 (2,349)
Colima	-0.3019	(0.0647)	[-0.4288, -0.1750]	57,600 (2,400)

Notes: Each row reports the Callaway and Sant'Anna (2021) ATT after excluding all municipalities from the named state. N reports municipality-year observations with number of municipalities in parentheses. Stability across exclusions indicates no single state drives the result.

F. Standardized Effect Sizes

Table 6: Standardized Effect Sizes for Main Outcomes

Outcome	Specification	$\hat{\beta}$	SD(X)	SD(Y)	SDE	Classification
asinh(Tree cover loss, ha)	CS-DiD, Table 2 Col. 1	−0.3024	—	2.359	−0.128	Large negative
Tree cover loss (ha)	CS-DiD, Table 2 Col. 2	−21.5	—	598.9	−0.036	Null
Loss rate (per 1000 ha)	CS-DiD, Table 2 Col. 3	−0.4757	—	3.288	−0.145	Large negative

Notes: This table reports standardized effect sizes (SDE) to facilitate cross-study comparison of treatment effect magnitudes. For binary (0/1) treatments, SDE $= \hat{\beta}/\text{SD}(Y)$ and the SD(X) column is marked "—". SD(Y) values are unconditional standard deviations computed over the full analysis sample (all municipality-years), before conditioning on fixed effects.

Research question: Does Mexico's Sembrando Vida agroforestry subsidy affect deforestation through its "available land" eligibility requirement? **Caution:** Parallel trends assumption is violated (placebo $p < 0.001$); estimates are not credible as causal effects. **Treatment:** Binary municipality-level indicator for state-level Sembrando Vida program operation, staggered across 2019–2021 cohorts. **Data:** Hansen/UMD Global Forest Change v1.12 (30m satellite), 2001–2024, approximately 2,400 Mexican municipalities. **Method:** Staggered DiD with Callaway–Sant'Anna estimator, never-treated controls, doubly-robust estimation, multiplier bootstrap inference (1,000 iterations). **Sample:** All Mexican municipalities with positive baseline forest area (treecover2000 $\geq 25\%$ canopy).

Classification thresholds: large negative (< -0.10), small negative (-0.10 to -0.05), null (-0.05 to 0.05), small positive (0.05 to 0.10), large positive (> 0.10). A reader unfamiliar with the paper should be able to interpret this table on its own.