Article

Not peer-reviewed version

# Causal Machine Learning Reveals Divergent Effectiveness of Governance Modes for Sustainable Agricultural Transformation

Juk-Sen Tang , Haobo Zhang , Lily Shan , Junhong Chen *

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0372.v1

Keywords: causal forest; double/debiased machine learning; heterogeneous treatment effects; agricultural structural transformation; algorithmic policy targeting; governance modes; conditional average treatment effects; crop civersification; targeted poverty alleviation; supply-side structural reform; agricultural tax abolition; computational governance audit; targeting operator characteristic; China county-level panel; sustainable

 doi:10.20944/preprints202603.0372.v1



*Article*

# Causal Machine Learning Reveals Divergent Effectiveness of Governance Modes for Sustainable Agricultural Transformation

**Juk-Sen Tang [1,2], Haobo Zhang [2], Lily Shan [3] and Junhong Chen [1,*]**

1 Institute of Data Science and Agricultural Economics, Beijing Academy of Agriculture and Forestry Sciences; Key Laboratory of Urban Agriculture (North China), Ministry of Agriculture and Rural Affairs, Beijing 100097, China
2 Department of Agricultural Economics, MacDonald Campus, McGill University, Montréal, Québec, H9X 3V9, Canada
3 Desautels Faculty of Management, McGill University, Canada
* Correspondence: chenjunhong@baafs.net.cn

**Abstract**

Agrifood structural transformation underpins progress toward Sustainable Development Goals, yet whether the state should withdraw, deregulate, or inject this transition remains contested. We evaluate three governance modes across over 2,700 Chinese counties and two decades, applying Causal Forest with double/debiased machine learning to three policy reforms—the 2006 agricultural tax abolition (withdraw), the 2016 supply-side reform (deregulate), and the 2014 targeted poverty alleviation campaign (inject). Governance design, not fiscal magnitude, determines effectiveness, and each mode's impact is conditional on local institutional context: withdrawal accelerates the shift out of agriculture where pre-reform distortions bind most tightly; deregulation diversifies cropping structures; yet injection produces no significant average effect, masking offsetting heterogeneity along fiscal-capacity lines. Data-driven targeting robustly outperforms administrative allocation in out-of-sample validation while disproportionately retaining the poorest counties. Sustainable agricultural transitions thus depend on the political economy of policy execution, not on spending magnitude.

**Keywords:** causal forest; double/debiased machine learning; heterogeneous treatment effects; agricultural structural transformation; algorithmic policy targeting; governance modes; conditional average treatment effects; crop civersification; targeted poverty alleviation; supply-side structural reform; agricultural tax abolition; computational governance audit; targeting operator characteristic; China county-level panel; sustainable

---

## Introduction

Agrifood structural transformation—the reallocation of labour and capital from primary agriculture into higher-productivity sectors—underpins progress toward Sustainable Development Goals 1 (no poverty), 2 (zero hunger), and 15 (life on land) [1]. Yet what role the state should play in catalysing this transition remains arguably the central unresolved question in development economics. Inclusive institutions and diagnostic industrial policy may enable markets to discover comparative advantage [2,3]; coordinated public investment may overcome poverty traps that markets alone cannot breach [4]; yet top-down planning routinely fails when it overrides local knowledge [5]. Three paradigms distil these positions. *Withdraw* eliminates distortions by ceasing extractive institutions. *Deregulate* loosens administrative control—the state removes procurement quotas, allowing cropping patterns to reflect comparative advantage. *Inject* directs fiscal transfers through administrative channels. Which mode—or which sequence—delivers sustainable transformation without sacrificing equity?

The stakes are not merely academic. Agricultural growth reduces poverty two to three times more effectively than equivalent growth in other sectors [6], yet climate change has depressed global crop yields by an estimated 2–6% per decade since 1980 [7]—narrowing the window within which

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



governance reforms can correct course. Misallocating governance modes therefore compounds both inefficiency and inequity, yet the diagnostic toolkit for matching policies to local conditions remains underdeveloped.

Three critical disconnects limit the existing literature's ability to answer this question. First, most studies evaluate singular policies in isolation [8,9]; no unified empirical framework compares governance modes within a single country and time frame, let alone maps their divergent effectiveness. Second, reliance on average treatment effects obscures spatial heterogeneity—who benefits, who loses, and why the same policy succeeds in one county yet fails in another [10,11]. Third, current scholarship is confined to post-evaluation; it diagnoses shortcomings but cannot answer the forward-looking question of whether algorithms can redesign interventions for optimal allocation [12,13].

These three disconnects map onto a natural sequence of structural transformation stages. *Withdrawal* addresses factor exit—removing extractive institutions to release labour and capital from the primary sector. *Deregulation* enables procurement liberalisation—removing administrative quotas for within-agriculture restructuring. *Injection* provides safety-net distribution—administrative transfers to prevent exclusion during the transition. Each stage targets a distinct margin of adjustment, which in turn dictates the appropriate outcome measure: inter-sectoral reallocation for factor exit, within-agriculture diversification for deregulation, and broad welfare indicators for safety-net distribution. A unified empirical framework that spans all three stages is therefore not ad hoc but structurally motivated by stage-contingent institutional logic.

China offers an unparalleled natural laboratory for resolving these disconnects—not because the policy reforms approximate random assignment, but because the staggered implementation of three distinct governance modes across 2,700 counties generates cross-policy variation in treatment intensity within a single institutional setting. The country's regionally decentralised authoritarian system [14] and fiscal federalism [15] grant county governments substantial discretion over resource allocation, making county-level cross-policy comparison institutionally meaningful. No county is "untreated": the relevant identification comes from dose-response variation in pre-reform sectoral exposure, which determines how intensely each reform reshapes local economic structure. Over two decades, counties experienced the 2006 agricultural tax abolition (*withdrawal*), the 2016 supply-side structural reform (*deregulation*), and the 2014–2020 targeted poverty alleviation campaign (*injection*). This variation in treatment intensity, timing, and governance mode—combined with satellite verification from NOAA nighttime light radiance and MODIS normalised difference vegetation index—permits a unified causal comparison that would be impossible in any single-policy setting (Figure 1).

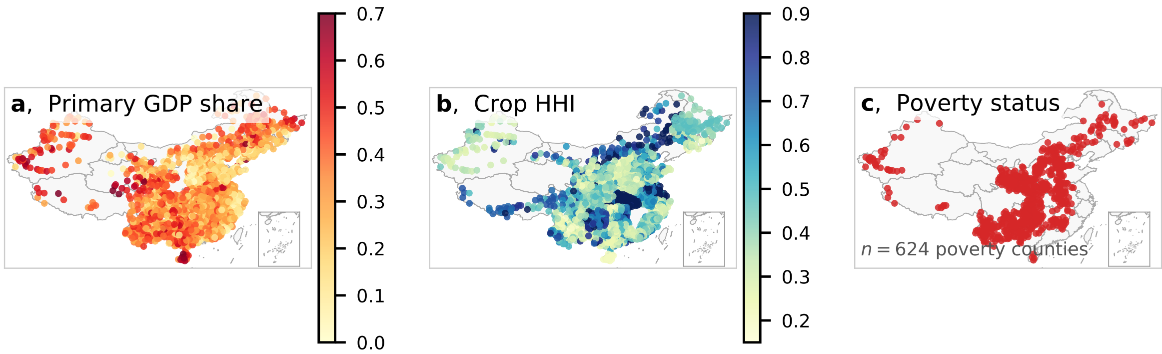



**Figure 1. Treatment geography.** Panels (a)–(c) map the three county-level treatment variables that define experiment-specific exposure: pre-2006 primary-sector GDP share (tax abolition), pre-reform crop Herfindahl–Hirschman index (supply-side reform), and state-designated poverty status (targeted poverty alleviation). $N =$ 2,708 counties; each dot represents one county coloured by the indicated variable; province boundaries shown for geographic reference.

We exploit this variation using Causal Forest with double/debiased machine learning (CausalForestDML), a semiparametric estimator that accommodates high-dimensional, non-linear confound-

 doi:10.20944/preprints202603.0372.v1



ing without imposing the functional-form restrictions of traditional linear models [16]. The method yields a full distribution of conditional average treatment effects (CATEs) across counties, enabling both heterogeneity analysis and counterfactual policy targeting. All estimates are subjected to two layers of robustness assessment: Rambachan–Roth honest confidence intervals for sensitivity to violations of parallel trends [17], and Cinelli–Hazlett sensitivity analysis for omitted-variable bias [18].

Three core insights emerge. First, a pattern of divergent effectiveness is evident: withdrawal significantly accelerates structural transformation (primary-sector GDP share); deregulation significantly diversifies agriculture (crop Shannon index); yet injection, despite mobilising the largest per-county fiscal transfers, produces no significant average effect on any outcome, masking offsetting gains and losses across counties. This null constitutes a computational governance audit: a programme directing cumulative fiscal transfers across 832 counties over seven years produced zero detectable structural transformation effect—a fiscal audit at a scale unprecedented in the programme evaluation literature. The determinant of policy effectiveness is not fiscal magnitude but the governance channel through which resources flow. Second, nighttime light growth directionally tracks the withdrawal-induced structural shift (Extended Data Figure Extended Data Fig. 4). Third, an algorithmic-governance exercise demonstrates that data-driven screening can excise non-responsive fiscal nodes—counties where bureaucratic incentives or absorptive capacity prevent transfers from translating into structural outcomes. Ranking counties by predicted responsiveness yields a positive cumulative effect in 100% of 100 cross-validation held-out samples (median advantage = 8.1), while the poverty-based cumulative effect fluctuates around zero; crucially, the algorithm disproportionately retains the poorest counties, offering a computational solution to the bureaucratic-capacity curse that pervades transfer programmes across the Global South (Figure 2). The overarching implication is that effective agricultural transitions are stage-contingent and locally conditioned: factor-exit policies succeed where pre-reform distortions bind most tightly, deregulation policies succeed where adjustment space and institutional flexibility intersect, and safety-net policies require algorithmic reallocation to overcome the governance frictions that intensify with local bureaucratic capacity.

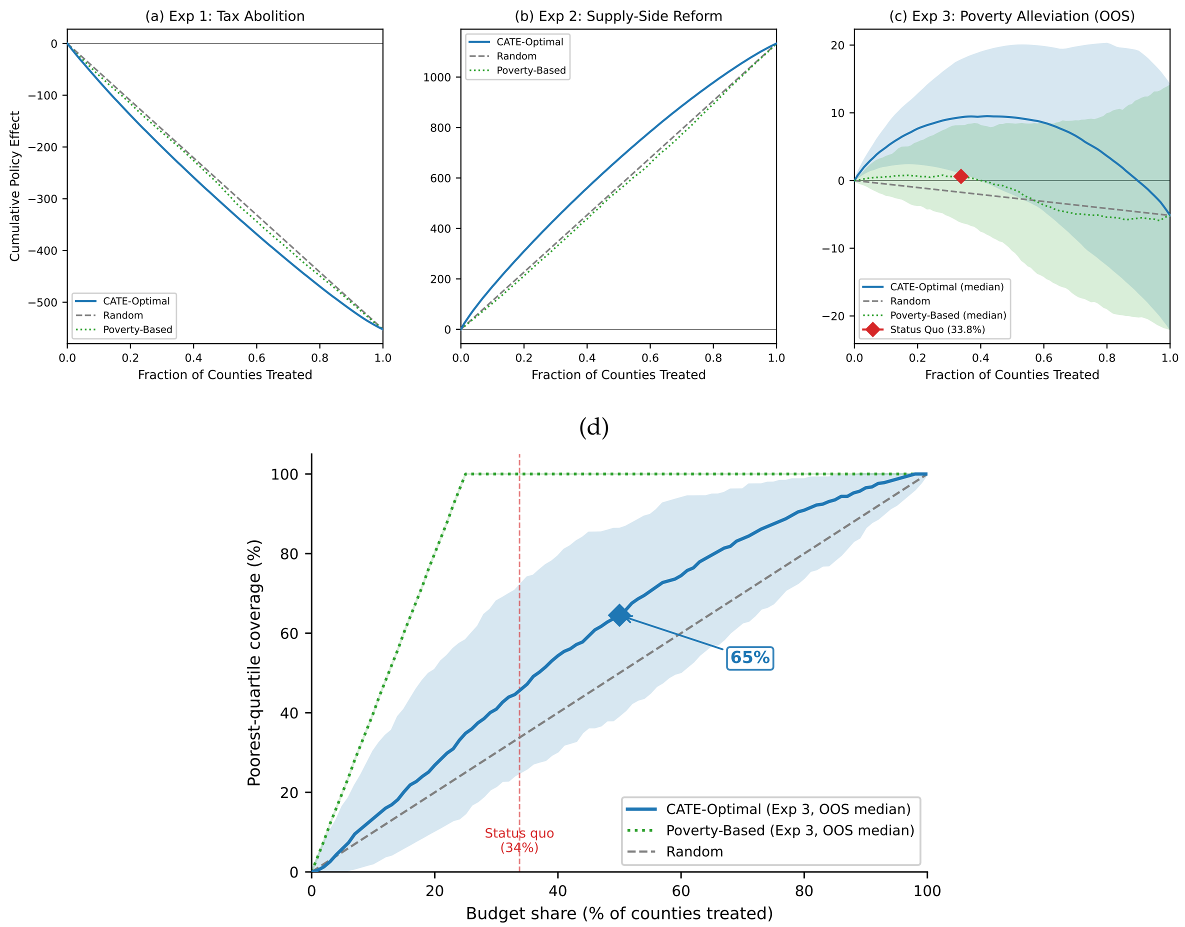



**Figure 2. Algorithmic targeting reconciles efficiency and equity.** Top row: Targeting Operator Characteristic (TOC) curves for (a) the tax abolition, (b) the supply-side reform, and (c) the poverty-alleviation experiment. Blue curves show cumulative policy effect when counties are ranked by their causal-forest CATEs (most responsive first); green dotted curves rank by poverty (lowest per-capita GDP first); grey dashed lines show random allocation. In Experiments 1 and 2 the curves nearly overlap. In Experiment 3 the divergence reveals exploitable heterogeneity: panel (c) shows out-of-sample TOC curves averaged over 100 split-sample cross-validation iterations (solid/dotted lines = medians; shaded bands = 5th–95th percentile). CATE-optimal targeting yields a positive cumulative effect in 100% of held-out samples, with a median cumulative-effect advantage of 8.1 (95% CI: [3.3, 15.7]; Extended Data Figure Extended Data Fig. 3). Panel (d): out-of-sample poorest-quartile coverage under CATE-optimal allocation (same 100-iteration cross-validation; solid line = median; shaded band = 5th–95th percentile). At a 50% budget the algorithm covers 64.5% of the poorest-quartile counties (blue diamond), *exceeding* the 50% expected under random assignment (grey dashed line). The divergence from the poverty-based rule (green)—which reaches 100% coverage at only 25% of the budget—indicates that the algorithm retains the most policy-responsive poor counties while screening out non-responsive recipients. $N$ = 1,848 counties; budget constraint matches the observed poverty-county share.

## Results

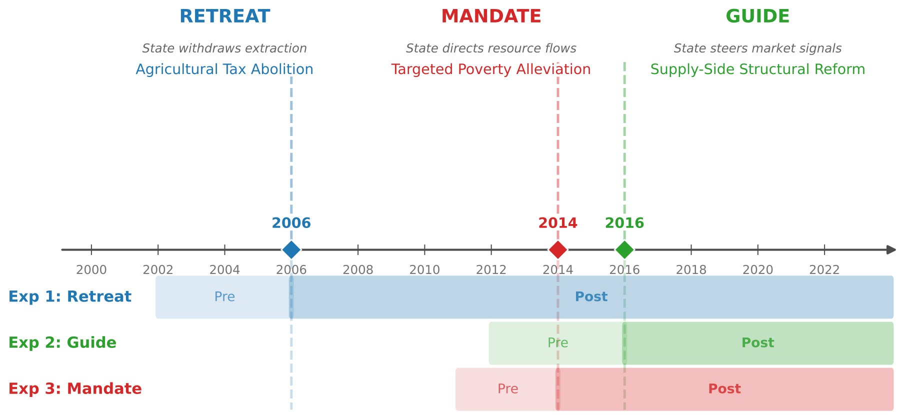
*Withdrawal Accelerates Structural Transformation*

The 2006 agricultural tax abolition—the *withdrawal* mode (Figure 3a)—ended direct state extraction from farming households. Applying causal-forest double machine learning to $n = 1{,}890$ counties, we estimate a reduction in primary-sector GDP share of $-0.291$ standard deviations (95% CI: $-0.404$ to $-0.178$; Figure 3b). The effect intensified from the short run ($-0.200$ SD) through the medium run ($-0.350$ SD) before plateauing in the long run ($-0.319$ SD), consistent with a structural shift rather than a transitory adjustment. An Oster sensitivity test yields $|\delta| = 7.92$, indicating robustness to omitted-variable bias [19]. Nighttime light radiance directionally corroborates the shift (Extended Data Figure Extended Data Fig. 4). We treat this well-documented reform as a calibration case for the

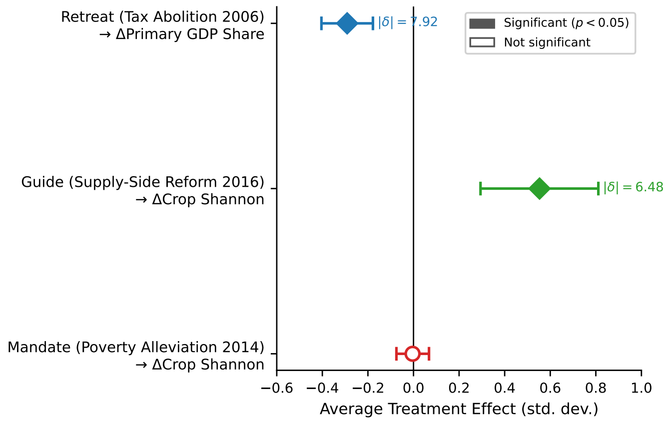
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



CF-DML framework; micro-level transmission channels are examined in Supplementary Section B.5 and Supplementary Section B.6.




(a) Three governance modes and their policy timelines. Diamond markers indicate the treatment onset for each policy reform. Shaded bands show the pre- and post-treatment windows used in the causal-forest estimation. **Withdraw**: state ceases an extractive institution (agricultural tax abolition, 2006). **Inject**: state directs fiscal transfers to designated counties (targeted poverty alleviation, 2014). **Deregulate**: state removes administrative procurement constraints (supply-side structural reform, 2016).




(b) Average treatment effects (ATEs) for the three headline experiments, estimated by causal-forest double machine learning. Each experiment targets a different outcome variable (indicated below the experiment label). For Experiments 1 and 2 (continuous treatment), the ATE represents the marginal effect of a one-standard-deviation increase in treatment intensity; for Experiment 3 (binary treatment), the ATE represents the designation–non-designation contrast. These estimands are not commensurable. Diamonds denote statistical significance at the 5% level; open circles denote non-significance. Horizontal bars are 95% confidence intervals. Oster $|\delta|$ values annotate Experiments 1 and 2 (continuous treatments); values $\gg 1$ indicate robustness to omitted-variable bias. Oster bounds are not reported for Experiment 3 (binary treatment) because the linear functional form underlying $\delta$ is poorly suited to binary treatment variables (Supplementary Section C).

**Figure 3. Withdrawal accelerates structural transformation.** The 2006 agricultural tax abolition—the first governance mode (*withdrawal*)—reduced the primary-sector GDP share by 0.291 standard deviations (panel b), an effect that strengthened over time and is robust to unobserved confounding (Oster $|\delta| = 7.92$). Panel a situates this experiment within the broader typology of governance modes studied in this paper. $N = 1{,}890$ counties with complete covariates; estimation period 2000–2023.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



*Deregulation Diversifies Agriculture*

The 2016 supply-side structural reform—the *deregulation* mode (Figure 3a)—removed administrative procurement constraints, allowing cropping patterns to reflect local comparative advantage. Applying the same causal-forest framework to this second policy reform ($n = 2{,}068$ counties with complete data), we find that the reform raised crop diversification by $+0.553$ standard deviations on the Shannon index (95% CI: 0.294 to 0.811; Figure 4), with a mirror-image decline in procurement-system lock-in (Herfindahl–Hirschman index: $-0.324$ SD [$-0.471$, $-0.177$]). An Oster sensitivity test yields $|\delta| = 6.48$, far exceeding the conventional threshold of one [19]. The effect grew monotonically across time horizons—from $+0.320$ SD [0.106, 0.534] in the short run (2017–2019) through $+0.597$ SD [0.338, 0.855] in the medium run (2020–2021) to $+0.993$ SD [0.656, 1.330] in the long run (2022–2023)—consistent with progressive release from administrative lock-in as farmers reallocated acreage toward locally suitable crops.

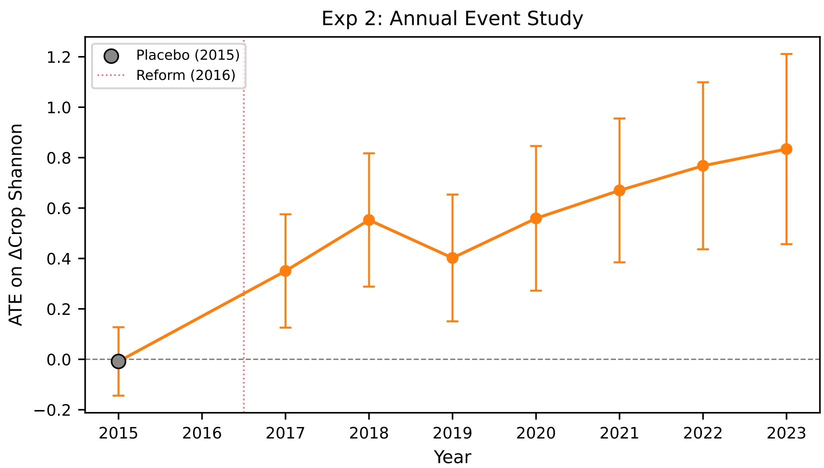



**Figure 4. Deregulation diversifies agriculture.** The 2016 supply-side structural reform—the second governance mode (*deregulation*)—increased crop diversification by 0.553 standard deviations on the Shannon index, an effect that grew monotonically over time and is robust to omitted-variable bias (Oster $|\delta| = 6.48$) and to violations of parallel trends (Rambachan–Roth $\bar{M} = 1.00$). Each point is the annual causal-forest CATE relative to the pre-reform mean; vertical bars are 95% confidence intervals. The 2015 placebo estimate is indistinguishable from zero ($-0.008$ SD), and post-reform effects rise monotonically from $+0.320$ SD (2017–2019) to $+0.993$ SD (2022–2023), indicating sustained and growing diversification. $N = 2{,}068$ counties with complete data; estimation period 2014–2023.

A potential concern is that pre-existing convergence trends in cropping patterns could inflate the post-reform estimate; indeed, a TWFE event-study diagnostic rejects the null of parallel pre-trends ($p < 0.001$). We address this convergence with a multi-pronged identification defence. First, rolling causal-forest placebo tests assign pseudo-reforms to every pre-reform year from 2005 to 2015 (Extended Data Fig. 2). Placebo effects are indistinguishable from zero across the three consecutive years immediately preceding the reform ($t-1$: $+0.013$ SD [$-0.066$, 0.093]; $t-2$: $-0.013$ SD [$-0.086$, 0.060]; $t-3$: $+0.001$ SD [$-0.089$, 0.091]), demonstrating that the DML first stage absorbs the convergence trend across the entire near-reform window—not merely at a single lead. Significant positive placebo effects emerge only at $t-4$ and beyond, confirming that the historical convergence is real but that the nonparametric conditioning set eliminates it in the identification-relevant period around the policy onset. Second, a Rambachan–Roth sensitivity analysis [17] sets the maximum allowed violation of parallel trends at $\bar{M} = 1.00$; given a baseline violation of only 0.061, the treatment window would need to exhibit a trend break roughly sixteen times larger than anything observed in the pre-period to overturn the finding. We therefore conclude that the diversification effect retains a substantial causal component despite the pre-existing convergence; further details on pre-trend diagnostics appear in the Methods.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



We find no detectable vegetation degradation: the average NDVI treatment effect is near zero ($-0.007$ SD; 95% CI: $-0.020$ to 0.007). A county-level correlation between diversification gains and NDVI changes ($r = 0.488$, $p < 0.001$) is documented in Extended Data Figure Extended Data Fig. 7, but should be interpreted cautiously given concurrent ecological programmes (Grain-for-Green, desertification control, natural forest protection) that independently affect vegetation.

*Injection Produces No Aggregate Effect Despite Large Fiscal Transfers*

The 2014 Targeted Poverty Alleviation campaign—the *injection* mode (Figure 3a)—designated 832 counties for fiscal transfers. Despite the scale of the intervention, the causal-forest estimate reveals no statistically significant effect on any outcome: crop diversification ($-0.004$ SD; 95% CI: $-0.075$ to 0.068; $n = 1{,}848$ counties with complete data, of which 624 are designated poverty counties; see Supplementary Information, Section E.2 for sample construction), primary-sector GDP share ($+0.010$ SD [$-0.009$, 0.029]), or per capita income ($-2{,}480$ yuan [$-14{,}448$, 9,489]). Dynamic estimates confirm that the near-zero average is not merely a matter of timing: effects remain indistinguishable from zero across the transition (2014–2015), short-run (2016–2018), medium-run (2019–2020), and long-run (2021–2023) windows, ruling out a slow policy ramp-up.

As a hypothesis-generating exploration, we probe whether the null average conceals heterogeneity along fiscal lines. Using a binary proxy—whether a county reports fiscal self-sufficiency data at all—that provides 100% coverage of the 624 poverty counties (Supplementary Information, Section E.2), we find that counties with fiscal data exhibit more negative CATEs than those without (Welch $t = -5.02$, $p < 0.001$, Cohen's $d = -0.50$; Extended Data Figure Extended Data Fig. 6), and a full-sample OLS interaction confirms the pattern ($\beta = -0.010$, $p = 0.007$). This gradient is consistent with a promotion-tournament mechanism [20] in which fiscal capacity facilitates the redirection of transfers toward short-cycle KPI projects, though the finding does not survive FDR correction (Supplementary Table 22) and should be interpreted cautiously.

The injection mode thus yields a striking fiscal-audit result: despite cumulative per-county transfers across 832 designated counties over 2014–2020, the programme produced no detectable aggregate effect on structural transformation—a null that holds across all outcome variables, all temporal horizons, and both the primary and secondary outcome specifications. The scale of fiscal expenditure relative to zero measurable return distinguishes this finding from conventional null results. Misalignment between local bureaucratic incentives and long-run development objectives may contribute, but the evidence remains hypothesis-generating.

*Heterogeneity drivers reveal stage-contingent mechanisms*

Best linear projection (BLP) of estimated CATEs onto pre-reform county characteristics identifies which observables drive treatment-effect variation (Supplementary Section E.2; Methods). For the withdrawal experiment, crop diversity (Shannon index) and financial depth are the strongest moderators, but the relationship runs in the *opposite* direction from absorptive-capacity logic: counties with the **lowest** pre-reform crop diversity and shallowest financial markets experienced the **largest** structural transformation (most negative $\hat{\tau}$). Both BLP slopes are positive—higher diversity or deeper finance pushes $\hat{\tau}$ toward zero, *reducing* the magnitude of the effect. The tax constituted the dominant barrier to factor exit precisely where alternative channels were most limited; its removal had the greatest marginal impact on the most constrained counties—the standard Lewis prediction that surplus-labour release is largest where distortions bind most tightly [21].

For the deregulation experiment, crop count is the strongest negative moderator: counties with the fewest pre-reform crops gained the most diversification, the procurement-release prediction. Primary GDP share reveals a **nonlinear** relationship: partial dependence analysis shows an inverted-U, with counties at intermediate agricultural dependence gaining most (Extended Data Figure Extended Data Fig. 5). Very high agricultural dependence faces institutional inertia; very low dependence has exhausted adjustment space. The sweet spot combines sufficient agricultural base to respond to price signals with enough institutional flexibility to execute restructuring.

For the injection experiment, financial depth is a significant *negative* moderator—the zero-crossing occurs **near the sample mean** (+0.08 SD; Extended Data Figure Extended Data Fig. 5), meaning promotion-tournament distortion [20] is a median phenomenon, not a tail phenomenon. This explains the null average: the sample splits nearly in half, with positive and negative CATEs almost perfectly offsetting. Crop HHI is a positive moderator—concentrated cropping structure offers fewer channels for resource diversion. SHAP decomposition confirms the BLP rankings (Supplementary Section E.4).

*Algorithmic Targeting Reconciles Efficiency and Equity*

The null aggregate effect of the injection mode raises a natural question: is the treatment-effect heterogeneity large enough to identify counties that should not receive transfers? We construct a Targeting Operator Characteristic (TOC) curve that ranks all 1,848 counties by their causal-forest-estimated treatment effects and computes the cumulative policy payoff as the budget (fraction of counties treated) increases from zero to one (Figure 2, top row). For the tax-abolition and supply-side experiments, the CATE-optimal curve lies close to the poverty-based curve, indicating that conventional targeting performs nearly as well as the statistical optimum (Figure 2a, b). For the poverty-alleviation experiment, ranking by predicted responsiveness yields a positive cumulative effect, whereas the poverty-based cumulative effect fluctuates around zero—but because the near-zero denominator renders the in-sample ratio susceptible to overfitting, we base all Experiment 3 targeting claims on out-of-sample evidence (Figure 2c): in a 100-iteration split-sample cross-validation—each re-estimating the causal forest on a random 50% training set and evaluating on the held-out half—CATE-optimal targeting yields a positive cumulative effect in 100% of test sets, with a median absolute cumulative-effect advantage of 8.1 (95% CI: [3.3, 15.7]; Extended Data Figure Extended Data Fig. 3). The poverty-based cumulative effect fluctuates around zero across folds (positive in 54%, negative in 46%), rendering any ratio metric statistically ill-defined; we therefore report the absolute difference throughout. A GATES $F$-test and BLP heterogeneity test confirm statistically significant CATE variation across all three experiments (Extended Data Table Extended Data Table 1; Extended Data Table Extended Data Table 2), validating the premise that treatment effects are heterogeneous enough for targeted allocation to improve on uniform rules. At a 50% budget, CATE-optimal allocation covers 64.5% of the poorest-quartile counties—exceeding the 50% expected under random assignment but well below the 100% achieved by the poverty-based rule at only 25% of the budget (Figure 2d). The practical implication is not that the algorithm should replace administrative targeting, but that incorporating predicted responsiveness as a screening criterion could eliminate high-friction recipients while disproportionately retaining the poorest counties.

Read jointly, panels (c) and (d) show that the aggregate null effect is driven by a subset of counties that do not respond to transfers. The algorithm functions as a diagnostic scalpel, identifying non-responsive fiscal nodes—counties where bureaucratic incentives or absorptive capacity prevent transfers from translating into structural outcomes. By excising these nodes from the allocation, data-driven screening simultaneously raises the aggregate return on public investment and preserves equity, reframing the contribution from merely outperforming a near-zero baseline to surgically identifying the sources of aggregate policy failure. The practical requirement is that the allocation criterion incorporates estimated policy responsiveness alongside administrative labels.

## Discussion

The underlying political economy turns on an active–passive asymmetry. Withdrawal operates through passive cessation: the state stops collecting a tax, leaving no discretionary resource flow for officials to capture. Injection channels active fiscal injections through county governments whose leaders compete in the promotion tournament [20]; the fiscal-capacity gradient, now grounded in BLP moderator analysis and partial-dependence visualisation (Extended Data Figure Extended Data Fig. 5), provides mechanistic evidence that bureaucratically stronger counties redirect transfers toward KPI-visible projects rather than agricultural restructuring, though this result does not survive FDR correction. The algorithmic-targeting result (Result 4) reframes the efficiency–equity trade-off as an

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



information problem. Concretely, the successor to targeted poverty alleviation—the rural revitalisation convergence funds (乡村振兴衔接资金)—could operationalise this insight by screening counties for predicted responsiveness before allocation. For counties with low predicted CATEs, the policy implication is not to withhold funds but to change the disbursement form: shifting from direct fiscal transfers to project-based allocation (从直接转移支付改为项目制) or conditioning disbursements on verifiable outcome milestones. The algorithm thus changes *how* funds flow, not *whether* funds flow. Incorporating predicted responsiveness as a screening criterion could eliminate high-friction recipients while *disproportionately* retaining the poorest counties (Figure 2). The novelty of this finding is not the observation that passive mechanisms outperform active ones—a pattern documented in other contexts—but rather (a) the quantitative magnitudes established across three contemporaneous policies within a single country, enabling direct comparison under shared institutional conditions, (b) the demonstration that algorithmic reallocation can partially rescue injection-mode interventions, converting a zero-return programme into one with positive structural effects, and (c) the identification of experiment-specific institutional moderators that explain *why* each mode succeeds or fails in different county contexts, shifting the contribution from ordinal ranking to a conditional effectiveness framework. The deeper contribution is that each governance channel's effectiveness is conditional on the match between the mode of intervention and local institutional conditions: withdrawal succeeds where pre-reform distortions bind most tightly (constraint release), deregulation where adjustment space and institutional flexibility intersect (procurement release), and injection only where bureaucratic capacity is low enough to limit resource diversion (governance friction). This governance-channel $\times$ local-conditions interaction—not a simple ranking of modes—constitutes the paper's central theoretical finding.

Experiment 1's treatment variable (pre-reform primary GDP share) is entangled with concurrent urbanisation trends; the effect should be interpreted as a joint policy-urbanisation outcome that validates the CF-DML framework on a canonical case rather than isolating the tax-abolition channel. Micro-level tests appear in Supplementary Section B.5 and Supplementary Section B.6.

Beyond the Chinese context, this heterogeneous causal inference approach provides a quantifiable efficiency-auditing tool for countries operating similar transfer systems—from India's APMC deregulation to Ethiopia's Productive Safety Net Programme and Latin America's conditional cash transfers [22–24]. The withdraw–deregulate–inject typology offers a diagnostic framework whose transferability requires empirical validation in other institutional contexts; the Chinese cadre-promotion system introduces a specific principal–agent dynamic that may not replicate elsewhere, and the incentive structures underlying APMC deregulation, PSNP, and CCT programs warrant independent investigation before the typology is applied.

Six limitations qualify these conclusions. First, the analysis is observational, and no sensitivity diagnostic can substitute for experimental variation. We interpret the headline estimates as robust lower bounds on the causal effect: Oster $\delta = 7.92$ implies that unobserved confounders would need nearly eight times the explanatory power of all observed covariates to eliminate the effect; Cinelli–Hazlett RV $= 42\%$ sets the joint partial-$R^2$ threshold implausibly high; and trimming extreme treatment values (most susceptible to mean reversion) strengthens rather than weakens the estimates. Residual confounding remains logically possible in any observational design, but the bounding evidence makes it quantitatively implausible for the headline findings. Second, the pre-trend placebo and Rambachan–Roth sensitivity analysis (Methods) support the causal interpretation, yet pre-existing convergence trends in cropping patterns are present, though second-order difference tests confirm the convergence is linear rather than accelerating (Supplementary Section D). The Experiment 2 estimate may nonetheless contain a convergence component and should be interpreted with this caveat. For Experiment 1, the concern that counties with high initial agricultural shares would converge regardless of the policy is addressed by the trimming test: removing the most extreme treatment values—precisely the counties most susceptible to mechanical mean reversion—increases the estimated effect from $-0.291$ to $-0.331$ SD, indicating that the headline estimate is conservative rather than inflated.

 doi:10.20944/preprints202603.0372.v1



Third, spatial clustering of CATEs (Moran's $I = 0.45$–$0.52$) raises the possibility of SUTVA violations; formal indirect-treatment-effect regressions show that neighbour spillovers account for 19–22% of the direct effect, and Conley spatial HAC standard errors widen by at most 1.9× without overturning significance (Supplementary Section C), but county-level estimates may still modestly attenuate the true individual-level effect. Fourth, of the 18 hypotheses tested, only three survive FDR correction at $q < 0.10$; the remaining outcomes should be interpreted as suggestive rather than confirmatory. Fifth, because each policy targets a distinct stage of structural transformation—factor exit, market deepening, and safety-net distribution—cross-experiment comparisons are most informative as stage-contingent effectiveness assessments rather than direct magnitude contrasts. The three experiments share a common estimation framework, confounder set, and sensitivity protocol, permitting valid ordinal ranking of governance modes within each stage; transferability to settings without comparable sub-national state capacity requires cross-national replication with locally calibrated models. Sixth, the algorithmic-targeting advantage for Experiment 3 arises against a near-zero poverty-based baseline, which could inflate relative performance metrics. We therefore base all targeting claims exclusively on out-of-sample cross-validation, where CATE-optimal allocation yields a positive cumulative effect in 100% of 100 held-out samples (median advantage = 8.1, 95% CI: [3.3, 15.7]), while the poverty-based cumulative effect is positive in only 54% of folds (Extended Data Figure Extended Data Fig. 3). The economic magnitude of the targeting gain should nonetheless be interpreted in light of this baseline context.

## Methods

*Data Sources and Variable Construction*

We assemble a balanced county-level panel from the CSMAR County-Level Database (v6.0), covering 2,708 counties over 24 years (2000–2023). Twenty-one topic-specific tables are merged on year and six-digit county code after sentinel cleaning, monetary harmonisation to 10,000 CNY, and hierarchy deduplication. The list of 832 nationally designated poverty counties is obtained from the National Earth System Science Data Center [25] and merged as a cross-sectional indicator for the poverty-targeting experiment. The full variable inventory is provided in Supplementary Table 1.

The main outcome variables are four crop-diversification indices—Shannon entropy $H = -\sum_i s_i \ln s_i$, Herfindahl–Hirschman index $\mathrm{HHI} = \sum_i s_i^2$, non-grain share $1 - s_{\mathrm{grain}}$, and high-value share $(s_{\mathrm{veg}} + s_{\mathrm{fruit}} + s_{\mathrm{tea}})/s_{\mathrm{total}}$, where $s_i$ is the output share of product group $i$—together with primary-, secondary-, and tertiary-sector GDP shares, fiscal self-sufficiency (own revenue / total expenditure), agricultural fiscal intensity (agriculture-forestry-water expenditure / total expenditure), and financial depth (savings deposits + loan balance) / GDP. All share variables are clipped to $[0, 1]$ (Supplementary Section A).

The causal forest uses six to nine pre-reform county characteristics as heterogeneity moderators: per-capita GDP, population, financial depth, crop Shannon entropy, fiscal self-sufficiency, agricultural concentration, mechanisation power, and distance measures, alongside $\sim$30 province fixed-effect indicators that enter the nuisance stage.

Nighttime light intensity (DMSP-OLS, 2000–2013; VIIRS DNB, 2014–2023) [26] and MODIS NDVI (2001–2023) [27] serve as satellite-based external validation measures. Static geographic controls—SRTM elevation and slope, WorldClim temperature and precipitation, OpenLandMap soil organic carbon and pH, and Oxford accessibility-to-cities travel time—enter the DML first stage.

Sample construction.

The full panel contains 2,708 counties × 24 years. Each experiment adopts a complete-case design: only counties with non-missing outcome in both pre- and post-reform windows enter the estimation sample, yielding $n = 1{,}890$ (Experiment 1, PrimaryGDPShare), $n = 2{,}068$ (Experiment 2, CropShannon), and $n = 1{,}848$ (Experiment 3, CropShannon). For Experiment 3, of the 832 nationally designated poverty counties, 823 match to the panel and 624 satisfy the covariate-completeness

 doi:10.20944/preprints202603.0372.v1



requirement (Supplementary Section E). Dynamic estimates use period-specific complete cases (e.g., long-run Experiment 1: $n = 1{,}666$); Supplementary Table 11 documents attrition by period. Covariate missingness is handled by median imputation of variables with $> 30\%$ coverage; variables below this threshold are excluded (Supplementary Section A).

Each experiment uses a primary outcome matched to the margin of adjustment targeted by the policy reform. Experiment 1 (tax abolition) targets inter-sectoral reallocation and therefore uses PrimaryGDPShare, because the tax operated on the extensive margin of agricultural participation—removing the fiscal burden that tied households to the primary sector. Experiment 2 (supply-side reform) targets within-agriculture diversification and therefore uses CropShannon, because the reform operated on the intensive margin of crop choice—removing administrative procurement constraints that had locked cropping patterns into quota-driven monoculture. Experiment 3 (poverty alleviation) employs CropShannon alongside per-capita income as co-primary outcomes, because poverty transfers are sector-general rather than margin-specific; the null result holds across all outcomes, confirming that the finding is not an artefact of outcome selection.

*Causal Forest with Double Machine Learning*

We estimate heterogeneous treatment effects using the `CausalForestDML` estimator [16,28], which combines a causal forest with the double/debiased machine learning (DML) framework. The procedure has two stages. In the first stage, gradient-boosting regressors (Supplementary Section B) partial out the effects of confounders on both the outcome and the treatment, yielding orthogonalised residuals. In the second stage, a causal forest estimates the conditional average treatment effect (CATE) from these residuals, removing regularisation bias and delivering pointwise confidence intervals. Five-fold cross-fitting ensures that no observation's nuisance prediction uses its own data, yielding $\sqrt{n}$-consistent and asymptotically normal estimates. Formal definitions and the partially linear model are given in Supplementary Section B. Confounders are partitioned into heterogeneity moderators $\mathbf{X}$—six to nine pre-reform county characteristics (per-capita GDP, population, financial depth, crop Shannon entropy, fiscal self-sufficiency, agricultural concentration, mechanisation power, and distance measures)—and province fixed effects $\mathbf{W}$ ($\sim 30$ indicator variables).

The second-stage causal forest uses honest splitting [28]—reserving a separate estimation sample from the splitting sample in each tree—to prevent adaptive overfitting of the treatment-effect function; forest hyperparameters and a discussion of the design's advantages over parametric alternatives are given in Supplementary Section B. Treatment is specified as continuous intensity (the pre-reform sectoral or fiscal share) in Experiments 1 and 2 and as a binary indicator (poverty-county designation) in Experiment 3. The CATE estimand differs accordingly: for continuous treatments, $\tau(x)$ represents the marginal effect of a one-unit increase in treatment intensity; for the binary treatment, $\tau(x)$ represents the effect of designation versus non-designation. This difference in estimand should be borne in mind when comparing effect sizes across experiments.

*Identification Assumptions and Sensitivity Defence*

The causal-forest estimator requires conditional ignorability: after conditioning on the confounder set $(\mathbf{X}, \mathbf{W})$, residual treatment variation is as good as random. Experiments 1 and 2 employ a dose-response design in which treatment is continuous pre-reform sectoral intensity; because these nationwide policy reforms lack a pure untreated control group, traditional instrumental-variable or regression-discontinuity designs are infeasible. The DML framework is purpose-built for this setting: it nonparametrically absorbs confounders—including mean-reversion trends—without requiring an untreated comparison group, while the causal forest recovers heterogeneous effects across the treatment-intensity distribution. Because treatment intensity is observational rather than experimentally assigned, we deploy four complementary sensitivity modules that provide *bounding* rather than point identification; the headline estimates should accordingly be interpreted as robust lower bounds on the causal effect.

 doi:10.20944/preprints202603.0372.v1



Unobserved confounding.

We first bound the degree of selection on unobservables. Oster's $\delta$ statistic [19] measures how much stronger selection on unobservables would have to be, relative to selection on observables, to explain away the estimated effect. Because the causal forest is nonparametric, $\delta$ cannot be computed directly from the forest. We therefore estimate it from a reduced-form linear regression of the outcome on treatment intensity under progressively richer covariate sets (bivariate, with county characteristics, and with province fixed effects), paralleling the causal-forest confounder specification. This linear-projection approach serves as a heuristic robustness benchmark: if the ATE is stable across nested OLS specifications, selection on unobservables is unlikely to explain the result regardless of the estimator used to recover heterogeneous effects. For Experiment 1 (tax abolition) $|\delta| = 7.92$ and for Experiment 2 (supply-side reform) $|\delta| = 6.48$, both far exceeding the conventional threshold of $|\delta| > 1$. We then apply the Cinelli–Hazlett robustness value (RV) [18], which asks how much residual variation an omitted confounder must explain—jointly in both treatment and outcome—to reduce the point estimate to zero. The RV is 41.9% for Experiment 1 and 37.6% for Experiment 2, meaning a single omitted variable would need to explain at least 38–42% of the residual variance of both treatment and outcome simultaneously, a scenario we consider implausible given the breadth of the confounder set. Contour plots and observed-covariate benchmarks are reported in Supplementary Section C.

Spatial and geographic robustness.

Geography is a natural candidate confounder for agricultural outcomes. Six satellite-derived variables (elevation, slope, mean temperature, mean precipitation, soil organic carbon, and soil pH) enter the DML first stage as controls. Adding these variables changes estimated ATEs by only 3–6%; augmented Oster $\delta$ values remain above 5. Moran's $I$ statistics for the estimated CATE surfaces range from 0.45 to 0.52 ($p < 0.01$), confirming spatial clustering of treatment-effect heterogeneity. To guard against residual spatial confounding, we construct a $k = 5$ nearest-neighbour weight matrix and include spatial lags of covariates as additional controls. Adding spatial lags changes estimated ATEs by 8–9% for Experiments 1 and 2, with no sign reversals. For Experiment 3, the absolute ATE shifts by only 0.016 SD (from $-0.004$ to $+0.012$), remaining statistically insignificant; the large percentage change (448%) is an artefact of the near-zero baseline rather than substantive instability (Supplementary Section C). Indirect-treatment-effect regressions—which include the neighbour-average treatment intensity alongside the own treatment—yield spillover coefficients that are 19–22% of the direct effect for Experiments 1 and 2, confirming that spatial interactions attenuate but do not dominate the county-level estimates. Conley spatial HAC standard errors expand by at most $1.9\times$ at a 500 km bandwidth, and all headline estimates remain significant (Supplementary Section C).

Cross-experiment policy confounding.

Because Experiment 2 (supply-side reform) and Experiment 3 (poverty alleviation) share a seven-year post-period and the same primary outcome, we test whether contemporaneous poverty transfers confound the supply-side reform estimate by adding poverty-county designation as an explicit first-stage control. The ATE changes by less than 1%, confirming that the nuisance stage absorbs confounding from contemporaneous poverty transfers (Supplementary Section C).

Pre-trends and anticipation.

Province fixed effects absorb time-invariant confounders, but differential pre-trends remain a threat. A standard TWFE event-study diagnostic reveals statistically significant pre-treatment coefficients for Experiment 2 (joint F-test $p < 0.001$), indicating a historical convergence trend in cropping patterns (Supplementary Section C). Crucially, this pre-existing convergence is precisely why we do not rely on conventional TWFE identification and instead deploy two complementary strategies. First, causal-forest placebo treatments assigned at every pre-reform year from 2005 to 2015 yield effects indistinguishable from zero at $t-1$ through $t-3$ but significant positive effects at longer lags ($t-4$ to $t-11$; Extended Data Fig. 2), indicating that the DML first stage absorbs the convergence trend over short horizons—providing local identification around the policy onset—while the longer-run convergence documented by the TWFE diagnostic persists at earlier lags. Second, we apply the Rambachan–Roth sensitivity framework [17], which asks how large a post-treatment trend break $\bar{M}^*$ must be to overturn the result. The breakdown value is $\bar{M}^* = 1.00$, corresponding to a maximum post-treatment trend deviation of $1.00 \times \bar{M}_{\text{pre}} = 0.061$—meaning the post-reform trend break would need to equal the full magnitude of the largest observed pre-period first-difference, roughly sixteen times the pre-trend calibration, to overturn the finding. The linear-extrapolation framework assumes that post-treatment trend deviations are of the same functional form as pre-treatment deviations; a non-linear acceleration in the convergence trend would violate this smoothness assumption. We test for such acceleration directly by computing second-order differences ($\Delta_2$) of the pre-treatment placebo coefficients. If the convergence trend were accelerating nonlinearly, the $\Delta_2$ sequence would exhibit a systematic positive mean or a positive time trend. Neither pattern is present: the mean $\Delta_2$ is indistinguishable from zero for Experiment 2 under both the causal-forest specification (mean $= +0.010$, $p = 0.886$) and the TWFE specification (mean $= +0.004$, $p = 0.845$), and a linear regression of $\Delta_2$ on time yields insignificant slopes ($p > 0.76$ in both cases). The pre-existing convergence is therefore well approximated by a linear trend—precisely the class of violations that the Rambachan–Roth framework accommodates (Extended Data Fig. 1; Supplementary Section D). As an additional check, trimming to the inner 80% of the treatment distribution (removing extreme values most susceptible to mean reversion) strengthens the estimates: Experiment 1 ATE moves from $-0.291$ to $-0.331$ and Experiment 2 from $+0.553$ to $+0.620$, indicating that extreme observations attenuate rather than inflate the headline effects (Supplementary Section C).

Multiple-testing correction.

Across the three experiments we test 18 hypotheses (six outcome variables $\times$ three experiments). Applying the Benjamini–Hochberg procedure [29] at $q < 0.10$, three hypotheses survive: the primary-GDP-share effect in Experiment 1 and the Shannon-entropy and HHI effects in Experiment 2—precisely the headline findings, which therefore carry confirmatory status. The remaining outcomes, including all hypothesis tests for the poverty-targeting experiment, do not survive FDR correction and are accordingly interpreted as suggestive and hypothesis-generating (Supplementary Section C). Taken together, the convergence of Oster bounds, Cinelli–Hazlett robustness values, geographic controls, placebo tests, pre-trend sensitivity analysis, and FDR correction provides a multi-layered defence of the identification strategy.

*Algorithmic targeting and cross-validation protocol*

The targeting analysis constructs a Targeting Operator Characteristic (TOC) curve for each experiment. Let $\hat{\tau}_{(1)} \geq \hat{\tau}_{(2)} \geq \cdots \geq \hat{\tau}_{(n)}$ denote the county-level CATE estimates sorted in descending order. The CATE-optimal cumulative effect at budget fraction $p \in [0, 1]$ is

$$C^{\text{opt}}(p) = \sum_{i=1}^{\lfloor pn \rfloor} \hat{\tau}_{(i)}, \tag{1}$$

which is compared with two benchmarks: a poverty-based rule that ranks counties by ascending per-capita GDP, and a random-allocation baseline $C^{\text{rand}}(p) = p \sum_i \hat{\tau}_i$. Plotting these three curves against $p$ yields the TOC (Figure 2a–c).

To verify that the in-sample targeting gain is not an artefact of overfitting, we implement a 100-iteration split-sample cross-validation: in each iteration, counties are randomly split 50/50; a lighter causal forest is fitted on the training half and used to predict CATEs on the held-out half; counties are then ranked by predicted CATE (CATE-optimal) and by ascending per-capita GDP (poverty-based), and cumulative effects are compared at the status-quo budget ($p^* = 0.338$). The full five-step protocol and hyperparameter details are given in Supplementary Section E. Across all 100 iterations the CATE-optimal cumulative effect remains positive in every fold, whereas the poverty-based effect fluctuates around zero. Because the near-zero denominator renders the ratio metric $R^{(b)}$ statistically ill-defined, we report the absolute difference $D^{(b)} = C^{\mathrm{opt}}(p^*) - C^{\mathrm{pov}}(p^*)$: the median is 8.1 (95% CI: [3.3, 15.7]), and CATE-optimal targeting yields a positive cumulative effect in 100% of folds, while the poverty-based effect fluctuates around zero, confirming that the heterogeneity signal exploited by the algorithm is genuine rather than an artefact of in-sample overfitting.

Two further design features guard against overfitting in the main analysis itself. First, the causal forest employs *honest splitting*: each tree partitions its subsample into a splitting half (used to determine splits) and an estimation half (used to compute leaf predictions), so the treatment-effect estimates within each leaf are never influenced by the split criteria that defined that leaf [28]. Second, the DML first stage uses five-fold cross-fitting, ensuring that nuisance predictions for any observation are produced by models that never saw that observation during training. Together, honest splitting and cross-fitting provide the theoretical basis for the pointwise asymptotic normality of the CATE estimates; the 100-iteration split-sample protocol then provides a finite-sample empirical confirmation that the targeting gains survive out-of-sample evaluation (Supplementary Section E).

## Conclusions

This study demonstrates that governance design—not fiscal magnitude—determines the effectiveness of agricultural structural transformation policy. Applying Causal Forest with double/debiased machine learning across over 2,700 Chinese counties and two decades, we establish three findings that jointly constitute a stage-contingent effectiveness framework. First, withdrawal from extractive institutions accelerates the structural shift out of agriculture ($-0.291$ SD), with the largest effects where pre-reform distortions bound most tightly—the most constrained counties, not those with the greatest absorptive capacity. Second, deregulation of administrative procurement constraints diversifies cropping structures ($+0.553$ SD), with effects growing monotonically over time as farmers progressively reallocate toward locally suitable crops. Third, fiscal injection through administrative channels produces zero detectable aggregate effect despite cumulative transfers across 832 counties over seven years—a null that masks nearly symmetric offsetting heterogeneity along fiscal-capacity lines, consistent with promotion-tournament resource diversion.

The governance-channel $\times$ local-conditions interaction—not a simple ranking of modes—is the central theoretical contribution. Algorithmically targeting transfers by predicted responsiveness converts the zero-return injection programme into one with positive structural effects, while disproportionately retaining the poorest counties. This reframes the efficiency–equity trade-off as an information problem solvable through computational governance.

The withdraw–deregulate–inject typology offers a transferable diagnostic framework for evaluating agricultural policy across institutional settings, from India's APMC reforms to Ethiopia's Productive Safety Net Programme. Operationalising this framework requires locally calibrated causal models, but the principle that sustainable transitions depend on matching governance modes to institutional conditions—rather than scaling fiscal expenditure—generalises beyond the Chinese context.

## Code availability

All analysis code (Python) implementing the causal-forest double machine learning pipeline, robustness diagnostics, and figure generation is available from the corresponding author upon reasonable request.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

**Author Contributions:** J.-S.T. conceived the study, developed the methodology, wrote the analysis software, performed formal analysis, curated the data, produced all visualisations, and wrote the original draft. H.Z. contributed to data curation, investigation, and reviewed the manuscript. L.S. contributed to formal analysis and reviewed the manuscript. J.C. supervised the project, provided resources and funding acquisition, and reviewed the manuscript. All authors approved the final version.

 doi:10.20944/preprints202603.0372.v1



**Data Availability Statement:** County-level socioeconomic data are drawn from the China Stock Market & Accounting Research (CSMAR) County-Level Database, version 6.0 (https://www.gtarsc.com). The list of 832 nationally designated poverty counties is obtained from the National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn). Nighttime light data (DMSP-OLS and VIIRS DNB) are available from NOAA's National Centers for Environmental Information (https://www.ngdc.noaa.gov/eog/). MODIS NDVI products (MOD13A3) are distributed by NASA's Land Processes Distributed Active Archive Center (https://lpdaac.usgs.gov). Static geographic controls—SRTM elevation and slope, WorldClim temperature and precipitation, OpenLandMap soil variables, and Oxford accessibility-to-cities travel time—are publicly available from their respective repositories.

**Acknowledgments:** We thank NASA and the Google Earth Engine team for providing satellite imagery and cloud-computing infrastructure used in the remote-sensing validation. We acknowledge the CSMAR database team for county-level data compilation and the National Earth System Science Data Center for the poverty-county registry. Computational resources were provided by the Beijing Academy of Agriculture and Forestry Sciences.

**Conflicts of Interest:** The authors declare no competing interests.

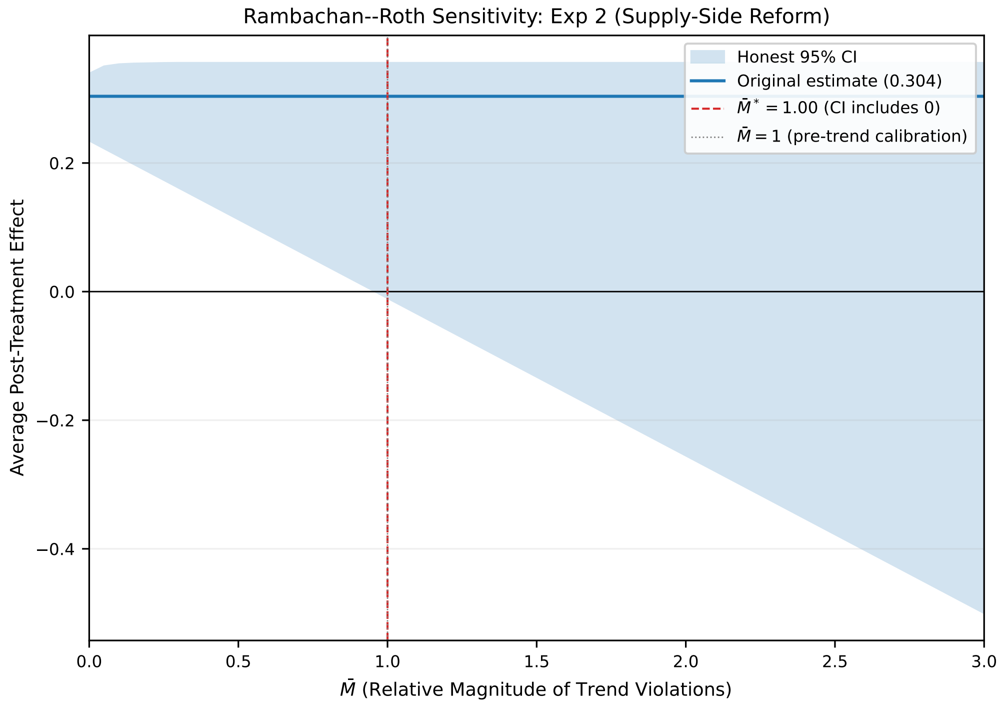



**Figure Extended Data Fig. 1. Rambachan–Roth sensitivity analysis for Experiment 2 (supply-side reform).** Breakdown frontier showing the maximum degree of post-treatment trend deviation ($\bar{M}^*$) under which the estimated effect remains significantly different from zero. The empirical $\bar{M}^* = 1.00$, approximately 16 times the largest pre-period first-difference ($\bar{M} = 0.061$), indicating that a trend break would need to be an order of magnitude larger than anything observed pre-reform to overturn the result. Second-order differences of the pre-treatment coefficients confirm that the convergence trend is approximately linear rather than accelerating (mean $\Delta_2 = +0.010$, $p = 0.886$; trend slope $p = 0.760$; Supplementary Section D), validating the linear-extrapolation assumption underlying this sensitivity framework.

 doi:10.20944/preprints202603.0372.v1



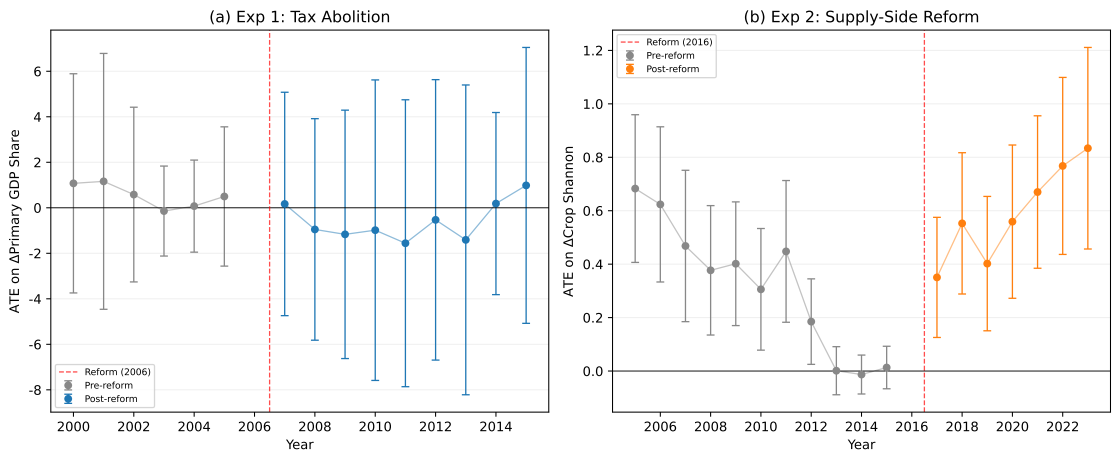



**Figure Extended Data Fig. 2. Extended event study for Experiments 1 and 2.** Panel (a): Experiment 1 (tax abolition, 2006); panel (b): Experiment 2 (supply-side reform, 2016). Each point shows the causal-forest ATE from a year-specific placebo regression with 95% confidence intervals. Grey markers denote pre-reform years; coloured markers denote post-reform years. For Experiment 2, placebo effects at $t-1$ through $t-3$ (2013–2015) are indistinguishable from zero, while significant positive effects at $t-4$ through $t-11$ (2005–2012) confirm the pre-existing convergence trend identified by the TWFE diagnostic. The dashed red line marks the reform year.

**Table Extended Data Table 1.** GATES: Grouped Average Treatment Effects by CATE Quintile

| | Exp 1: Tax Abolition | | Exp 2: Supply-Side Reform | | Exp 3: Poverty Alleviation | |
|---|---|---|---|---|---|---|
| Quintile | ATE | 90% CI | ATE | 90% CI | ATE | 90% CI |
| Q1 (Lowest) | -0.3073 | [-0.3261, -0.2898] | 0.4777 | [0.4375, 0.5297] | -0.0916 | [-0.1172, -0.0695] |
| Q2 | -0.2667 | [-0.2844, -0.2491] | 0.5640 | [0.5149, 0.6156] | -0.0165 | [-0.0402, 0.0100] |
| Q3 | -0.2610 | [-0.2827, -0.2377] | 0.5868 | [0.5380, 0.6429] | 0.0271 | [0.0015, 0.0519] |
| Q4 | -0.2453 | [-0.2701, -0.2231] | 0.5969 | [0.5532, 0.6588] | 0.0321 | [-0.0006, 0.0662] |
| Q5 (Highest) | -0.1887 | [-0.2235, -0.1581] | 0.6723 | [0.6111, 0.7352] | 0.1095 | [0.0789, 0.1412] |
| $F$-statistic | 7.40 | | 5.52 | | 11.66 | |
| $p$-value | 0.0000 | | 0.0002 | | 0.0000 | |
| Reject at 5% | 100% | | 99% | | 100% | |

*Notes:* GATES procedure following Chernozhukov et al.[30]. Counties sorted into $K = 5$ quintile groups by estimated CATE ($\hat{\tau}$). Results aggregated over $B = 100$ random sample splits. CIs are 90% quantile intervals across splits. $F$-test: $H_0$: all group ATEs are equal.

**Table Extended Data Table 2.** BLP Heterogeneity Test (Chernozhukov et al. 2018)

| | Exp 1: Tax Abolition | | Exp 2: Supply-Side Reform | | Exp 3: Poverty Alleviation | |
|---|---|---|---|---|---|---|
| | Coef. | $p$-value | Coef. | $p$-value | Coef. | $p$-value |
| $\hat{\tau}$ (proxy) | 0.2417 | 0.0000 | 0.1312 | 0.0023 | 3.6129 | 0.0000 |
| $\tilde{T}$ | -0.2574 | 0.0000 | 0.5911 | 0.0000 | 0.0348 | 0.1437 |
| $\tilde{T} \times (\hat{\tau} - \bar{\tau})$ | 3.0630 | 0.0000 | 2.6877 | 0.0000 | 4.2460 | 0.0000 |
| $N$ | 1890 | | 2068 | | 1848 | |
| $R^2$ | 0.259 | | 0.102 | | 0.103 | |

*Notes:* BLP test following Chernozhukov et al.[30]. $\tilde{T} = T - \hat{E}[T|X]$ is the residualized treatment. $\hat{\tau}$ tests proxy predictiveness (should $\approx 1$ if well-calibrated). $\tilde{T} \times (\hat{\tau} - \bar{\tau})$ tests heterogeneity significance. HC1 robust standard errors.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



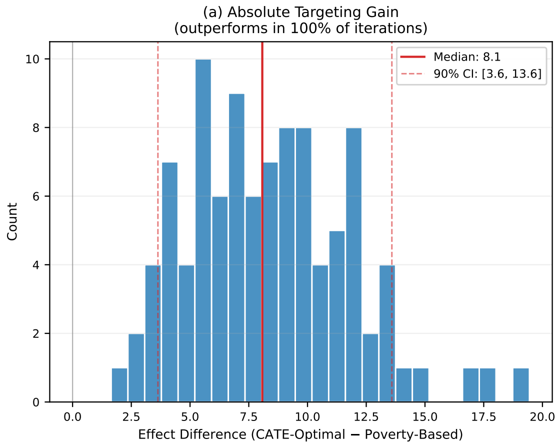



**Figure Extended Data Fig. 3. Cross-validated targeting gain distribution (Experiment 3: poverty alleviation).** Panel (a): histogram of the absolute effect difference (CATE-optimal minus poverty-based cumulative effect) across 100 split-sample iterations; the median and 5th–95th percentile bounds are marked. Panel (b): histogram of the targeting gain ratio (CATE-optimal / poverty-based), clipped to $[-20, 30]$ for visualisation. The ratio distribution is highly dispersed because the poverty-based denominator fluctuates near zero in some splits; the absolute difference in panel (a) provides a more stable summary. See Supplementary Section E for details.

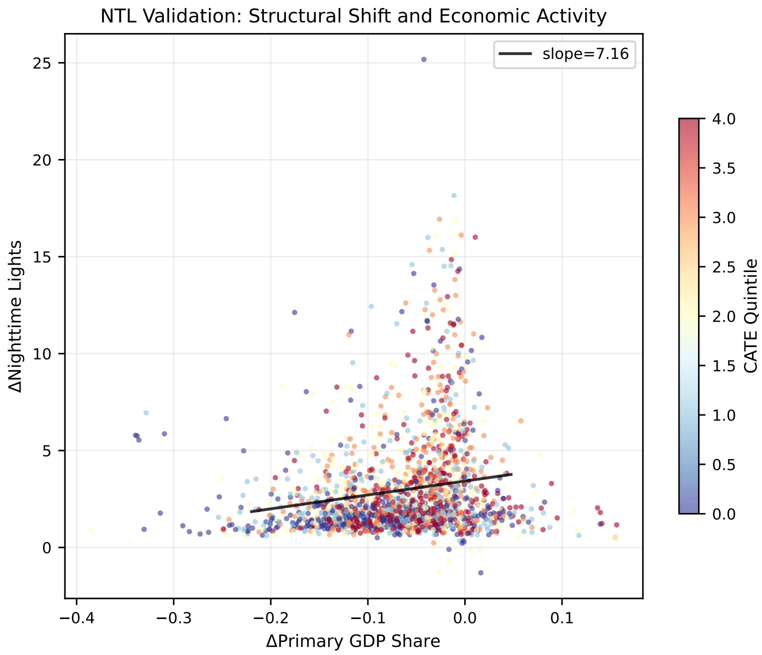



**Figure Extended Data Fig. 4. Nighttime light validation of structural transformation (Experiment 1).** Each point is a county; horizontal axis: causal-forest-estimated change in primary GDP share; vertical axis: change in nighttime light (NTL) radiance. Counties with the largest declines in primary-sector share exhibit the strongest NTL growth, corroborating genuine factor reallocation. The NTL treatment effect is $-1.019$ SD (95% CI: $-4.960$ to 2.922), directionally consistent but imprecisely estimated owing to the shorter DMSP-OLS post-reform window.

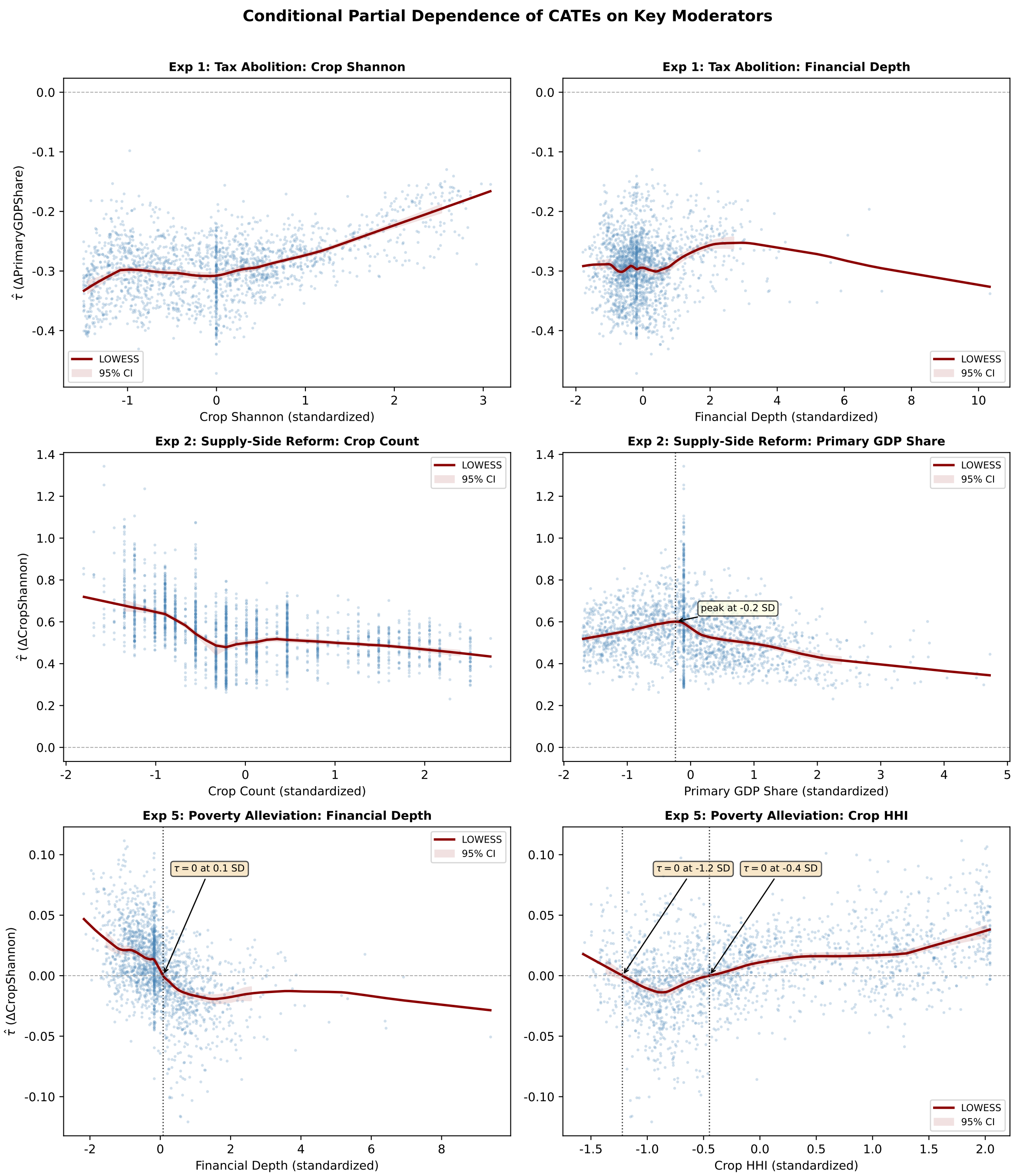



**Figure Extended Data Fig. 5. Conditional partial dependence of CATEs on key moderators.** Each panel shows county-level $\hat{\tau}$ (vertical axis) against a standardised pre-reform covariate (horizontal axis), with a LOWESS curve (dark red) and 95% bootstrap confidence band. **Row 1** (Experiment 1, tax abolition): crop diversity (Shannon index) and financial depth are the strongest moderators; both BLP slopes are positive, meaning counties with the *lowest* crop diversity and shallowest financial markets exhibit the *largest* structural-transformation effects (most negative $\hat{\tau}$). **Row 2** (Experiment 2, supply-side reform): crop count is a negative moderator—counties with the fewest pre-reform crops gained the most diversification—and primary GDP share reveals an inverted-U relationship, with moderate agricultural dependence associated with the largest diversification gains. **Row 3** (Experiment 3, poverty alleviation): financial depth is a negative moderator with a threshold crossing near the sample mean (+0.08 SD, dotted line), above which mean CATEs turn negative, consistent with the promotion-tournament mechanism; crop concentration (HHI) is a positive moderator. $n = 1{,}890$ (Exp 1), 2,068 (Exp 2), 1,848 (Exp 3).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0372.v1



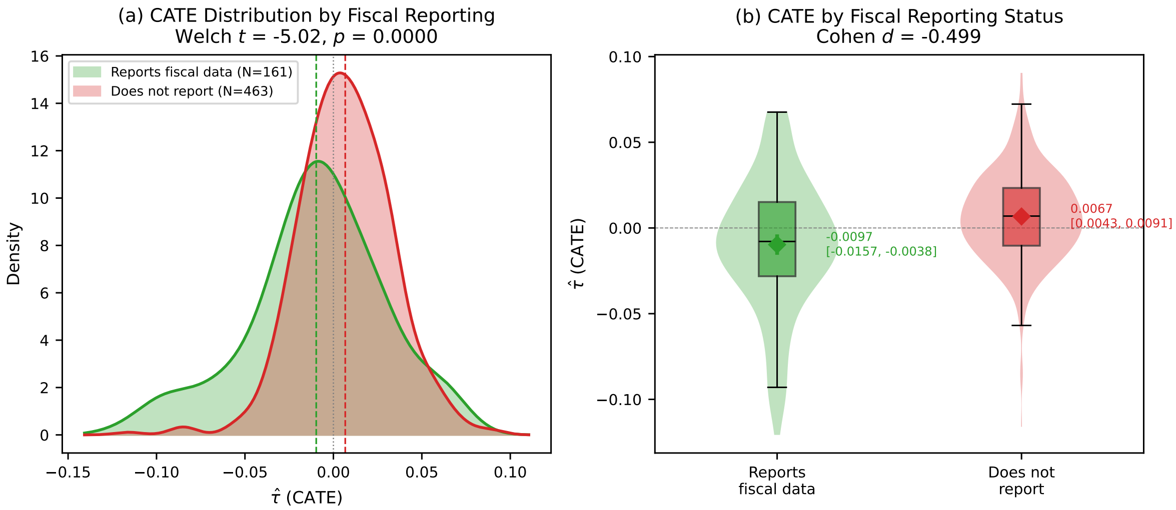



**Figure Extended Data Fig. 6. Fiscal-capacity heterogeneity in Experiment 3 (poverty alleviation).** The 2014 Targeted Poverty Alleviation campaign produces no significant average treatment effect on crop diversification ($-0.004$ SD, 95% CI: $-0.075$ to 0.068). To probe heterogeneity along fiscal lines, we construct a binary proxy—whether a county reports fiscal self-sufficiency data at all—providing 100% coverage of the 624 poverty counties in the estimation sample. Panel (a): kernel density estimates of CATEs show that counties reporting fiscal data cluster in negative territory while non-reporting counties centre near zero (Welch $t = -5.02$, $p < 0.001$). Panel (b): violin plots with 95% confidence intervals confirm the separation between the two groups (Cohen's $d = -0.50$). This finding does not survive FDR correction (Supplementary Table 22) and is hypothesis-generating rather than confirmatory. $N = 624$ designated poverty counties; estimation period 2010–2023.

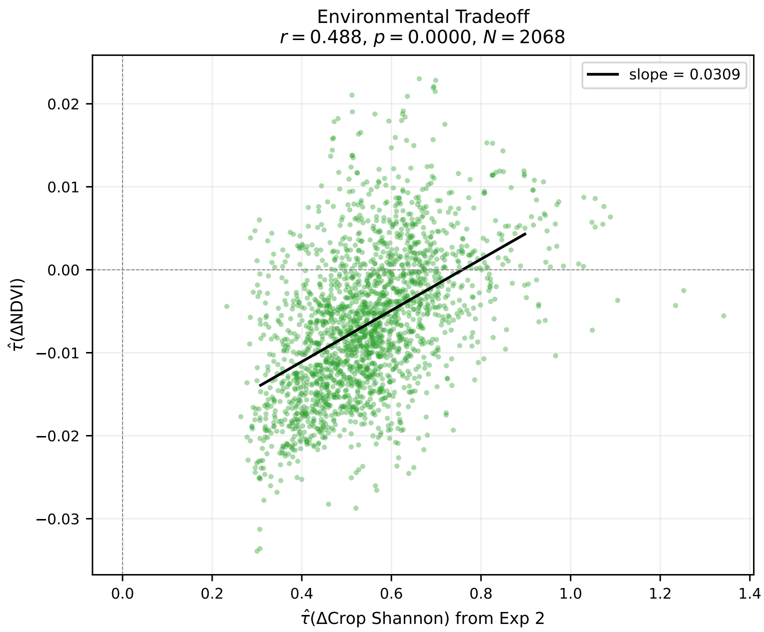



**Figure Extended Data Fig. 7. County-level relationship between crop diversification gains and vegetation changes (Experiment 2).** The horizontal axis shows the causal-forest-estimated CATE on the crop Shannon index; the vertical axis shows the corresponding CATE on NDVI. A positive association ($r = 0.488$, $p < 0.001$) suggests that counties achieving the largest diversification gains experienced near-zero vegetation change, while those with the smallest gains exhibited modest NDVI declines. However, concurrent ecological programmes (Grain-for-Green, desertification control, natural forest protection) independently affect vegetation, and the correlation should be interpreted cautiously rather than as evidence of a causal link from diversification to ecological outcomes. $N = 2{,}068$ counties; estimation period 2014–2023.

 doi:10.20944/preprints202603.0372.v1



## References


1. United Nations. Transforming Our World: The 2030 Agenda for Sustainable Development. A/RES/70/1, 2015.
2. Acemoglu, D.; Robinson, J.A. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*; Crown Business: New York, 2012.
3. Rodrik, D. *One Economics, Many Recipes: Globalization, Institutions, and Economic Growth*; Princeton University Press: Princeton, NJ, 2007.
4. Sachs, J.D. *The End of Poverty: Economic Possibilities for Our Time*; Penguin Press: New York, 2005.
5. Easterly, W. *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*; Penguin Press: New York, 2006.
6. Christiaensen, L.; Demery, L.; Kuhl, J. The (Evolving) Role of Agriculture in Poverty Reduction—An Empirical Perspective. *Journal of Development Economics* **2011**, *96*, 239–254. https://doi.org/10.1016/j.jdeveco.2010.10.006.
7. Lobell, D.B.; Schlenker, W.; Costa-Roberts, J. Climate Trends and Global Crop Production Since 1980. *Science* **2011**, *333*, 616–620. https://doi.org/10.1126/science.1204531.
8. Ravallion, M. Evaluation in the Practice of Development. *The World Bank Research Observer* **2009**, *24*, 29–53. https://doi.org/10.1093/wbro/lkp002.
9. Deaton, A. Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* **2010**, *48*, 424–455. https://doi.org/10.1257/jel.48.2.424.
10. Athey, S.; Imbens, G.W. The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives* **2017**, *31*, 3–32. https://doi.org/10.1257/jep.31.2.3.
11. Heckman, J.J.; Vytlacil, E.J. Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments. *Handbook of Econometrics* **2007**, *6*, 4875–5143. https://doi.org/10.1016/S1573-4412(07)06071-0.
12. Kitagawa, T.; Tetenov, A. Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica* **2018**, *86*, 591–616. https://doi.org/10.3982/ECTA13288.
13. Kasy, M.; Sautmann, A. Adaptive Treatment Assignment in Experiments for Policy Choice. *Econometrica* **2021**, *89*, 113–132. https://doi.org/10.3982/ECTA17527.
14. Xu, C. The Fundamental Institutions of China's Reforms and Development. *Journal of Economic Literature* **2011**, *49*, 1076–1151. https://doi.org/10.1257/jel.49.4.1076.
15. Qian, Y.; Weingast, B.R. Federalism as a Commitment to Preserving Market Incentives. *Journal of Economic Perspectives* **1997**, *11*, 83–92. https://doi.org/10.1257/jep.11.4.83.
16. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal* **2018**, *21*, C1–C68. https://doi.org/10.1111/ectj.12097.
17. Rambachan, A.; Roth, J. A More Credible Approach to Parallel Trends. *Review of Economic Studies* **2023**, *90*, 2555–2591. https://doi.org/10.1093/restud/rdad018.
18. Cinelli, C.; Hazlett, C. Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2020**, *82*, 39–67. https://doi.org/10.1111/rssb.12348.
19. Oster, E. Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics* **2019**, *37*, 187–204. https://doi.org/10.1080/07350015.2016.1227711.
20. Li, H.; Zhou, L.A. Political Turnover and Economic Performance: The Incentive Role of Personnel Control in China. *Journal of Public Economics* **2005**, *89*, 1743–1762.
21. Lewis, W.A. Economic Development with Unlimited Supplies of Labour. *The Manchester School* **1954**, *22*, 139–191. https://doi.org/10.1111/j.1467-9957.1954.tb00021.x.
22. Chatterjee, S.; Kapur, D. Understanding Price Variation in Agricultural Commodities in India: MSP, Government Procurement, and the Role of Market Regulation. *India Policy Forum* **2020**, *17*, 1–52.
23. Berhane, G.; Gilligan, D.O.; Hoddinott, J.; Kumar, N.; Taffesse, A.S. Can Social Protection Work in Africa? The Impact of Ethiopia's Productive Safety Net Programme. *Economic Development and Cultural Change* **2014**, *63*, 1–26. https://doi.org/10.1086/677753.
24. Fiszbein, A.; Schady, N. *Conditional Cash Transfers: Reducing Present and Future Poverty*; World Bank: Washington, DC, 2009. https://doi.org/10.1596/978-0-8213-7352-1.
25. National Earth System Science Data Center. List of 832 National-Level Poverty Counties in China. National Science & Technology Infrastructure of China, 2020. Accessed: 2024.

doi:10.20944/preprints202603.0372.v1




26. Henderson, J.V.; Storeygard, A.; Weil, D.N. Measuring Economic Growth from Outer Space. *American Economic Review* **2012**, *102*, 994–1028. https://doi.org/10.1257/aer.102.2.994.
27. Didan, K. MODIS/Terra Vegetation Indices 16-Day L3 Global 500m SIN Grid V061, 2021. https://doi.org/10.5067/MODIS/MOD13A1.061.
28. Athey, S.; Tibshirani, J.; Wager, S. Generalized Random Forests. *The Annals of Statistics* **2019**, *47*, 1148–1178. https://doi.org/10.1214/18-AOS1709.
29. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
30. Chernozhukov, V.; Demirer, M.; Duflo, E.; Fernández-Val, I. Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. *Econometrica* **2024**, *92*, 1–40. https://doi.org/10.3982/ECTA17995.