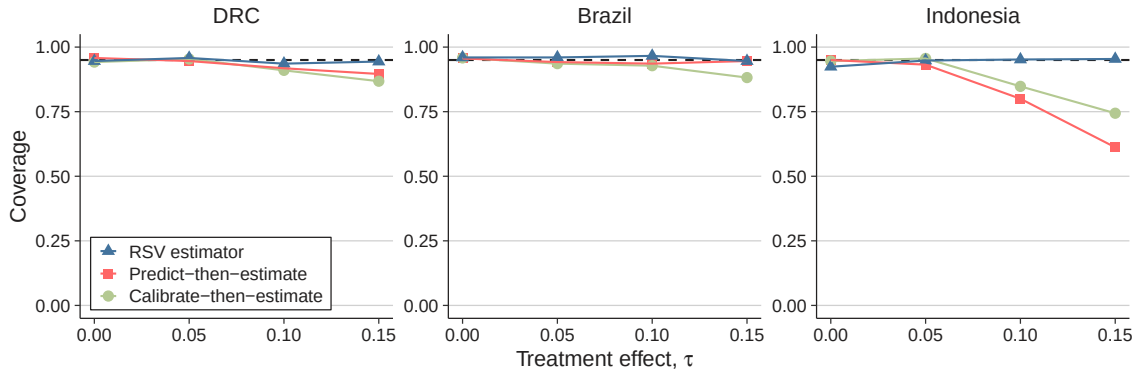
Supplemental Appendix

Causal Inference with Satellite Imagery: A Comparison of Methods for Forest Conservation Data

By HAYA ALSHARIF, ASHESH RAMBACHAN, RAHUL SINGH AND DAVIDE VIVIANO*

A. APPENDIX FIGURES AND TABLES

Panel A. No experimental outcomes



Panel B. Limited experimental outcomes

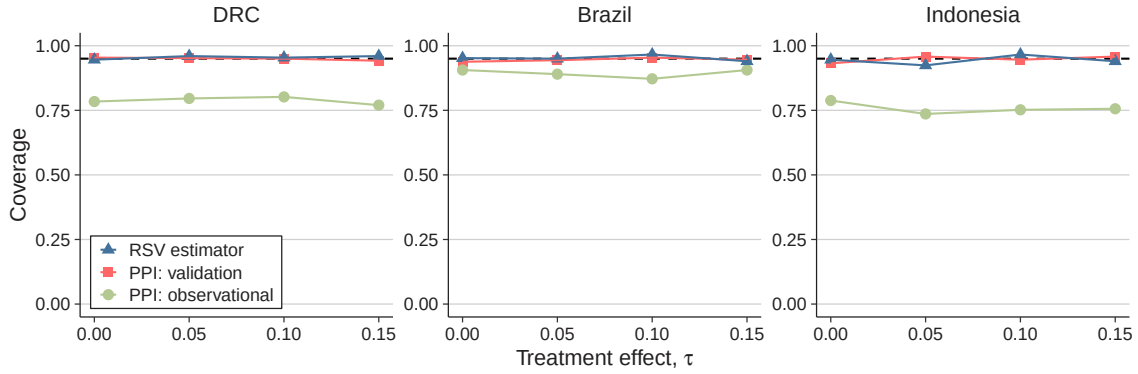


FIGURE A1: COMPARISON OF COVERAGE IN BOTH DATA ENVIRONMENTS.

Note: Coverage is the coverage rate of a 95% confidence interval across 500 simulations. Panel A: The experimental sample has $n_e = 1000$ units and the observational sample has $n_o = 1000$ units. Panel B: The experimental sample has $n_e = 750$ units, the validation sample has $n_{eo} = 250$ units, and the observational sample has $n_o = 1000$ units. See Appendices B and C for details.

* Alsharif: MIT, Department of Economics (haya1@mit.edu); Rambachan: MIT, Department of Economics (asheshr@mit.edu); Singh: Harvard University, Department of Economics and Society of Fellows (rahul_singh@fas.harvard.edu); Viviano: Harvard University, Department of Economics (dviviano@fas.harvard.edu).

B. Semi-Synthetic Simulations With No Experimental Outcomes

We now describe in more detail the design of the semi-synthetic simulations without experimental outcomes in Section I.A of the main text. Let α_e denote the baseline outcome probability $P(Y(0) = 1 \mid S = e)$ in the experimental sample and let α_o denote the baseline outcome probability $P(Y(0) = 1 \mid S = o)$ in the observational sample. For experimental sample size n_e, observational sample size n_o, and treatment effect τ, we perform B replications. We vary $\tau \in \{0, 0.05, 0.10, 0.15\}$.

In each replication, we first draw n_e experimental units. For experimental unit i, we draw treatment independently according to $D_i \sim Bernoulli(0.5)$ and draw the economic outcome as $Y_i \mid D_i \sim Bernoulli(\alpha_e + \tau D_i)$. We then draw the remotely sensed variable from its empirical conditional distribution, sampling $R_i \mid Y_i = y \sim P(R_i \mid Y_i = y)$. We set the sample indicator $S_i = e$ and observe only (D_i, R_i); crucially, we delete the outcome Y_i to mimic the setting where no experimental outcomes are collected. We next draw n_o observational units. For each observational unit i, we draw the outcome as $Y_i \sim \text{Bernoulli}(\alpha_o)$ and again draw the remotely sensed variable from its empirical conditional distribution, $R_i \sim P(R \mid Y = y_i)$ as before. We set $S_i = o$ and observe only (Y_i, R_i). This design enforces the stability and no direct effects assumptions by construction since R_i is independent of D_i given Y_i and the conditional distribution of R_i given Y_i is identical across samples.

We then compare three estimators. The `predict-then-estimate` approach plugs the remotely sensed variable R_i into a regression on the experimental sample, estimating τ as the coefficient β_1 from the regression $R_i = \beta_0 + \beta_1 D_i + \varepsilon_i$. The `calibrate-then-estimate` first fits a logistic regression of Y_i on R_i using the observational sample to obtain predicted outcomes $\hat{Y}_i$. It then regresses these predicted outcomes on treatment in the experimental sample to estimate the treatment effect. `RSV` is implemented using $L = 2$ cross-fitting.

We again note that prediction-powered inference (`PPI`) approaches are not feasible in this data environment, since there exist no labeled observations with variation in treatment assignment D. This data environment deliberately omits all experimental outcomes to represent the most cost-effective and econometrically challenging data collection strategy. This distinction highlights a key advantage of Rambachan, Singh and Viviano (2025): by leveraging the post-outcome structure of remotely sensed variables, they can conduct valid program evaluation even when experimental outcomes are entirely unavailable.

To calibrate the simulation to realistic forest coverage levels, we aggregate the forest coverage data to 1-degree resolution for the DRC, Brazil, and Indonesia. Within each country, we calculate the average outcome within each cell, then compute the 55^{th}, 65^{th}, and 75^{th} percentiles of these cell-level averages across all cells in the country. Throughout the simulations, we set α_e equal to the 65^{th} percentile forest coverage in each country, reflecting experimental sites with moderately high baseline forest cover. For the main text results, we set α_o equal to the 55^{th} percentile, representing observational samples from areas with slightly lower baseline forest coverage.

We focus on sample sizes of $n_e = 1000$ and $n_o = 1000$ in the main text. The additional results below report performance across varying observational sample sizes $n_o \in \{500, 1000, 2000\}$ to demonstrate robustness to the amount of auxiliary data available. We similarly report results varying α_o across the 55^{th}, 65^{th}, and 75^{th} percentiles to examine sensitivity to distribution shift between samples. Each simulation performs $B = 500$, and we summarize normalized bias, coverage probabilities, and standard errors across these replications. For brevity, we present detailed tables only for the DRC, though we find qualitatively similar results across Brazil and Indonesia.

TABLE B1: BIAS IN THE FIRST DATA ENVIRONMENT: DEMOCRATIC REPUBLIC OF THE CONGO.

Panel A. $n_o = 500$

τ	Predict-then-estimate			Calibrate-then-estimate			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.054	0.054	0.054	−0.032	−0.001	−0.030	−0.041	0.065	−0.002
0.05	−0.303	−0.303	−0.303	−0.291	−0.145	−0.254	−0.017	−0.053	−0.065
0.10	−0.487	−0.487	−0.487	−0.498	−0.335	−0.471	−0.009	0.032	−0.036
0.15	−0.699	−0.699	−0.699	−0.859	−0.534	−0.527	0.011	0.102	0.093

Panel B. $n_o = 1000$

τ	Predict-then-estimate			Calibrate-then-estimate			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.054	0.054	0.054	−0.023	0.038	0.041	0.038	−0.007	0.026
0.05	−0.303	−0.303	−0.303	−0.297	−0.213	−0.174	0.019	0.017	−0.015
0.10	−0.487	−0.487	−0.487	−0.565	−0.376	−0.427	0.001	−0.101	0.013
0.15	−0.699	−0.699	−0.699	−0.768	−0.552	−0.648	−0.040	−0.028	−0.045

Panel C. $n_o = 2000$

τ	Predict-then-estimate			Calibrate-then-estimate			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.054	0.054	0.054	0.032	−0.064	0.005	0.018	0.008	−0.031
0.05	−0.303	−0.303	−0.303	−0.148	−0.093	−0.186	0.014	−0.080	0.010
0.10	−0.487	−0.487	−0.487	−0.529	−0.362	−0.325	0.022	0.005	0.008
0.15	−0.699	−0.699	−0.699	−0.759	−0.583	−0.687	−0.088	−0.014	−0.015

Notes: Normalized bias is the average bias divided by the average standard error across $B = 500$ simulations. Each row summarizes results for a treatment effect τ, and each column corresponds to a choice of α_o. See Appendix B for discussion.

TABLE B2: COVERAGE IN THE FIRST DATA ENVIRONMENT: DEMOCRATIC REPUBLIC OF THE CONGO.

Panel A. $n_o = 500$

τ	Predict-then-estimate			Calibrate-then-estimate			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.958	0.958	0.958	0.952	0.960	0.972	0.942	0.944	0.944
0.05	0.946	0.946	0.946	0.926	0.942	0.948	0.950	0.954	0.952
0.10	0.918	0.918	0.918	0.932	0.958	0.920	0.940	0.970	0.940
0.15	0.896	0.896	0.896	0.876	0.902	0.900	0.956	0.970	0.956

Panel B. $n_o = 1000$

τ	Predict-then-estimate			Calibrate-then-estimate			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.958	0.958	0.958	0.942	0.938	0.948	0.946	0.960	0.962
0.05	0.946	0.946	0.946	0.954	0.944	0.948	0.958	0.944	0.944
0.10	0.918	0.918	0.918	0.910	0.944	0.918	0.936	0.940	0.938
0.15	0.896	0.896	0.896	0.868	0.904	0.884	0.944	0.956	0.948

Panel C. $n_o = 2000$

τ	Predict-then-estimate			Calibrate-then-estimate			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.958	0.958	0.958	0.948	0.954	0.962	0.938	0.946	0.926
0.05	0.946	0.946	0.946	0.950	0.950	0.940	0.940	0.950	0.948
0.10	0.918	0.918	0.918	0.906	0.926	0.938	0.940	0.946	0.948
0.15	0.896	0.896	0.896	0.856	0.906	0.910	0.952	0.936	0.918

Notes: Coverage is the coverage rate of a 95% confidence interval across $B = 500$ simulations. Each row summarizes results for a treatment effect τ, and each column corresponds to a choice of α_o. See Appendix B for discussion.

C. Semi-Synthetic Simulations With Limited Experimental Outcomes

We now describe in more detail the design of the semi-synthetic simulations with limited experimental outcomes in Section I.B of the main text. The simulation design follows the same data-generating process as Appendix B, with one key modification: we now collect outcomes for a subset of experimental units.

We fix the total experimental sample size at $n_e = 1000$ and the observational sample size at $n_o = 1000$. Within the experimental sample, we vary the number of units with observed outcomes $n_{eo} \in \{100, 250, 500\}$, representing scenarios where researchers collect traditional outcome measurements for 10%, 25%, or 50% of experimental units while relying on remotely sensed data for the remainder. For the n_{eo} experimental units with observed outcomes, we observe (D_i, Y_i, R_i); for the remaining $n_e - n_{eo}$ experimental units, we observe only (D_i, R_i) as before.

This data environment enables comparison with `PPI` methods, which require some labeled observations with variation in the treatment to correct for errors in the remotely sensed variable. We compare three estimators across the $B = 500$ replications: the RSV estimator; a `PPI` estimator using only the n_{eo} validation units; and the `PPI` estimator using both the n_{eo} validation units and the n_o observational units. The benchmark regression uses only the n_{eo} experimental units with observed outcomes, estimating the treatment effect by regressing Y_i on D_i without using the remotely sensed variable R_i or the observational sample.

We report normalized bias, coverage probabilities, and standard errors across replications, presenting detailed results for the DRC. Qualitatively similar patterns emerge for Brazil and Indonesia.

TABLE C1: BIAS IN THE SECOND DATA ENVIRONMENT: DEMOCRATIC REPUBLIC OF THE CONGO.

Panel A. $n_{eo} = 100$

τ	PPI: validation			PPI: observational			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	−0.024	0.005	−0.038	1.306	−0.006	−1.023	−0.015	0.042	−0.044
0.05	0.033	0.150	0.031	1.280	0.086	−0.989	0.009	−0.040	0.044
0.10	−0.049	0.050	0.011	1.192	−0.000	−0.952	0.005	0.019	0.045
0.15	−0.008	0.023	0.017	1.284	−0.035	−1.055	0.026	−0.044	−0.030

Panel B. $n_{eo} = 250$

τ	PPI: validation			PPI: observational			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.013	−0.042	0.002	1.151	−0.051	−0.881	−0.015	0.047	0.029
0.05	−0.025	−0.056	0.001	1.114	0.009	−1.022	0.060	−0.083	−0.048
0.10	−0.063	−0.054	−0.033	1.122	0.025	−0.981	−0.015	−0.040	−0.063
0.15	−0.015	−0.006	−0.057	1.177	−0.022	−0.950	−0.001	0.068	0.037

Panel C. $n_{eo} = 500$

τ	PPI: validation			PPI: observational			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	−0.014	0.037	0.086	0.895	−0.022	−0.805	−0.028	0.047	−0.045
0.05	0.064	−0.039	0.097	0.922	−0.019	−0.748	−0.025	−0.082	−0.083
0.10	−0.010	−0.073	−0.074	0.968	−0.025	−0.722	−0.044	−0.037	0.041
0.15	0.045	0.020	−0.051	0.964	0.026	−0.711	0.031	0.011	−0.036

Notes: Normalized bias is the average bias divided by the average standard error across $B = 500$ simulations. Each row summarizes results for a treatment effect τ, and each column corresponds to a choice of α_o. See Appendix C for discussion.

TABLE C2: COVERAGE IN THE SECOND DATA ENVIRONMENT: DEMOCRATIC REPUBLIC OF THE CONGO.

Panel A. $n_{eo} = 100$

τ	PPI: validation			PPI: observational			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.934	0.956	0.938	0.738	0.960	0.814	0.952	0.966	0.940
0.05	0.940	0.958	0.956	0.742	0.952	0.854	0.934	0.934	0.946
0.10	0.944	0.948	0.956	0.804	0.940	0.830	0.934	0.962	0.938
0.15	0.956	0.940	0.956	0.752	0.924	0.810	0.942	0.944	0.934

Panel B. $n_{eo} = 250$

τ	PPI: validation			PPI: observational			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.954	0.964	0.944	0.784	0.940	0.856	0.946	0.946	0.956
0.05	0.952	0.954	0.950	0.796	0.942	0.820	0.960	0.940	0.944
0.10	0.950	0.944	0.958	0.802	0.936	0.838	0.954	0.934	0.962
0.15	0.942	0.942	0.952	0.770	0.930	0.860	0.960	0.952	0.952

Panel C. $n_{eo} = 500$

τ	PPI: validation			PPI: observational			RSV estimator		
	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}	55^{th}	65^{th}	75^{th}
0.00	0.946	0.942	0.960	0.870	0.952	0.868	0.950	0.942	0.944
0.05	0.952	0.962	0.950	0.850	0.944	0.894	0.948	0.960	0.944
0.10	0.926	0.968	0.956	0.832	0.956	0.888	0.932	0.940	0.958
0.15	0.942	0.952	0.934	0.848	0.950	0.884	0.958	0.958	0.950

Notes: Coverage is the coverage rate of a 95% confidence interval across $B = 500$ simulations. Each row summarizes results for a treatment effect τ, and each column corresponds to a choice of α_o. See Appendix C for discussion.

TABLE C3: STANDARD ERRORS IN THE SECOND DATA ENVIRONMENT: DEMOCRATIC REPUBLIC OF THE CONGO.

Panel A. $n_{eo} = 100$

τ	PPI: validation			PPI: observational			RSV estimator		
	55th	65th	75th	55th	65th	75th	55th	65th	75th
0.00	0.045	0.045	0.045	0.038	0.039	0.039	0.034	0.034	0.034
0.05	0.046	0.046	0.045	0.039	0.039	0.039	0.034	0.034	0.034
0.10	0.045	0.045	0.045	0.039	0.039	0.039	0.033	0.033	0.033
0.15	0.045	0.045	0.045	0.039	0.039	0.039	0.033	0.033	0.033

Panel B. $n_{eo} = 250$

τ	PPI: validation			PPI: observational			RSV estimator		
	55th	65th	75th	55th	65th	75th	55th	65th	75th
0.00	0.040	0.039	0.040	0.037	0.037	0.037	0.033	0.033	0.034
0.05	0.040	0.039	0.040	0.037	0.037	0.037	0.033	0.033	0.034
0.10	0.039	0.040	0.039	0.037	0.037	0.037	0.033	0.033	0.033
0.15	0.039	0.039	0.039	0.037	0.037	0.037	0.033	0.033	0.033

Panel C. $n_{eo} = 500$

τ	PPI: validation			PPI: observational			RSV estimator		
	55th	65th	75th	55th	65th	75th	55th	65th	75th
0.00	0.043	0.043	0.044	0.042	0.042	0.042	0.034	0.033	0.033
0.05	0.043	0.043	0.044	0.042	0.042	0.043	0.033	0.033	0.034
0.10	0.043	0.043	0.043	0.042	0.042	0.042	0.033	0.033	0.033
0.15	0.043	0.043	0.043	0.042	0.042	0.042	0.033	0.033	0.033

Notes: Average standard errors are calculated across $B = 500$ simulations. Each row summarizes results for a treatment effect τ, and each column corresponds to a choice of α_o. See Appendix C for discussion.

D. Semi-Synthetic Simulations with Alternative Outcome Definition

We now present semi-synthetic simulations using an alternative definition of the binary outcome Y. We explore a different specification where the threshold τ equals the average forest cover across grid cells in each country. This alternative definition could correspond to a country-specific forest code that designates areas with sufficient forest cover—relative to the country's mean—for additional environmental protection.

Using this alternative outcome, we again split the data for each country into two halves. On one half, we train predictors for the new outcome using MOSAIKS embeddings of satellite images. The resulting predictions $R \in [0,1]$ are our RSVs, and we again find they are highly predictive of the forest cover outcome on the held-out sample: AUC of 0.994 for the DRC, AUC of 0.993 for Brazil, and AUC of 0.977 for Indonesia. We use the held-out sample for the semi-synthetic evaluation, simulating treatment assignments and treatment effects and sampling from the empirical distribution of R conditional on the simulated outcome as described in Section I.A and Section I.B of the main text.

For the first data environment, Figure D1 displays normalized bias for the three estimators: `predict-then-estimate`, `calibrate-then-estimate`, and `RSV`. The results parallel the main text. As before, `RSV` exhibits negligible normalized bias across all countries.

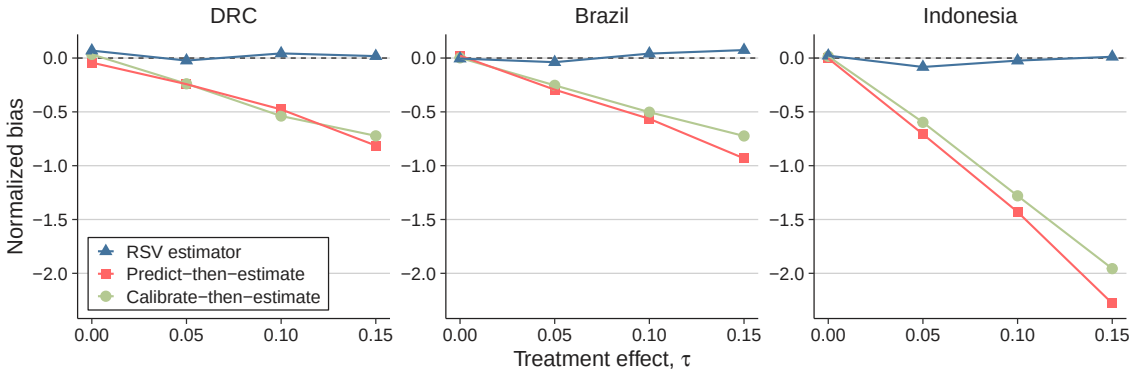


FIGURE D1: COMPARISON OF BIAS IN THE FIRST DATA ENVIRONMENT: COUNTRY-SPECIFIC AVERAGE THRESHOLD

Note: Normalized bias is the average bias divided by the average standard error across 500 simulations. The experimental sample has $n_e = 1000$ units and the observational sample has $n_o = 1000$ units. See Appendix D for discussion.

For the second data environment, Figures D2 and D3 reproduce our results from the main text with the new outcome. We again find that prediction-powered inference (`PPI`) using the observational sample remains badly biased. `RSV`, by contrast, exhibits minimal bias across all three countries. Furthermore, we find that `RSV` delivers precision gains by comparing it against the benchmark regression using the validation sample and against `PPI` using the validation sample.

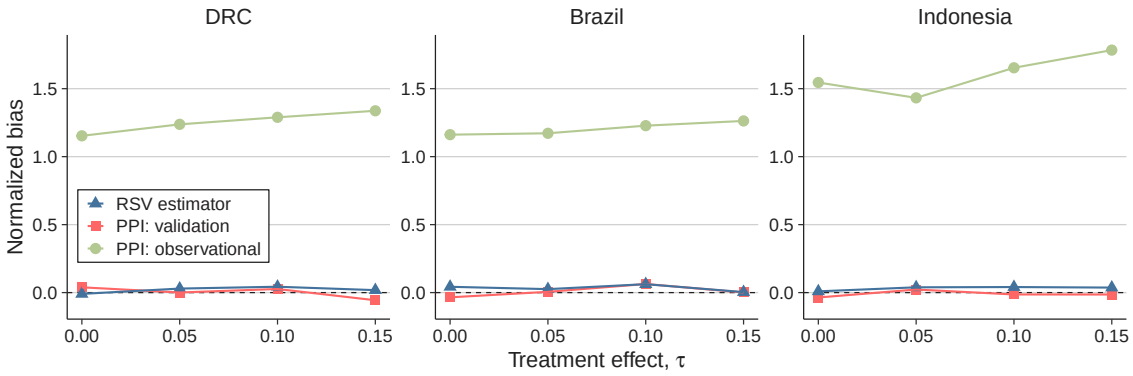


FIGURE D2: COMPARISON OF BIAS IN THE SECOND DATA ENVIRONMENT: COUNTRY-SPECIFIC AVERAGE THRESHOLD

Note: Normalized bias is the average bias divided by the average standard error across 500 simulations. The experimental sample has $n_e = 750$ units, the validation sample has $n_{eo} = 250$ units, and the observational sample has $n_o = 1000$ units. See Appendix D for discussion.

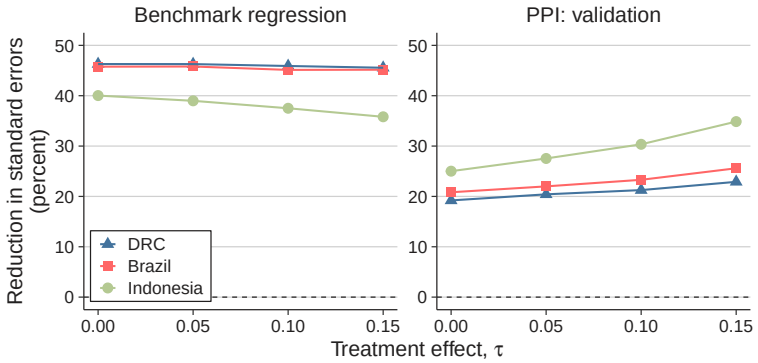


FIGURE D3: COMPARISON OF STANDARD ERRORS IN THE SECOND DATA ENVIRONMENT: COUNTRY-SPECIFIC AVERAGE THRESHOLD

Note: Reduction in standard errors is the percentage decrease in standard errors achieved by `RSV` relative to benchmark regression (left panel) and `PPI` (right panel), averaged across 500 simulations with $n_e = 750$ units in the experimental sample, $n_{eo} = 250$ units in the validation sample, and $n_o = 1000$ units in the observational sample. See Appendix D for discussion.

*

REFERENCES

Rambachan, Ashesh, Rahul Singh, and Davide Viviano. 2025. "Program Evaluation with Remotely Sensed Outcomes." Preprint, arXiv. https://arxiv.org/abs/2411.10959.