Program Evaluation with Remotely Sensed Outcomes

Ashesh Rambachan
MIT

Rahul Singh
Harvard

Davide Viviano*
Harvard

First draft: November 2024
This draft: June 2026

arXiv:2411.10959v5 [econ.EM] 10 Jun 2026

Abstract

We study causal inference in experiments and quasi-experiments, where the economic outcome is imperfectly measured by a remotely sensed variable. Remotely sensed variables are low-cost, scalable, and predictive of economic outcomes; examples include satellite imagery and mobile phone activity. We model the remotely sensed variable as post-outcome: variation in the economic outcome causes variation in the remotely sensed variable. For example, changes in environmental quality cause changes in satellite imagery, not vice versa. Under this assumption, we nonparametrically identify the causal parameter by combining experimental and observational data, and develop a method for robust $n^{-1/2}$ inference.

Keywords: Causal inference, data fusion, satellite imagery.

*Email: asheshr@mit.edu, rahul_singh@fas.harvard.edu, dviviano@fas.harvard.edu. We thank Isaiah Andrews, Josh Angrist, Arun Chandrasekhar, Raj Chetty, Kevin Chen, Ben Deaner, Melissa Dell, Cristophe Gaillac, Paul Goldsmith-Pinkham, Seema Jayachandran, Namrata Kala, Sylvia Klosin, Ben Olken, Jonathan Roth, Jesse Shapiro, Matthieu Stigler, and Alex Tetenov, as well as audiences at Berkeley, Bocconi, CEMFI, CREST/Sciences Po, Harvard/MIT, the NBER Summer Institute, the Online Causal Inference Seminar, Opportunity Insights, Princeton, Stanford, University of Bologna, University of Chicago, University of Southern California, University of Toronto, University of Geneva, and Yale for helpful discussions. Haya Alsharif, Peter Chen, Marvin Lob, Leonard Mushunje, Miriam Nelson, Kevin Wang, and Sammi Zhu provided excellent research assistance. Davide Viviano gratefully acknowledges funding from the Harvard Griffin Fund in Economics and NSF Grant SES 2447088. Replication code is available here. We provide the `remoteoutcome` R package to implement our method.

1 Introduction

Traditional program evaluation relies on surveys to measure outcomes, but many economic outcomes are costly or infeasible to collect at scale. Researchers increasingly rely on remotely sensed variables (Appendix Figure A.1). Examples include night lights as measures of local economic activity, roofing materials as measures of housing quality, mobile phone transactions as measures of poverty or consumption, and satellite images as measures of pollution, deforestation, fires, flooding, and local poverty.[1] These variables are cheap, scalable, and predictive of economic outcomes in observational data, but imperfect.[2] We study how to identify treatment effects in experiments and quasi-experiments where the outcome is unobserved but remotely sensed variables are available.

To formalize this setting, we consider two samples, each containing the remotely sensed variable. The experimental sample contains a randomly assigned treatment and the remotely sensed variable, but the outcome is missing due to logistical or cost constraints. The observational sample links the outcome to the remotely sensed variable, but comes from a different, non-experimental context. For this reason, the treatment variable is missing or confounded in the observational sample, and treatment effects may differ across the two populations.

What the two samples share is the remotely sensed variable, which we view as a "post-outcome" measurement: variation in the outcome causes variation in the remotely sensed variable. For example, while evaluating financial incentives to reduce crop burning, the experimental sample contains randomized contracts and satellite images of fields, while the observational sample links satellite images to surveyor measurements of whether fields were burned. In this example, crop burning causes changes in satellite images, not vice versa.

Our main assumption is *measurement stability*: the conditional distribution of the remotely sensed variable, given the outcome and treatment, is the same across samples. This assumption is plausible when the same measurement technology is used across samples; for example, the same satellite generates the images in both datasets. Therefore, we study a

[1] Night lights (Chen and Nordhaus, 2011; Henderson et al., 2012; Asher et al., 2021). Roofing material (Marx et al., 2019; Michaels et al., 2021; Huang et al., 2021). Mobile phone transactions (Blumenstock et al., 2015; Aiken et al., 2025). Satellite images of pollution (Currie et al., 2023); deforestation (Jayachandran et al., 2017; Assuncao et al., 2023); fires (Jack et al., 2025; Balboni et al., 2025); flooding (Chen et al., 2017; Patel, 2024); local poverty (Jean et al., 2016). See Donaldson and Storeygard (2016), Burke et al. (2021), and Jack and Walker (2023) for reviews.

[2] Recent work documents that remotely sensed measures of economic outcomes can be noisy (see, e.g., Bluhm and Krause, 2022; Fowlie et al., 2019; Corral et al., 2025; Josephson et al., 2026).

setting where the treatment mechanism and outcome mechanism vary across samples, but the *sensing* mechanism is the same.

Our primary contribution is an identification formula for the treatment effect that requires no outcome data in the experimental sample. It exploits stability to combine the experimental and observational samples. Intuitively, the experimental sample identifies how the treatment affects the remotely sensed variable, while the observational sample identifies how the outcome affects the remotely sensed variable. The treatment effect is identified by a conditional moment restriction, in which the remotely sensed variable relates experimental treatment variation to observational outcome variation (Theorems 1–2).

By characterizing a set of conditional moment restrictions, we also provide diagnostic tests for the key identifying assumptions. One diagnostic test evaluates whether the remotely sensed variable is sufficiently informative about outcome variation, similar to testing for weak instruments. Another diagnostic test evaluates whether different representations of the remotely sensed variable yield the same treatment effect estimate, similar to an overidentifying restrictions test.

As a secondary contribution, we develop a method for $n^{-1/2}$ inference on the treatment effect that is robust to misspecification and that accommodates complex machine learning. We explain how to use predicted outcomes for valid inference. Moreover, we explain how to use predicted outcomes, treatments, and sample indicators for more efficient inference through a connection to optimal instruments (Chamberlain, 1987; Newey, 1993). Unlike the standard inference guarantees for double machine learning (Chernozhukov et al., 2018), $n^{-1/2}$ inference remains valid even if all of these predictors are misspecified and lack rate conditions. From a practical perspective, this result allows complex neural networks to compress high-dimensional, unstructured data.

Our framework extends in many directions (Section 5). One extension (Section 5.1) considers not only an experimental sample and an observational sample but also a validation sample: units for which we observe the randomized treatment, the outcome, and the remotely sensed variable. The validation sample may be non-randomly selected, with different treatment effects than the experimental sample. Another extension (Section 5.2) replaces the experimental sample with a quasi-experimental sample: instead of randomized treatments, we may have instrumental variables or a difference-in-differences design.

In the setting we study, we show that some existing methods are biased. Fifty percent of papers in general interest economics journals from 2015–2024 that use remotely sensed outcomes

employ a two-step method: (i) train a predictor of the outcome from the remotely sensed variable in the observational sample; (ii) use it to predict outcomes in the experimental sample, and report a difference of predicted outcome means.[3] Without covariates, this method suffers from attenuation bias (Proposition 1), even if samples are randomly selected (Proposition C.2). With covariates and heterogeneity, this method may flip the sign of the average treatment effect (Footnote 8). Another important method is called prediction-powered inference (PPI) (Angelopoulos et al., 2023), which adds bias correction terms estimated from an appropriate auxiliary sample. In our setting, it can be biased when the difference in outcome means in the experimental sample does not equal the difference in outcome means in the observational sample (Proposition 2 and Remark 6). Such a distribution shift is plausible when the two samples come from different populations, time periods, or regions.

We illustrate our method in calibrated simulations and in real-world applications, pertaining to: (i) forest cover in Uganda (Hansen et al., 2013; Jayachandran et al., 2017); (ii) household poverty in India (Muralidharan et al., 2016, 2023), and (iii) crop burning in India (Jack et al., 2025). Across settings, we assess our identifying assumptions using our diagnostic tests. In simulations, our method is approximately unbiased with nominal coverage. By contrast, we find the two-step method and PPI are generally biased. In the household poverty application, our method yields estimates close to the "oracle" difference-in-means estimate obtained using experimental outcomes. In the crop burning application, our analysis suggests that the two-step method underestimates the effectiveness of the intervention by 47 percent.

1.1 Related Work

Our causal model differs from existing models along three dimensions: (i) the causal direction between the auxiliary variable and the outcome, (ii) robustness to misspecification, and (iii) data requirements. We provide formal comparisons in Section 3.2.

In the surrogacy model (Prentice, 1989; Athey et al., 2025; Kallus and Mao, 2024), the auxiliary variable—called a surrogate—is pre-outcome: it is a "short-term" measurement that mediates between the treatment and "long-term" outcome. By contrast, in our model, the remotely sensed variable is post-outcome: the outcome mediates between the treatment and our measurement. We prove that misusing a remotely sensed variable as a surrogate leads

[3]We survey American Economic Association (AEA) journals, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*. The remaining fifty percent use a similar logic, without an explicit formula.

to attenuation bias (Proposition 1). The negative control literature extends the surrogacy model to address unobserved confounding (Ghassami et al., 2022; Imbens et al., 2025). In these extensions, misusing a remotely sensed variable incurs bias.

Several recent papers propose alternative methods for inference with remotely sensed outcomes. Proctor et al. (2023) propose a multiple imputation method that models the outcome as a function of the remotely sensed variable and imputes it across samples. It is akin to the surrogacy methods discussed above and in Section 3.2.1. Alix-Garcia and Millimet (2023); Torchiana et al. (2025) instead model the misclassification process for binary and panel outcomes, requiring correct specification of a model relating the remotely sensed variable to the outcome. By contrast, our method does not require correct specification of any model relating the remotely sensed variable to the outcome. This distinction is important because the remotely sensed variable may be high-dimensional or unstructured.

One major difference between our method and PPI methods (Angelopoulos et al., 2023, 2024; Kluger et al., 2025) is the data requirement. PPI methods require joint observations of the treatment, outcome, and remotely sensed variable, in both treatment arms; these methods are infeasible without such observations. Sanford et al. (2025); Lu et al. (2025); Pelletier et al. (2026); Fong and Tyler (2021); Allon et al. (2023); Egami et al. (2023); Carlson and Dell (2025); Ji et al. (2025) share this data requirement. By contrast, our method remains feasible when the experimental sample excludes outcomes, and when the observational sample excludes treatments (Theorem 1).

Another difference is that our method requires stability of the sensing mechanism, whereas we show that PPI applied with an observational sample implicitly requires stability of the outcome mechanism (Proposition 2). By contrast, our method allows for the observational sample to come from a different population, time period, or region, with different baseline outcomes and treatment effects. See Remark 6 for comparisons in other data environments.

Compared to the vast literature on data combination (e.g., Cross and Manski, 2002; Ridder and Moffitt, 2007; Bareinboim and Pearl, 2016; D'Haultfœuille et al., 2025) and nonclassical measurement error (e.g., Chen et al., 2011; Schennach, 2020), we place a different main assumption. Several influential works handle measurement error in moment condition models by assuming that the conditional distribution of the variable of interest, given the imperfect measurement, is stable across samples (Chen et al., 2005, 2008; Graham et al., 2016), akin to the surrogacy model. Our main assumption is the opposite: the conditional distribution of the imperfect measurement, given the variable of interest, is stable. Our assumption is natural

when we believe the sensing mechanism, e.g. the satellite technology, is stable across samples rather than the outcome mechanism. This distinction leads to a distinct identifying formula.

Our framework is closer to nonclassical measurement error models with instruments or repeated measurements (Hu and Schennach, 2008; Hu, 2008). However, their setting has no observational sample in which the outcome is observed, and therefore requires an additional normalization to pin down how the imperfect measurement is distributed around the true outcome. In our setting, the outcome is observed in the observational sample, and so we do not require an additional assumption beyond stability of the sensing mechanism.

Our framework is also related to the "shadow-variable" model for recovering the outcome distribution in a single sample with nonignorable missingness (e.g., Miao and Tchetgen Tchetgen, 2016; Miao et al., 2024). A shadow variable is an auxiliary variable that is relevant to the missing outcome yet excluded from the missingness mechanism. The main difference is that we study a two-sample causal inference problem, rather than a single-sample missing-data problem studied in this literature. Our framework includes scenarios where the treatment is only observed in one sample while the outcome is only observed in the other sample, so data combination is necessary for identification of the treatment effect; the single-sample shadow-variable approach would not, by itself, identify it.

Finally, a rich literature studies mismeasured or missing covariates (e.g., Fan et al., 2014; D'Haultfœuille et al., 2024), possibly imputed with machine learning (Battaglia et al., 2024). By contrast, in our setting, only the outcome is missing yet a post-outcome variable is present.

2 Model and Identifying Assumptions

2.1 Setting and Causal Parameter

The researcher observes units in two samples, indicated by the variable $S \in \{e, o\}$: an experimental sample ($S = e$) and an observational sample ($S = o$).

In the experimental sample ($S = e$), we observe pre-treatment covariates $X \in \mathcal{X}$ and a binary treatment $D \in \{0,1\}$. The outcome $Y \in \mathcal{Y}$, however, is entirely missing. In its place, we observe a remotely sensed variable $R \in \mathcal{R}$. (In Section 5.1, we allow the outcome to be only partially missing in the experimental sample.) We place no restriction on the form of the remotely sensed variable: examples include unstructured data such as satellite images or digital traces, a pretrained embedding of unstructured data, or the output of a pretrained

predictor. The researcher would like to use the remotely sensed variable R as an imperfect measurement of the outcome Y in the experimental sample.

The parameter of interest is the effect of the treatment D on the outcome Y in the experimental sample. Though the outcome Y is unobserved in the experimental sample, we may still define its potential outcomes $Y(d)$ and the causal parameter.[4]

Definition 1 (Causal parameter)**.** The average treatment effect (ATE) in the experimental sample is $\tau := \mu(1) - \mu(0)$, where $\mu(d) := \mathbb{E}\{Y(d) \mid S = e\}$.

Without further assumptions, when R is an imperfect measurement of Y, it is generally impossible to point identify this causal parameter (Horowitz and Manski, 1995).

Motivated by empirical work in environmental and development economics, we leverage an auxiliary sample that we refer to as the observational sample ($S = o$). For these units, we observe the covariates X, outcome Y, and remotely sensed variable R. We may or may not observe the treatment D. If we do observe the treatment and there is variation in the treatment (though it may be confounded), we refer to this scenario as having "complete" observational cases. If we do not observe the treatment or if the treatment is deterministic in the observational sample, we set $D = 0$ for all units in the observational sample, and we refer to this scenario as having "incomplete" observational cases. Incomplete observational cases will require stronger assumptions for point identification.

Table 1: Summary of the Data Environment.

	Sample Indicator S		
	Experimental	Observational: Complete	Observational: Incomplete
Covariates X	✓	✓	✓
Treatment D	✓	✓	Missing or Deterministic
Outcome Y	Missing	✓	✓
Remotely Sensed Variable R	✓	✓	✓

Notes: ✓ denotes that the variable is observed. When treatment is missing or deterministic in the observational sample, we encode it as $D = 0$ in the observational sample.

Table 1 summarizes the data environment. Each unit is characterized by the vector $(S, X, D, Y(0), Y(1), R)$, which we assume to be independent and identically distributed across units. For units in the experimental sample ($S = e$), we observe (X, D, R). For units in the

[4]This definition rules out spillovers. We extend our framework to allow for spillovers in Section 5.3.

observational sample ($S=o$), we observe (X,D,Y,R) if there are complete cases or (X,Y,R) if there are incomplete cases.

Example 1 (Environmental programs)**.** Consider a randomized experiment evaluating whether incentives for environmental conservation D, such as payments for ecosystem services (PES), reduce harmful environmental behaviors Y, such as deforestation or crop burning. Directly measuring these outcomes is costly: monitoring tree cover requires on-site surveyors (Jayachandran et al., 2017), and recording crop management practices requires field visits during the narrow post-harvest/pre-tilling window when crop burning occurs (Jack et al., 2025). By contrast, satellite imagery R is cheap and predictive of these environmental outcomes (e.g., Hansen et al., 2013; Burke et al., 2021; Walker et al., 2022). In summary, in the experimental sample, only the treatment assignment D and remotely sensed variable R are easily observed.

Therefore, the researcher turns to an observational sample linking the remotely sensed variable R to the outcome Y. For example, Jayachandran et al. (2017) collect ground-based forest measurements linked to satellite images, for a sample of locations distinct from the experimental units, collected at different dates than the images used in their experimental analysis. Jack et al. (2025) collect ground-based, random spot checks for crop burning, but the spot checks could not be synchronized with the narrow post-harvest/pre-tilling window (see Section 7 for discussion). Neither is well-suited to serve as the primary basis for treatment effect estimation in the experiments. In our framework, both qualify as valid observational samples. ▲

Example 2 (Household poverty)**.** Consider a randomized experiment evaluating whether an anti-poverty program D, such as unconditional cash transfers (e.g., Egger et al., 2022; Aiken et al., 2025) or biometrically authenticated payments (Muralidharan et al., 2023), reduces village-level poverty Y. Directly measuring poverty requires detailed household surveys on consumption, assets, and income in dispersed rural villages. By contrast, remotely sensed variables R, such as satellite imagery (Jean et al., 2016; Huang et al., 2021; Rolf et al., 2021) and digital trace data from mobile phones (Blumenstock et al., 2015; Aiken et al., 2022, 2025), are widely available and predictive of poverty. The researcher has access to an observational sample linking R to poverty measurements Y (e.g., census statistics or existing household surveys in other regions). Our framework combines these samples to identify the causal effect of the anti-poverty program. ▲

2.2 Identifying Assumptions

We present three assumptions under which the experimental and observational samples can be combined to identify the causal parameter.

Our identifying assumptions allow (X,Y,R) to be discrete or continuous. For exposition, we slightly abuse notation: for a random variable W, we use $f_W(\cdot\,|\,...)$ to denote its (conditional) probability mass function if W is discrete, or its (conditional) density if W is continuous. Formally, $f_W(\cdot\,|\,...)$ is the Radon-Nikodym derivative.

2.2.1 Experimental Unconfoundedness

Assumption 1 (Experimental unconfoundedness)**.** Suppose the following:

i. Stable unit treatment value (SUTVA): $Y = DY(1)+(1-D)Y(0)$ almost surely.

ii. Randomization: $D \perp\!\!\!\perp \{Y(0),Y(1)\}\,|\,X,S=e$.

iii. Overlap: $\Pr(D=1\,|\,X,S=e)$ is bounded away from zero and one almost surely.

Assumption 1 imposes standard conditions for causal inference in randomized experiments. In many applications involving remotely sensed variables, such as Examples 1 and 2, the assumption is satisfied by design: the treatment is randomly assigned to well-separated, aggregate units. In Section 5.2, we extend our analysis to allow quasi-experimental designs in the experimental sample, such as instrumental variables and difference-in-differences. In Section 5.3, we extend our analysis to allow some spillovers across units in the experimental sample.

2.2.2 Stability of the Remotely Sensed Variable

Because the outcome is not observed in the experiment, our aim is to learn the relationship between the remotely sensed variable and the outcome in the observational sample and then to "transport" it to the experimental sample. Our next assumption justifies this intuition.

Assumption 2 (Stability of the remotely sensed variable)**.** Suppose the following:

i. Stability: $S \perp\!\!\!\perp R\,|\,X,D,Y$.

ii. Common support: for some outcome support $\mathcal{Y}$, $\Pr(Y\in\mathcal{Y}\,|\,S=e,X)=1$ almost surely, and $f_Y(y\,|\,S=o,X)$ is bounded away from zero almost surely for all $y\in\mathcal{Y}$.

iii. Coverage: $f_R(r\,|\,S,X,D,Y)$ is bounded away from zero almost surely for all $r\in\mathcal{R}$.

iv. Two samples: $\Pr(S=e \mid X)$ is bounded away from zero and one almost surely.

Assumption 2(i) is the main assumption of our framework: conditional on (X,D,Y), the distribution of the remotely sensed variable R is *stable* across the experimental and observational samples. It formalizes the idea that the sensing mechanism (i.e., how the outcome generates the remotely sensed variable) is invariant across samples: $f_R(r \mid S=e,X,D,Y) = f_R(r \mid S=o,X,D,Y)$. This assumption is most plausible when the sensing mechanism is a shared technology across the samples (e.g., the same satellite) and when factors that affect the measurement process are similar across the samples (e.g., the same season, soil conditions, and land-use patterns). In Sections 6-7, we provide evidence illustrating that Assumption 2(i) is plausible across three applications in environmental and development economics.

Assumption 2(i) does not require treatment effects to be stable across samples. The treatment mechanism $\Pr(D=1 \mid S,X)$ and the outcome mechanism $f_{Y(d)}(y \mid S,X,D)$ may depend on the sample S.

As mentioned earlier, our framework is agnostic to whether R is (i) unstructured data, (ii) a pretrained embedding of unstructured data, or (iii) the output of a pretrained predictor applied to unstructured data. Stability of (i) implies stability of (ii) and (iii).[5]

The remaining conditions of Assumption 2 are mild regularity conditions. Assumption 2(ii) requires that the outcome in the observational sample has a common (or larger) support than the outcome in the experiment. Assumption 2(iii) ensures that the distribution of the remotely sensed variable does not degenerate for any stratum. Assumption 2(iv) requires that both samples are observed with positive probability across all covariate values.

Example 1 (continued): Returning to Example 1, satellite images detect environmental outcomes through physical processes: burned cropland alters soil reflectance, and deforestation changes color saturation. Assumption 2(i) requires that these processes are stable across samples. It does not require that PES contracts have the same effects on environmental outcomes across samples. ▲

Example 2 (continued): In Example 2, satellite images detect poverty through observable markers such as roofing materials or the footprint of the house, while mobile phone data detect poverty through activity patterns. Assumption 2(i) requires that these sensing mechanisms are stable across samples. Again, it does not require that anti-poverty programs have the

[5]Let U be some unstructured data and $f(\cdot)$ be a fixed measurable function with $R=f(U)$. Stability of the unstructured data $S \perp\!\!\!\perp U \mid X,D,Y$ implies stability of the pretrained output $S \perp\!\!\!\perp R \mid X,D,Y$.

same effects, outcome mechanisms, or treatment mechanisms across samples. ▲

2.2.3 Complete versus Incomplete Observational Cases

Assumptions 1-2 suffice to identify the causal parameter when there are complete observational cases: when the (possibly confounded) treatment is observed and varies in the observational sample. However, when observational cases are incomplete—i.e., when the treatment is missing or deterministic in the observational sample—an additional assumption is needed for point identification.

Assumption 3 (Observational completeness)**.** Suppose that *either* condition holds:

i. Complete observational cases: $\Pr(D=1 \mid S=o,X,Y)$ is bounded away from zero and one almost surely; *or*

ii. No direct effect: $D \perp\!\!\!\perp R \mid X,Y$.

Assumption 3 requires only one of these two conditions to hold.

Complete observational cases: Assumption 3(i) requires access to complete cases: joint observations of (X,D,Y,R) in which D varies in the observational sample. The treatment need not be randomized; it may suffer from unobserved confounding in the observational sample. The treatment D may also have a direct effect on the remotely sensed variable R. Under Assumption 3(i), no further causal assumptions are needed.

Incomplete observational cases: If Assumption 3(i) fails, we have incomplete cases: no observations of (X,D,Y,R) where D varies (for at least some values of (X,Y)). Without joint observations of the outcome and treatment, an additional causal restriction is necessary. Assumption 3(ii) fills this gap, requiring that the treatment D affects the remotely sensed variable R only via its effect on the outcome Y. In Example 1, it requires that the PES contract affects the satellite image only via its effect on crop burning; the contract has no direct effect on the specific infrared bands used to measure charred soil. The assumption becomes more plausible when Y is a vector of outcomes, which may capture all mechanisms through which the treatment affects the remotely sensed variable.[6]

Assumption 2(i) and Assumption 3(ii) imply that $(S,D) \perp\!\!\!\perp R \mid X,Y$. These assumptions are jointly testable, as we discuss in Section 4. Section 5.5 extends our framework to the setting with both incomplete observational cases and direct effects.

[6] It is possible to have Assumption 3(i) hold for some covariate values $X=x$ and Assumption 3(ii) hold for other covariate values $X=x'$. We leave this extension to future research.

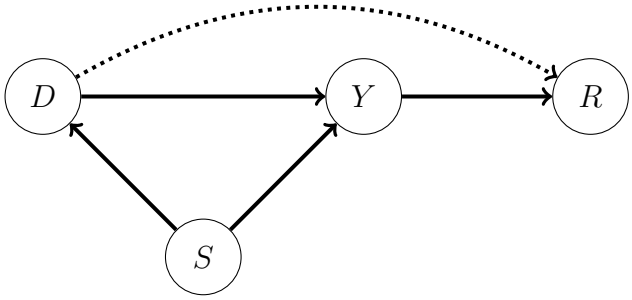


Figure 1: Causal graph for remotely sensed variables under Assumptions 3(i) versus 3(ii). Complete cases allow the dotted line. Assumption 2 rules out the line from S to R.

Figure 1 illustrates our identifying assumptions as a causal graph. The treatment affects the outcome, which in turn affects the remotely sensed variable. Under our identifying assumptions, the remotely sensed variable R is *post-outcome*: R is caused by Y. By Assumption 2, the sample indicator does not change the conditional distribution of the remotely sensed variable given the treatment and outcome. Depending on which version of Assumption 3 is imposed, the treatment may have a direct effect on the remotely sensed variable, as illustrated by the dotted line. See also Appendix Table A.1.

3 Identification via Stability of Remotely Sensed Variables

Our primary contribution is to nonparametrically identify the treatment effect by combining the experimental and observational samples, appealing to stability of the sensing mechanism. Under our assumptions, the treatment effect is identified by a conditional moment restriction, in which the remotely sensed variable relates experimental treatment variation to observational outcome variation. We compare our identification strategy to alternative approaches, clarifying how our assumptions and data requirements differ. In our setting, several commonly used methods are infeasible, biased, or inefficient.

3.1 Main Result

We derive a formula for data combination, without restricting either the distribution or the dimension of the remotely sensed variable.

In this section, we impose that the outcome is discrete with finite support $\mathcal{Y} = \{y_1, \ldots, y_K\}$ for $K \geq 2$. A discrete outcome simplifies our discussion of identification, estimation, and inference. It is also realistic in some empirical applications, where the outcome is often binary (e.g., crop

burning versus no crop burning). Section 5.6 extends our results to continuous outcomes.

Throughout, we assume that the support of R is at least as large as the support of Y. Specifically, R may be discrete or continuous, and we are agnostic about its complexity. We separately consider incomplete and complete observational cases, under Assumption 3(ii) and 3(i), respectively.

3.1.1 Incomplete Observational Cases: No Direct Effects

We express the distribution of the remotely sensed variable in the experimental sample as a mixture (i.e., linear combination) of distributions in the observational sample. The mixture weights are the potential outcome probabilities we wish to identify.

Lemma 1 (Identification as conditional densities)**.** Suppose Assumptions 1, 2, and 3(ii) hold. Then, for any $d \in \{0,1\}$,

$$f_R(R \mid S=e,X,D=d) = \sum_{k=1}^{K} f_R(R \mid S=o,X,Y=y_k)\mu_k(d,X), \quad \mu_k(d,X) := \Pr\{Y(d)=y_k \mid S=e,X\}.$$

Proof. See Appendix B.1.1. □

By Lemma 1, we recover the treatment effect by combining (i) how the remotely sensed variable reflects the treatment in the experimental sample, and (ii) how the remotely sensed variable reflects the outcome in the observational sample.

While Lemma 1 identifies the treatment effect, it involves conditional densities of the remotely sensed variable. Estimating these densities could be challenging, especially when the remotely sensed variable is high-dimensional, as in contexts with unstructured data or pretrained embeddings. Therefore, we use Bayes' rule to rewrite Lemma 1 as a conditional moment equation. This transformation avoids estimation of the conditional densities.

Let $\theta(X) := (\mu_1(1,X)-\mu_1(0,X),\ldots,\mu_{K-1}(1,X)-\mu_{K-1}(0,X))^\top$ denote the vector of conditional treatment effects on outcome probabilities. Identification of $\theta(X)$ implies identification of the conditional average effect $\tau(X) := \mathbb{E}\{Y(1) - Y(0) \mid S = e, X\}$ since $\tau(X) = \sum_{k=1}^{K-1}(y_k - y_K)\theta_k(X)$.

Theorem 1 (Identification as conditional moment)**.** Under the conditions of Lemma 1,

$$\mathbb{E}\{\Delta^e(X) - \Delta^o(X)^\top \theta(X) \mid X,R\} = 0 \text{ almost surely, where}$$

$$\Delta^e(X) := \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e \mid X)} - \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e \mid X)},$$

$$\Delta^o_k(X) := \frac{1\{Y=y_k,S=o\}}{\Pr(Y=y_k,S=o \mid X)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o \mid X)} \text{ with } \Delta^o(X) = \left(\Delta^o_1(X),\ldots,\Delta^o_{K-1}(X)\right)^\top.$$

The scalar $\Delta^e(X)$ summarizes the treatment variation in the experimental sample, while the vector $\Delta^o(X)$ summarizes the outcome variation in the observational sample.

Proof. See Appendix B.1.2. □

Example 3 (Binary outcome, no covariates)**.** For intuition, consider the setting where the outcome is binary and there are no covariates: $K=2$, $y_1=1$, $y_2=0$, and $\mathcal{X}=\varnothing$. Treatment and outcome variation simplify to scalars Δ^e and Δ^o. Theorem 1 reduces to a ratio:

$$\tau=\frac{\mathbb{E}[\Delta^e\,|\,R]}{\mathbb{E}[\Delta^o\,|\,R]},\quad \Delta^e=\frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)}-\frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)},\quad \Delta^o=\frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)}-\frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}.$$

This ratio appears to be a new formula for data combination, where the numerator and denominator are from different samples. The numerator captures the effect of both D and Y on R, so we must divide it by the effect of Y on R in order to isolate the effect of D on Y. Crucially, there is no need to specify the conditional distribution of R given Y. ▲

More generally, Theorem 1 identifies the conditional average treatment effect, for each covariate stratum, through a system of conditional moments. Each conditional moment contrasts one possible outcome value against a baseline outcome value. These conditional moments imply unconditional moments.

Corollary 1 (Identification as representation)**.** Suppose the conditions of Lemma 1 hold, and consider any measurable function $H:\mathcal{X}\times\mathcal{R}\to\mathbb{R}^J$ with $J\geq K-1$. Then,

$$\mathbb{E}[H(X,R)\Delta^e(X)\,|\,X]=\mathbb{E}[H(X,R)\Delta^o(X)^\top\,|\,X]\theta(X).$$

If $\mathbb{E}[H(X,R)\Delta^o(X)^\top\,|\,X]$ has full column rank, then $\theta(X)$ and hence τ are identified.

Example 3 (continued): Return to the setting of binary Y and no X for intuition. Consider any measurable function $H:\mathcal{R}\to\mathbb{R}$. Corollary 1 simplifies to the ratio

$$\tau=\frac{\mathbb{E}[H(R)\Delta^e]}{\mathbb{E}[H(R)\Delta^o]}\text{ when }\mathbb{E}[H(R)\Delta^o]\neq 0.$$

The ratio resembles the Wald estimand from instrumental variable analysis, with many parallel insights. Any representation $H(R)$ of the remotely sensed variable R is valid, as long as the representation is relevant to outcome variation: $\mathbb{E}[H(R)\Delta^o]\neq 0$. A weak remotely sensed variable is one where $\mathbb{E}[H(R)\Delta^o]\approx 0$. A testable implication of our framework is that two representations $H(R)$ and $H'(R)$ should give similar causal estimates. ▲

Corollary 1 shows that we may introduce a representation $H(X,R)$ of the remotely sensed variable R. The representation may be arbitrary, as long as it satisfies a rank condition: $H(X,R)$ must generate sufficient variation to distinguish each outcome category. The representation must predict outcome variation in the observational sample well, which may be viewed as an analogue of an instrument relevance condition.

By Corollary 1, many representations suffice for identification. This has two important consequences. First, naive choices of the representation may be inefficient. To find an efficient representation, we therefore connect the literature on "representation learning" (e.g., Johannemann et al., 2019; Vafa et al., 2025) to classical results for conditional moment equalities (Chamberlain, 1987; Newey, 1993) in Section 4. Second, there are testable implications of our assumptions through overidentifying restrictions.

Remark 1 (Overidentifying restrictions and a specification test)**.** Our framework generates testable overidentifying restrictions. With incomplete observational cases, Corollary 1 shows that any representation $H(X,R)$ identifies the treatment effect as long as $\mathbb{E}[H(X,R)\Delta^o(X)^\top \mid X]$ has full column rank. If Assumptions 2 and 3(ii) hold, then different representations should yield the same treatment effect. Disagreement across representations constitutes evidence against our identifying assumptions. This can be formalized as a test of overidentifying restrictions using a J-statistic (e.g., Sargan, 1958; Hansen, 1982). An analogous test can be constructed with complete observational cases, using Corollary 2 below.

Remark 2 (Weak remotely sensed variables)**.** Just as instrumental variable methods require a sufficiently strong first stage, our method requires that the chosen representation is a sufficiently strong predictor of outcome variation in the observational sample. In Example 3, when $\mathbb{E}[H(R)\Delta^o]$ is close to zero, the ratio $\mathbb{E}[H(R)\Delta^e]/\mathbb{E}[H(R)\Delta^o]$ is weakly identified and conventional inference can be unreliable. Researchers can check for weak remotely sensed variables by assessing whether $\mathbb{E}[H(R)\Delta^o]$ is bounded away from zero, using existing frameworks that test for weak instruments.

3.1.2 Complete Observational Cases: Allowing for Direct Effects

Our identification result also holds for the setting of Assumption 3(i), which requires complete observational cases yet allows for direct effects. Complete observational cases do not require randomization of treatment D in the observational sample nor transportability of treatment effects between the observational and experimental sample. Let $\mu(d,X) := (\mu_1(d,X),...\mu_{K-1}(d,X))^\top$.

Theorem 2 (Identification as conditional moment)**.** Under Assumptions 1, 2, and 3(i),

$$\mathbb{E}[\widetilde{\Delta}^e(d,X)-\widetilde{\Delta}^o(d,X)^\top\mu(d,X)\,|\,X,R]=0 \text{ almost surely, for any } d\in\{0,1\}, \text{ where}$$

$$\widetilde{\Delta}^e(d,X):=\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e\,|\,X)}-\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o\,|\,X)},$$
$$\widetilde{\Delta}^o_k(d,X):=\frac{1\{Y=y_k,D=d,S=o\}}{\Pr(Y=y_k,D=d,S=o\,|\,X)}-\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o\,|\,X)}.$$

Proof. See Appendix B.1.3. □

Corollary 2 (Identification as representation)**.** Suppose the conditions of Theorem 2 hold, and consider any measurable function $H_d:\mathcal{X}\times\mathcal{R}\to\mathbb{R}^J$ with $J\geq K-1$. Then,

$$\mathbb{E}[H_d(X,R)\widetilde{\Delta}^e(d,X)\,|\,X]=\mathbb{E}[H_d(X,R)\widetilde{\Delta}^o(d,X)^\top\,|\,X]\mu(d,X) \text{ for any } d\in\{0,1\}.$$

If $\mathbb{E}[H_d(X,R)\widetilde{\Delta}^o(d,X)^\top\,|\,X]$ has full column rank, then $\mu(d,X)$ and hence τ are identified.

3.2 Comparison with Alternative Approaches

We compare our identification strategy to alternative approaches and clarify how our identifying assumptions and data requirements differ. Applying some existing methods to post-outcome variables may be infeasible, biased, or inefficient. To simplify the discussion, we will focus on Example 3, with a binary outcome $Y\in\{0,1\}$ and no covariates $X=\varnothing$, although our discussion extends to the more general setting.

3.2.1 A Common Practice: Surrogacy Methods with Post-Outcome Variables

For exposition, suppose Assumption 3(ii) holds (with $D=0$ almost surely in the observational sample): we have incomplete observational cases and no direct effects.

An intuitive method to combine the experimental and observational samples proceeds in two steps: (i) train a predictor $m(R)$ of the outcome Y from the remotely sensed variable R in the observational sample; (ii) apply the predictor $m(R)$ to the experimental sample and calculate treatment effects on the predicted outcomes. This method appears in roughly half the papers we surveyed in general interest economics journals from 2015–2024 that use remotely sensed outcomes. While intuitive, it is generally biased for the causal parameter τ.

This two-step method implicitly targets the estimand

$$\widetilde{\tau}=\widetilde{\mu}(1)-\widetilde{\mu}(0)=\mathbb{E}[m(R)\,|\,D=1,S=e]-\mathbb{E}[m(R)\,|\,D=0,S=e],\quad m(R):=\mathbb{E}[Y\,|\,R,S=o].$$

The first step estimates the function $m(R)$. The second step evaluates and averages $m(R)$ over the treated and untreated groups in the experimental sample.

As a starting point, suppose the remotely sensed variable fails to predict the outcome. The two-step method would return a precise estimate of zero, even though the remotely sensed variable provides no information about treatment effects in this case. Formally, if $\mathbb{E}[Y \mid R, S=o] = \mathbb{E}[Y \mid S=o]$, then $m(R)$ is constant and $\widetilde{\tau}=0$ regardless of the true treatment effect.

More generally, even if the remotely sensed variable is predictive of the outcome, the implicit target $\widetilde{\tau}$ is biased away from the experimental treatment effect τ.

Proposition 1 (Bias of a common practice)**.** If Assumptions 1, 2, and 3(ii) hold in Example 3:

i. $\widetilde{\mu}(d) = \mu(d) + \mathbb{E}[Y\{w_d(R,Y)-1\} \mid D=d, S=e]$, where $w_d(r,y) := \frac{\Pr(Y=y|S=o)}{\Pr(Y=y|D=d,S=e)} \frac{f_R(r|D=d,S=e)}{f_R(r|S=o)}$;

ii. there exists $\kappa \in [0,1)$ such that $\widetilde{\tau} = \kappa\tau$, as long as R is an imperfect predictor of Y (i.e., $\mathrm{Var}(Y \mid R, S=o) > 0$ almost surely).

Proof. See Appendix B.1.4 for the proof and its generalization to discrete Y. □

Proposition 1(i) characterizes the implicit estimand of the common practice. Under Assumptions 2 and 3(ii), the conditional distribution of the remotely sensed variable $f_R(r \mid Y, S=o)$ is stable across the experimental and observational samples. However, the common practice attempts to transport the reverse relationship—the predictions $m(R) = \mathbb{E}[Y \mid R, S=o]$—into the experimental sample. By Bayes' rule, this reversal induces bias due to differences in the marginal distributions of outcomes between the two samples.

Proposition 1(ii) further shows that the common practice has attenuation bias: it is biased towards zero, whenever the remotely sensed variable is an imperfect predictor of the outcome. Empirical papers using this method may systematically under-report the impacts of environmental and anti-poverty interventions. The attenuation factor κ reflects the strength of the correlation between the remotely sensed variable and the outcome.[7]

[7] Two related results appear in the literature. Pelletier et al. (2026) show that one component of the bias from using predicted outcomes in difference-in-differences is attenuation bias due to variance compression, though their remaining bias terms are application-specific and generally unsigned. Crossley et al. (2022) show that, in a linear model, imputing a dependent variable across two samples induces a Berkson measurement error that attenuates the regression coefficient. Proposition 1 fully characterizes the estimand and bias of the two-step method when the remotely sensed variable is post-outcome and without assuming a linear model.

In the presence of covariates, each conditional average treatment effect is attenuated, implying arbitrary sign flipping for the average treatment effect.[8] The two-step method transports outcome predictions within each covariate cell. However, under our identifying assumptions, the stable object is the sensing mechanism within each covariate cell.

We generalize Proposition 1 for several empirically relevant settings. Appendix C.1.1 provides an analogous result with complete observational cases (allowing for direct effects) under Assumption 3(i). Appendix C.1.2 shows that the attenuation bias arises even if units are randomly allocated into the experimental and observational samples.

Remark 3 (Surrogacy and negative control methods)**.** Proposition 1 provides a direct comparison to the surrogacy model. The implicit target $\widetilde{\tau}$ of the two-step method is the surrogate formula (Prentice, 1989; Athey et al., 2025), so the two-step method recovers the causal parameter τ under the standard surrogacy assumptions $(D,S) \perp\!\!\!\perp Y \mid R$. These surrogacy assumptions state that R fully mediates the effect of the treatment on the outcome; that is, the surrogate R is *pre-outcome*. By contrast, Assumptions 2 and 3 imply the opposite: the outcome mediates the effect of the treatment on the remotely sensed variable, either partially (Assumption 3(i)) or fully (Assumption 3(ii)); the remotely sensed variable R is *post-outcome*. Appendix Figure A.2 illustrates the difference via a causal graph.

The surrogacy method, and related methods that lead to the estimand $\widetilde{\tau}$, have been extended to include negative controls (e.g., Ghassami et al., 2022; Imbens et al., 2025). The surrogacy method and its negative control generalizations would all incur the bias formalized in Proposition 1 when applied to post-outcome variables.

Similar to our method, the single negative control method (Park et al., 2024) uses one auxiliary variable. Our method has two key differences. Unlike the single negative control method, we allow the treatment D to affect both the outcome Y and the auxiliary variable R when Assumption 3(i) holds. When no complete cases are available (i.e., the setting covered by Assumption 3(ii)), the single negative control method becomes infeasible.

Remark 4 (Imputation methods)**.** A recent approach to inference with remotely sensed variables is multiple imputation (Rubin, 1987; Proctor et al., 2023), which fits a model for

[8] Denote the conditional average treatment effect by $\tau(x) = \mathbb{E}\{Y(1) - Y(0) | S = e, X = x\}$ and the density by $p(x) = f_X(x|S = e)$. By the proof of Proposition 1, $\widetilde{\tau}(x) = \kappa(x)\tau(x)$ for $\kappa(x) = \frac{\mathrm{Var}\{m(R)|S=o,X=x\}}{\mathrm{Var}(Y|S=o,X=x)} \in [0,1)$. There exist $\tau(x)$ and $\kappa(x)$ such that the signs of $\tau = \int \tau(x)p(x)\mathrm{d}x$ and $\widetilde{\tau} = \int \kappa(x)\tau(x)p(x)\mathrm{d}x$ are reversed. For example, with a binary covariate X, if $\tau(1) > 0$ and $\tau(2) < 0$ with $\tau = \tau(1)p(1) + \tau(2)p(2) > 0$, then there exist small $\kappa(1)$ and large $\kappa(2)$ such that $\widetilde{\tau} = p(1)\kappa(1)\tau(1) + p(2)\kappa(2)\tau(2) < 0$.

the missing outcome, draws repeated imputations of the missing outcome, and combines the resulting estimates. Even with a correctly specified imputation model, multiple imputation targets the same estimand as the two-step method. For Example 3 with incomplete observational cases, the imputation model is $\Pr(Y=1 \mid R,S=o)$, so the multiple imputation estimand coincides with the two-step estimand $\widetilde{\tau}$. With complete observational cases, the imputation model additionally conditions on the treatment D, and the multiple imputation estimand coincides with the two-step estimand discussed in Appendix C.1.1. In our framework, bias can arise because multiple imputation implicitly transports the outcome model across samples, whereas identification instead relies on transporting the stable sensing mechanism.

3.2.2 Prediction-Powered Inference

We compare our approach with prediction-powered inference (PPI) (Angelopoulos et al., 2023; Pelletier et al., 2026). The PPI method takes as input a pretrained machine learning predictor of the outcome variable Y. Let $m_1(R)$ and $m_0(R)$ denote arbitrary predictors for treated and untreated units, respectively.

Like the common practice described above, PPI computes a difference of means using predicted outcomes for the observations of (D,R). It also adds bias correction terms, called "rectifiers," using complete observations of (D,Y,R). Within Example 3, the PPI estimand is

$$\widetilde{\tau}^{\text{PPI}} = \mathbb{E}[m_1(R) \mid D=1,S=e] - \mathbb{E}[m_0(R) \mid D=0,S=e] + \mathbb{E}[Y - m_1(R) \mid D=1,S=o] - \mathbb{E}[Y - m_0(R) \mid D=0,S=o].$$

First, suppose that we have incomplete observational cases. By definition, we have no joint observations of (D,Y,R) with variation in D. Consequently, it is not possible to observe Y for some treated units and also some untreated units. At least one PPI rectifier (i.e., $\mathbb{E}[Y - m_d(R) \mid D=d,S=o]$) cannot be computed, and so PPI cannot be implemented.

Second, suppose that we have complete observational cases, and we use the observational sample to construct the rectifiers. In this case, the relevant question is whether PPI rectifiers estimated in $S=o$ can be transported to $S=e$. PPI may be biased in several plausible scenarios: (i) the treatment is confounded in the observational sample; (ii) the experimental and observational samples have different treatment effects; (iii) the experimental and observational samples have different "baselines" (i.e., different untreated outcome means).[9] There may be bias even when complete observational cases are available and D has no direct effect on R.

[9] PPI is unbiased if and only if $\mathbb{E}\{Y - m_d(R) | D=d,S=e\} = \mathbb{E}\{Y - m_d(R) | D=d,S=o\}$ for $d \in \{0,1\}$. Each scenario we describe is a way to violate this condition. See Section 5.1 for further discussion.

Proposition 2 (Bias of transporting PPI rectifiers)**.** Suppose Assumptions 1, 2, 3(i) and 3(ii) hold in Example 3. Let $m_1(R)$ and $m_0(R)$ be any bounded and measurable functions. Let $\widetilde{\tau}^o := \mathbb{E}[Y \mid D=1, S=o] - \mathbb{E}[Y \mid D=0, S=o]$ denote the difference in means in the observational sample. Then

1. $\widetilde{\tau}^{\text{PPI}} = \tau + (1-\kappa_1)(\widetilde{\tau}^o - \tau) + (\kappa_1 - \kappa_0)\delta$ where $\kappa_d = \mathbb{E}[m_d(R) \mid Y=1] - \mathbb{E}[m_d(R) \mid Y=0]$ and $\delta = \mathbb{E}[Y \mid D=0, S=e] - \mathbb{E}[Y \mid D=0, S=o]$;

2. for any predictors $m_1(R)$, $m_0(R)$ and any values $\widetilde{\tau}^o$ and τ (with possibly $\widetilde{\tau}^o \neq \tau$), there exists a data-generating process with a relevant R (i.e., $R \not\perp\!\!\!\perp Y \mid D, S$) such that $\widetilde{\tau}^{\text{PPI}} = \widetilde{\tau}^o$.

Proof. See Appendix B.1.5. □

Proposition 2(i) characterizes the PPI estimand as a mixture of the desired estimand τ, the difference in means in the observational sample $\widetilde{\tau}^o$, and the difference in baselines across samples δ, with weights depending on the predictors $m_1(R)$ and $m_0(R)$. PPI is therefore biased whenever the difference in means in the observational sample differs from the treatment effect in the experimental sample (either due to heterogeneity of treatment effects or because of confounding) or whenever there are differences in baselines across samples. As a consequence, Proposition 2(ii) further shows that PPI can have arbitrary bias when it uses the observational sample: it may be targeting $\widetilde{\tau}^o > \tau$, $\widetilde{\tau}^o < \tau$, or something else entirely.

By contrast, our method relies on stability of the sensing mechanism. This distinction matters in applications like Example 1, where the observational sample is collected on different units and at different times than the experimental sample. Concretely, Jayachandran et al. (2017) use ground-based forest measurements at locations distinct from the experimental units and measured at different dates, while Jack et al. (2025) use randomized spot checks that cannot be synchronized with the narrow post-harvest/pre-tilling window. These mismatches make it plausible that baselines differ ($\delta \neq 0$), the treated–untreated contrasts differ across samples ($\widetilde{\tau}^o \neq \tau$), or both—precisely the conditions under which Proposition 2(i) shows PPI can be biased.

3.2.3 Measurement Error and Data Fusion

A large literature studies identification using noisy measurements of a latent variable as well as auxiliary data. Works such as Chen et al. (2005, 2008), Graham et al. (2016), Ridder and Moffitt (2007), Fan et al. (2014), and D'Haultfœuille et al. (2024); D'Haultfœuille et al. (2025) establish identification by assuming stability of the distribution of the latent

variable given the noisy measurement. In our notation, they require outcome stability: $f_Y(y \mid R,D,S=e) = f_Y(y \mid R,D,S=o)$.

We study a different regime, where the noisy measurement is post-outcome: the remotely sensed variable is caused by the outcome of interest. We instead assume that the sensing mechanism is stable: $f_R(r \mid Y,D,S=e) = f_R(r \mid Y,D,S=o)$. Meanwhile, we explicitly allow the treatment effects, treatment mechanism, and outcome mechanism to differ across samples. This distinction—in what information can be transported across samples—leads to a different identifying formula.

Finally, a closer point of contact is work by Hu and Schennach (2008); Hu (2008). Those authors also impose invariance of the noisy measurement distribution conditional on the latent variable. However, their data environment is different: the true variable is never directly observed, so identification requires a normalization that pins down the latent scale, e.g. conditionally mean zero measurement error (Hu and Schennach, 2008, Assumption 5). In our setting, the economic outcome is observed in the observational sample, eliminating the latent-scale normalization problem and yielding a different identification strategy based on conditional moment restrictions.

4 Estimation and Inference

As our secondary contribution, we show how our identification result guides the choice of representation for the remotely sensed variable, enabling precise inference on the treatment effect. We also prove that our confidence intervals are robust to mis-specification. We derive valid $n^{-1/2}$ inference without rate conditions or complexity restrictions on the researcher's remotely sensed variable-based predictions. In particular, we justify the use of complex deep learning algorithms that may be misspecified.

For exposition, throughout this section we focus on incomplete observational cases (Assumption 3(ii)) and no covariates ($X = \varnothing$). Estimation and inference for complete observational cases are similar, using the identification results in Section 3.1.2; see Appendix C.7.2 for details. We generalize our estimation algorithm to include covariates in Section 5.7.

4.1 Choice of the Representation

Consider the binary outcome setting (Example 3), where the treatment effect is identified as $\tau = \frac{\mathbb{E}\{H(R)\Delta^e\}}{\mathbb{E}\{H(R)\Delta^o\}}$ for any representation $H(R)$ satisfying $\mathbb{E}\{H(R)\Delta^o\} \neq 0$. Different choices

of representation have different efficiency properties. We now ask, which representation is a reasonable choice in practice?

To answer this question, we view τ as the coefficient in the regression model

$$\Delta^e = \tau \Delta^o + \epsilon, \quad \mathbb{E}(\epsilon \mid R) = 0, \tag{1}$$

where the residual $\epsilon := \Delta^e - \tau \Delta^o$ is mean zero conditional on R by Theorem 1. Results in Chamberlain (1987) and Newey (1993) imply that the semiparametrically efficient representation within a class of models satisfying (1) is $H^*(R) = \frac{\mathbb{E}(\Delta^o \mid R)}{\sigma^2(\tau,R)}$, where $\sigma^2(\tau,R) := \mathbb{E}\{(\Delta^e - \tau\Delta^o)^2 \mid R\}$. This "optimal instrument" $H^*(R)$ is a conditional expectation divided by a conditional variance.

The discrete case is similar. Applying results in Chamberlain (1987) and Newey (1993), the semiparametrically efficient representation for models satisfying the conditional moment restriction in Theorem 1 is

$$H^*(R) = \frac{\mathbb{E}(\Delta^o \mid R)}{\sigma^2(\theta,R)}, \quad \sigma^2(\theta,R) := \mathbb{E}[\{\Delta^e - (\Delta^o)^\top \theta\}^2 \mid R],$$

where now $H^*(R) \in \mathbb{R}^{K-1}$, $\mathbb{E}(\Delta^o \mid R) \in \mathbb{R}^{K-1}$, and $\sigma^2(\theta,R) \in \mathbb{R}$.[10]

This formula reveals that improving the efficiency of treatment effect estimates requires three types of predictions: (i) predictions of each outcome category from the remotely sensed variable, appearing in the numerator $\mathbb{E}(\Delta^o \mid R)$; (ii) prediction of the treatment from the remotely sensed variable, appearing in the denominator $\sigma^2(\theta,R)$; and (iii) prediction of the sample indicator from the remotely sensed variable, appearing in both the numerator and the denominator. That is, improving efficiency requires predicting not only the outcome Y (using the observational sample) but also the treatment D (using the experimental sample) and the sample indicator S (using all data).

In Section 4.2, we propose an algorithm based on the choice $H^*(R)$. In Section 4.3, we prove robust inference guarantees for a family of algorithms, allowing a family of representations, including the optimal choice $H^*(R)$ and the simple choice $\Pr(Y = y_k \mid S = o, R)$.

4.2 Estimation and Inference with Learned Representations

Our estimation algorithm follows three steps: divide the sample into TRAIN and TEST folds; learn the representation on TRAIN; and apply the learned representation to estimate the treatment effect on TEST. Algorithm 1 provides an overview of our estimation algorithm

[10] Although R can be high-dimensional, $H^*(R)$ is low dimensional because K is fixed, so results from Chamberlain (1987) directly apply.

Algorithm 1 Estimation with Incomplete Observational Cases (overview)

Require: Data $\{S_i, 1\{S_i=e\}D_i, 1\{S_i=o\}Y_i, R_i\}_{i=1}^n$ with $Y \in \mathcal{Y} = \{y_1, \ldots, y_K\}$.

1: Randomly split indices into a TRAIN fold and a TEST fold. Define the empirical mean operators $\mathbb{E}_{\text{TRAIN}}(\cdot) = \frac{1}{|\text{TRAIN}|}\sum_{i\in\text{TRAIN}}(\cdot)$ and $\mathbb{E}_{\text{TEST}}(\cdot) = \frac{1}{|\text{TEST}|}\sum_{i\in\text{TEST}}(\cdot)$.

Step 1: Learn a representation on TRAIN.

2: Using TRAIN, estimate $\widehat{\Pr}(Y=y_k \mid S=o, R)$, $\widehat{\Pr}(D=1 \mid S=e, R)$, and $\widehat{\Pr}(S=e \mid R)$.

3: Combine these estimates into a learned representation $\widehat{H}(R) \in \mathbb{R}^{K-1}$ that approximates the efficient representation $H^*(R)$ (see Algorithm E.1 for the explicit construction).

Step 2: Estimate $\widehat{\theta}$ on TEST.

4: On TEST, compute the sample analogues $(\widehat{\Delta}^e, \widehat{\Delta}^o)$ of the treatment and outcome variation.

5: Compute $\widehat{\theta} = \left[\mathbb{E}_{\text{TEST}}\left\{\widehat{H}(R)(\widehat{\Delta}^o)^\top\right\}\right]^{-1} \mathbb{E}_{\text{TEST}}\left\{\widehat{H}(R)\widehat{\Delta}^e\right\}$.

Step 3: Compute the ATE and standard error.

6: Compute $\widehat{\tau} = \sum_{j=1}^{K-1}(y_j - y_K)\widehat{\theta}_j$.

7: Bootstrap the standard error of $\widehat{\tau}$ on TEST, holding $\widehat{H}$ fixed.

8: If **cross-fitting**: Reverse the role of TRAIN and TEST, then re-estimate the ATE and its standard error. Average the ATEs and standard errors across the two iterations.

9: **Return:** $\widehat{\tau}$ and its confidence interval.

using sample splitting with TRAIN and TEST folds, as well as two-fold cross-fitting, though our proof in Appendix B.2 allows for cross-fitting with any fixed number of folds. Algorithm E.1 provides a detailed implementation of the estimation algorithm with covariates. The algorithm for complete observational cases is similar. Sample splitting and cross-fitting may be eliminated under complexity restrictions that tolerate simple machine learning procedures. See, for example, Chernozhukov et al. (2025) for a recent summary.

Within Algorithm 1, we propose a learned representation $\widehat{H}(R)$ based on $H^*(R)$ above. However, our guarantees below allow any learned representation $\widehat{H}(R)$ that has some limit $\widetilde{H}(R)$. For example, a researcher may prefer to use a pre-trained representation $\widehat{\Pr}(Y=y_k \mid S=o, R)$, sacrificing some precision to reduce computation.

4.3 Robustness to Misspecification

Remotely sensed variables take many forms—from unstructured data to pretrained embeddings to outputs of complex machine learning predictors—and our identification results accommodate all of these cases. We therefore analyze our estimator under weak regularity conditions.

For example, the prediction may be constructed from a complex machine learning algorithm, such as a deep convolutional neural network applied to satellite images or even a pre-trained model. In such settings, rates of convergence are often unknown. Even carefully crafted architectures positing a generative model would typically be misspecified.

For this reason, we do not require convergence rates for the machine learning algorithms used to construct predictions. Under cross-fitting, we only require that the predictions, and hence the representation estimator, have some probability limit; they may be misspecified. Under sample splitting, we can further relax this weak condition; see Remark 5 below.

Assumption 4 (Limit)**.** The learned representation has some mean-square limit $\widetilde{H}(R)$: $\mathbb{E}_R\left\{\|\widehat{H}(R)-\widetilde{H}(R)\|^2\right\}=o_p(1)$, where $\mathbb{E}\{\|\widetilde{H}(R)\|^2\}$ is finite, and possibly $\widetilde{H}(R)\neq H^*(R)$. This limit is correlated with outcome variation: $\mathbb{E}\{\widetilde{H}(R)(\Delta^o)^\top\}$ is nonsingular, and its smallest singular value is bounded away from zero.

While the overall structure of Algorithm 1 is familiar (Angrist et al., 1999; Chernozhukov et al., 2018), valid $n^{-1/2}$-inference on τ requires no rate conditions on the learned representation. The moment restriction in (1) is infinite-order Neyman orthogonal (Mackey et al., 2018; Chen et al., 2020), and therefore Assumption 4 requires no complexity restriction (Chernozhukov et al., 2018, 2023) and no rate conditions (Chamberlain, 1987; Newey, 1993).

Concretely, within Algorithm 1, all three predictors $\widehat{\Pr}(Y=y_k\,|\,S=o,R)$, $\widehat{\Pr}(D=1\,|\,S=e,R)$, and $\widehat{\Pr}(S=e\,|\,R)$ may be misspecified. Any other choice of learned representation may be misspecified. Nonetheless, we have valid $n^{-1/2}$ inference, which is a strong form of robustness.

In the following statements, we refer to $\Pr(D=d,S=e)$ for $d\in\{0,1\}$ and $\Pr(Y=y_k,S=o)$ for $y_k\in\{y_1,\ldots,y_K\}$ as the marginal probabilities.

Proposition 3 (Inference with known marginal probabilities)**.** Consider the cross-fitted estimator in Algorithm 1. Suppose Theorem 1's conditions and Assumption 4 hold. Suppose the marginal probabilities are known and bounded away from zero. Then, for $\lambda=(y_1-y_K,\ldots,y_{K-1}-y_K)^\top$, we have

$$\sqrt{n}(\widehat{\tau}-\tau)\rightsquigarrow\mathcal{N}(0,\lambda^\top ABA^\top\lambda),\quad A=[\mathbb{E}\{\widetilde{H}(R)(\Delta^o)^\top\}]^{-1},\quad B=\mathbb{E}[\{\Delta^e-(\Delta^o)^\top\theta\}^2\widetilde{H}(R)\widetilde{H}(R)^\top].$$

Moreover, if $\widetilde{H}(R)=H^*(R)$, then $\widehat{\tau}=\lambda^\top\widehat{\theta}$ is semiparametrically efficient for $\tau:=\lambda^\top\theta$ satisfying the conditional moment restriction $\mathbb{E}\{\Delta^e-(\Delta^o)^\top\theta\,|\,R\}=0$, with known marginal probabilities.

Proof. See Appendix B.2.1. □

Proposition 4 (Inference with estimated marginal probabilities)**.** Consider the cross-fitted estimator in Algorithm 1. Suppose Theorem 1's conditions and Assumption 4 hold. If the marginal probabilities and their estimators are bounded away from zero, then $\sqrt{n}(\widehat{\tau}-\tau) \rightsquigarrow \mathcal{N}(0, \lambda^\top A V A^\top \lambda)$ for V defined in Appendix Lemma B.9.

Proof. See Appendix B.2.2. □

When marginal probabilities are known, the asymptotic variance is standard. The asymptotic variance in Proposition 4 differs from the one in Proposition 3 because estimation of the marginal probabilities introduces an additional estimation error.

Standard errors may be calculated by either a bootstrap estimator or an analytic estimator. The former is described in Algorithm 1, and may be easier to code. The latter is described in Appendix Lemma B.9, and involves less computation.

Theorem 2 allows for direct effects of the treatment on the remotely sensed variable, provided there are complete observational cases. Since we have already derived the conditional moment equations, estimation and inference proceed in a similar fashion.

Remark 5 (Inference without a probability limit for $\widehat{H}$)**.** In the absence of Assumption 4, valid conditional inference is still possible by using sample splitting, with some caveats. With sample splitting, $\widehat{H}$ is a deterministic function conditional on TRAIN. Therefore, Algorithm 1 with sample splitting is asymptotically normal conditional on TRAIN: $|\text{TEST}|^{1/2}\sigma_{\widehat{H}}^{-1}(\widehat{\tau}-\tau)|\text{TRAIN} \rightsquigarrow \mathcal{N}(0,1)$, provided that $\widehat{H}(R)$ is uniformly bounded, $\mathbb{E}\{\widehat{H}(R)(\Delta^o)^\top|\text{TRAIN}\}$ has its smallest singular value bounded away from zero with probability approaching one, and the conditional asymptotic variance $\sigma_{\widehat{H}}^2$ is nondegenerate with probability approaching one. There are two caveats: (i) the convergence rate is $|\text{TEST}|^{-1/2}$ rather than $n^{-1/2}$; (ii) the variance $\sigma_{\widehat{H}}^2$ is random rather than nonrandom because it depends on the realization of TRAIN.

5 Extensions and Generalizations

Our framework for identification and estimation extends in several directions that are important for practice: (i) a "validation" sample with randomized treatments and observed outcomes; (ii) a quasi-experimental sample rather than an experimental sample; (iii) some spillovers across units; (iv) variables collected at different times; (v) sensitivity analyses under relaxations of our identifying assumptions; (vi) continuous outcomes; (vii) low-dimensional covariates. We provide a brief overview of each extension in this section, deferring formal details to the appendix.

5.1 A Validation Sample

In empirical applications, we may also observe outcomes for a (possibly small) subset of experimental units, which we call a "validation" sample. This third sample fits naturally into our framework. Define an extended sampling indicator $\widetilde{S} \in \{e,o,v\}$. As before, $\widetilde{S}=o$ denotes an observational unit and $\widetilde{S}=e$ denotes an experimental unit with missing outcome. Now, $\widetilde{S}=v$ denotes an experimental unit for which the outcome is observed.

With this notation, our identification and estimation results hold by replacing $S=e$ with $\widetilde{S} \in \{e,v\}$, and $S=o$ with $\widetilde{S} \in \{o,v\}$, in the appropriate expressions. Intuitively, units with $\widetilde{S}=v$ can be reused: because they contain (X,D,Y,R), they contribute to both treatment variation and outcome variation, and therefore inform both sides of our moment conditions.

This extension does not require assumptions about which experimental units have observed outcomes: the subset with $\widetilde{S}=v$ need not be random (e.g., it may be selected based on logistics or targeting). Our key requirement is that the remotely sensed variable is stable across all three samples: $\widetilde{S} \perp\!\!\!\perp R \,|\, X,D,Y$.[11] Only the sensing mechanism must be invariant.

Remark 6 below explains how a particular variation of PPI can be unbiased at the expense of (i) placing an extra assumption on the validation sample, and (ii) discarding the observational sample.[12] Simulations in Section 6 show that, in the presence of a validation sample, our method is more efficient.

5.2 A Quasi-Experimental Sample

Many empirical applications with remotely sensed variables exploit quasi-experimental variation (e.g., Burgess et al., 2012; Ratledge et al., 2022; Assuncao et al., 2023; Pelletier et al., 2026, among many others). Appendix C.2 extends our framework to these settings by modifying our assumption for the $S=e$ sample: we replace experimental unconfoundedness (Assumption 1) with instrumental variable or difference-in-differences assumptions. Under stability of the remotely sensed variable (Assumption 2), we show how researchers can combine a quasi-experimental sample with an observational sample to identify familiar causal parameters: the local average treatment effect (LATE) using instruments, or the average treatment effect on the treated (ATT) using difference-in-differences. Overall, our main identification technique

[11] In fact, by the proof of Lemma 1, it suffices that $f_R(r \,|\, \widetilde{S} \in \{e,v\},X,D,Y) = f_R(r \,|\, \widetilde{S} \in \{o,v\},X,D,Y)$.

[12] Only the validation sample, and not the observational sample, must be used for the rectifiers. This PPI variation is unbiased if and only if $\mathbb{E}\{Y - m_d(R) \,|\, D=d,\widetilde{S}=e\} = \mathbb{E}\{Y - m_d(R) \,|\, D=d,\widetilde{S}=v\}$ for $d \in \{0,1\}$. By contrast, our method is unbiased and also leverages the observational sample, so it is more efficient.

seamlessly integrates into identification arguments for quasi-experiments.

5.3 Spillovers

When evaluating environmental conservation and anti-poverty interventions, empirical researchers often discuss spillovers: how treatment of one unit may affect the outcome of another unit. In Appendix C.3, we extend our framework to allow for some simple types of spillovers. After expanding the potential outcome notation to reflect spillovers, we define the average direct treatment effect and the average global treatment effect. We identify the average direct treatment effect using stability (Assumption 2). Then, under the assumption that spillovers are either (i) within but not across higher levels of geography, or (ii) within a known geographic radius, we identify the average global treatment effect.

5.4 Time Variation

In practice, data are indexed by time. The remotely sensed variable and the outcome may be measured at different times. Moreover, the treatment variable may be defined as early adoption. In Appendix C.4, we confirm that our method accommodates such settings, and we clarify the interpretation of our reported estimate.

5.5 Sensitivity Analysis

Our identification result relies on two conditions beyond experimental unconfoundedness: stability of the remotely sensed variable (Assumption 2) and, with incomplete observational cases, no direct effects of the treatment on the remotely sensed variable (Assumption 3(ii)). While we provide evidence supporting these assumptions in our empirical applications (Sections 6 and 7), researchers may wish to assess how sensitive their conclusions are to departures from these assumptions. Appendix C.5 develops a partial identification framework for this purpose, which we summarize below.

First, we relax Assumption 3(ii): we maintain stability but allow the treatment to have a small direct effect on the remotely sensed variable. Suppose we are willing to bound the size of the direct effect and can calculate the covariance between the outcome variation and the representation $H(X,R)$. We show that the identified set for the average treatment effect is an interval whose size is increasing in the direct effect bound and decreasing in the calculated correlation.

Next, we relax Assumptions 2 and 3(ii) simultaneously: the treatment may have a direct effect on the remotely sensed variable, and the remotely sensed variable's conditional distribution may vary across samples. We provide an outer bound on the identified set for the average treatment effect. This outer bound recovers the identified set described above when Assumption 2 is imposed again.

We may report these identified sets for a range of plausible violation magnitudes to assess how robust causal findings are to departures from the identifying assumptions. For example, a researcher may report a "breakdown" analysis illustrating the smallest direct effect or stability violation required to overturn the study's conclusions (e.g., Masten and Poirier, 2020).

5.6 Continuous Outcomes

Appendix C.6 extends our results to settings with continuous outcomes. While the identification remains essentially the same, estimation is more complex because the statistical inverse problem of recovering the treatment effect becomes more difficult.

We propose a practical method: discretize the continuous outcome into finitely many bins and appeal to our results for discrete outcomes, while keeping track of the discretization error. Under a mild smoothness condition (that the conditional density of the remotely sensed variable $f_R(r \mid S=o,X,D,Y)$ is uniformly continuous in Y), we control the discretization error in Corollary C.5. This yields an approximate version of Theorem 1, with a discretization error that we bound explicitly. We can therefore apply our estimation algorithm to the discretized outcome and adjust inference for a non-vanishing discretization error, for example, by using bias-aware confidence intervals.

We also discuss the nonparametric generalization. Allowing the bin width to vanish amounts to kernel smoothing, which eliminates the discretization error at the cost of slower convergence rates and more subtle regularity conditions. We clarify how our relevance condition $\mathbb{E}\{H(R)\Delta^o\} \neq 0$ generalizes to a completeness condition when outcomes are continuous. Overall, continuous outcomes require stronger conditions for identification and estimation.

5.7 Pre-treatment Covariates

Section 3 derived identification with unrestricted, pre-treatment covariates. For ease of exposition, Section 4 derived inference with no covariates. We now describe how inference extends to low dimensional covariates. Future work may study inference with high dimensional covariates.

When covariates are discrete with finite support, inference is straightforward: apply Algorithm 1 within each covariate stratum, then average over covariate strata. Appendix C.7 presents this algorithm and shows that our asymptotic analysis suitably generalizes.

When covariates are low dimensional and continuous, inference requires standard nonparametric techniques. A researcher must estimate the conditional probabilities $\Pr(D=d,S=e\,|\,X)$ and $\Pr(Y=y,S=o\,|\,X)$ as smooth functions of X using kernels or series, then the conditional average treatment effects via analogous methods, before averaging over X; see Appendix C.7.3.

6 Calibrated Simulations

We conduct a semi-synthetic exercise calibrated to two real-world settings.

First, we study forest cover measurements from Hansen et al. (2013) for Uganda. These widely used measurements are derived from Landsat satellite imagery at 30-meter resolution, defining forest cover as the fraction of a grid cell containing vegetation taller than 5 meters. We define a binary outcome $Y \in \{0,1\}$ indicating whether at least 80% of a grid cell is forested, consistent with conservation policies that mandate forest cover retention.[13]

We define the experimental sample as grid cells within the area covered by the payments for ecosystem services experiment conducted by Jayachandran et al. (2017). The observational sample consists of grid cells in surrounding geographic bands. Appendix Figure F.1 provides a map, and Appendix F.1 provides additional details.

To illustrate how one might use the output of a pretrained predictor as the remotely sensed variable, we construct one for this setting: using grid cells from the rest of Uganda, we train a random forest to predict Y from MOSAIKS satellite embeddings (Rolf et al., 2021), which are low-cost pretrained embeddings of publicly available satellite images distinct from the Landsat images underlying the Hansen et al. (2013) data (see also Proctor et al., 2025). The random forest predictor achieves an area under the curve (AUC) of 0.967 on held-out grid cells. We define the remotely sensed variable R as the scalar output of this predictor.

Second, we use data from Muralidharan et al. (2016, 2023), who conducted a randomized experiment to evaluate the introduction of biometrically authenticated payment infrastructure (Smartcards) in Andhra Pradesh, India. We define a binary outcome $Y \in \{0,1\}$ indicating whether the village has only low- and middle-income households, using administrative data

[13] Examples include Brazil's Forest Code (Azevedo et al., 2017) as well as conservation programs that determine eligibility based on forest cover thresholds (e.g., Wunder et al., 2008).

from 2012–2013; see Appendix F.2 for details.

The experimental sample consists of treated, untreated, and buffer villages in the original evaluation. The observational sample consists of non-study villages; see Appendix Figure F.2.

We use this setting to illustrate our framework when the remotely sensed variable is high-dimensional. We link each village's geographic coordinates to nighttime luminosity measures (a vector in $\mathbb{R}^{50}$) (Asher et al., 2021) and MOSAIKS satellite embeddings (a vector in $\mathbb{R}^{4000}$) (Rolf et al., 2021). Both have been validated as predictors of poverty (Henderson et al., 2011; Jean et al., 2016; Huang et al., 2021). We define the remotely sensed variable R as their concatenation.

Several diagnostic tests in Appendix F.3 support the plausibility of our assumptions in both empirical settings.

6.1 Simulation Design

We first consider the most cost-effective data collection strategy: no experimental outcomes are collected, with incomplete observational cases (Assumption 3(ii)).

Let $\alpha_e := \Pr\{Y(0)=1 \mid S=e\}$ and $\alpha_o := \Pr\{Y(0)=1 \mid S=o\}$ denote the baseline outcome probabilities in the experimental and observational samples, respectively. The simulation design allows $\alpha_e \neq \alpha_o$, so that baseline outcomes may shift across samples. In the forest cover setting, α_e and α_o are calibrated from grid cells in the experimental and observational samples, respectively.[14] Similarly, in the Smartcards setting, α_e and α_o are calibrated from the experimental and observational samples, respectively. We vary the treatment effect across a range of values: $\tau \in \{0, 0.05, 0.10, 0.15, 0.20\}$. For each treatment effect value, we conduct 500 simulations.

Each simulation is generated with n_e experimental units and n_o observational units, according to a data-generating process satisfying both stability and no direct effects. For each experimental unit, we draw the potential outcomes $Y_i(0) \sim \text{Bernoulli}(\alpha_e)$, $Y_i(1) \sim \text{Bernoulli}(\alpha_e + \tau)$ and randomly assign treatment as $D_i \sim \text{Bernoulli}(0.5)$. The observed outcome is $Y_i = (1-D_i)Y_i(0) + D_i Y_i(1)$. We then draw the remotely sensed variable R_i from the empirical distribution of $R_i \mid Y_i$ in the real data, and delete Y_i, recording only (D_i, R_i). For each observational unit, we draw the outcome $Y_i \sim \text{Bernoulli}(\alpha_o)$. Then, we draw the remotely sensed variable R_i from the same empirical distribution, and record (Y_i, R_i). In the forest cover setting, we set $n_e = 1000$, roughly corresponding to the sample size in Jayachandran et al. (2017), and $n_o = 500$.

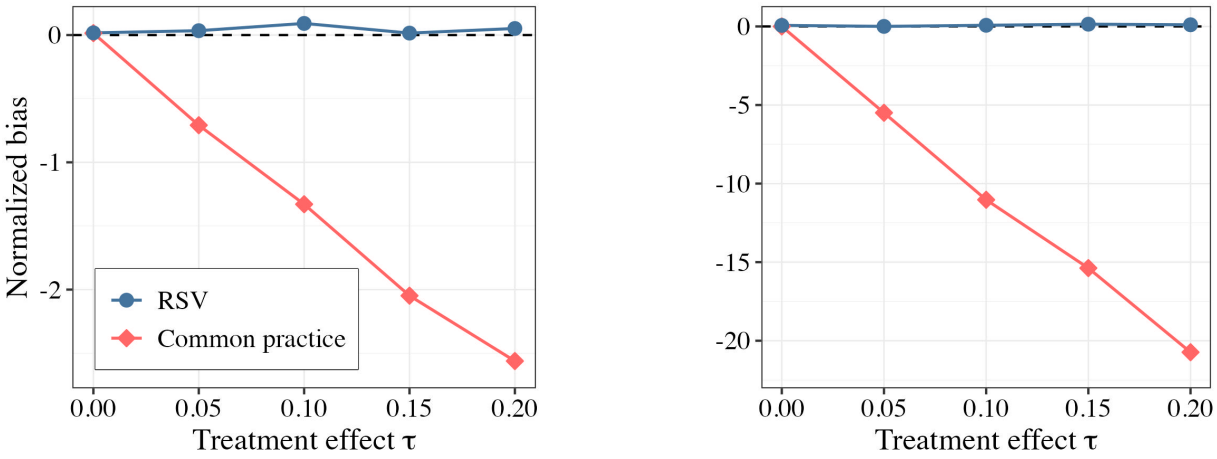
[14] In the main text, we report results calibrating α_o based on a two-kilometer band around the experimental sample. Appendix G.1 provides similar results in which α_o is calibrated using five- and ten-kilometer bands.

In the Smartcards setting, we set $n_e = 3000$ and $n_o = 1000$, and we also truncate the remotely sensed variable to $\mathbb{R}^{1050}$ for computational tractability. We report results for alternative sample sizes in Appendix G.1. Appendix G.2 provides additional details on the simulation design.

We compare two estimators: the commonly used two-step method (Section 3.2.1), which trains a predictor of Y from R in the observational sample and regresses predicted outcomes on D in the experimental sample; and our method (Algorithm 1). PPI methods (Section 3.2.2) are not feasible in this setting because there are no units with jointly observed (D,Y,R) with variation in D.

6.2 Results

First, we compare the quality of the point estimates by comparing their bias, across treatment effect values and empirical settings. Figure 2 visualizes the results. The commonly used two-step method exhibits large bias that is proportional to the treatment effect. This pattern confirms Proposition 1: this method targets $\kappa\tau$ for some $\kappa \in [0,1)$. Though the remotely sensed variables are good predictors of the outcome in both settings, accurate prediction alone does not ensure valid causal inference when R is post-outcome. By contrast, our method exhibits negligible bias across all treatment effect values and empirical settings.

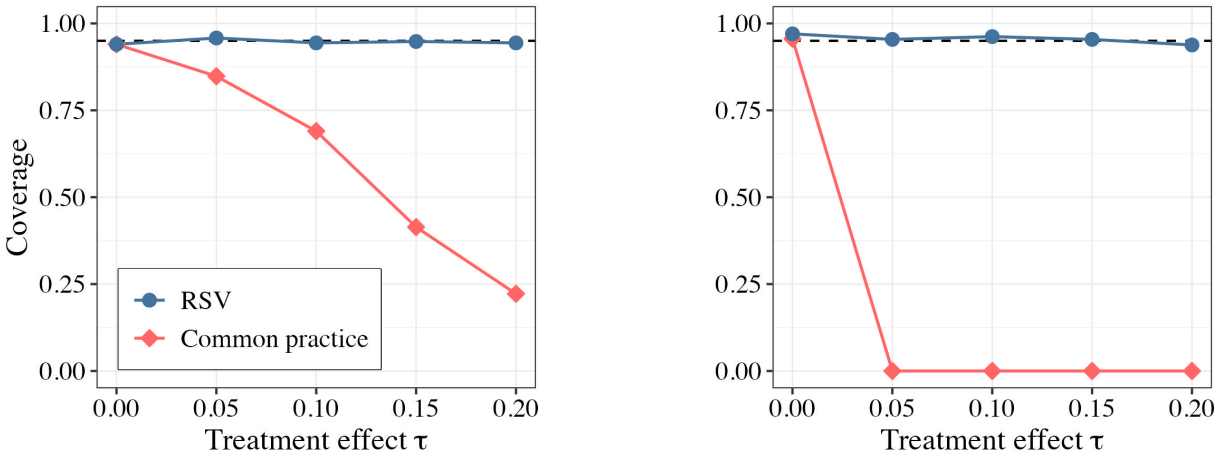


(a) Forest Cover in Uganda.

(b) Smartcards Experiment in India.

Figure 2: Our estimator outperforms the two-step estimator in terms of normalized average bias, defined as the average bias of the estimator divided by its standard deviation across simulations. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

Next, we compare the quality of the confidence intervals by comparing their coverage, across treatment effect values and empirical settings. Figure 3 visualizes the results. The 95%



(a) Forest Cover in Uganda.

(b) Smartcards Experiment in India.

Figure 3: Our confidence intervals achieve approximately nominal coverage across simulations. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

confidence intervals based on the commonly used method can undercover significantly—as low as 25% coverage in the forest cover setting and 0% coverage in the Smartcards setting, when the treatment effects are nonzero. By contrast, 95% confidence intervals based on our method achieve close to nominal coverage across all treatment effect values and empirical settings, confirming that the asymptotic theory in Section 4 provides reliable finite-sample guidance.

Remark 6 (Comparison with prediction-powered inference (PPI))**.** In the simulations above, PPI methods are infeasible because no experimental outcomes are observed. Appendix G.2 extends the simulations to include a small, randomly selected validation sample (Section 5.1), enabling comparison with PPI. Two findings emerge. First, a PPI method that uses both the validation sample and the observational sample in the rectifiers is biased whenever $\alpha_e \neq \alpha_o$, confirming Proposition 2 (see Appendix Figure G.5). Second, a PPI method that only uses the random validation sample in the rectifiers (disregarding the observational sample) is unbiased but inefficient; our estimator yields standard errors up to 40% smaller than PPI (see Appendix Figure G.6).[15] Therefore, our method can improve efficiency, even in the presence of a random validation sample. By exploiting the post-outcome structure of the remotely sensed variable, our method efficiently incorporates information from the observational sample.

[15]Only this variation of PPI is unbiased, since with a randomly selected validation sample $\mathbb{E}\{Y - m_d(R) \mid D=d, \widetilde{S}=e\} = \mathbb{E}\{Y - m_d(R) \mid D=d, \widetilde{S}=v\} \neq \mathbb{E}\{Y - m_d(R) \mid D=d, \widetilde{S}=o\}$. It is inefficient since it discards the sample $\widetilde{S}=o$. Even this variation of PPI would be biased if the validation sample were not randomly selected.

7 Applications with Remotely Sensed Outcomes

Next, we apply our framework to two real-world settings.

First, we revisit the Smartcards experiment of Muralidharan et al. (2016, 2023), who study the effect of biometrically authenticated payments on village-level poverty in Andhra Pradesh, India. Experimental outcomes were directly measured, so we can benchmark our method (using satellite images in the experimental sample) against an "oracle" difference-in-means (using direct outcome measurements in the experimental sample). Our method closely matches the benchmark, despite limited access to outcomes.

Second, we revisit the experiment of Jack et al. (2025), who study whether payments for ecosystem services (PES) contracts reduce crop burning by farmers in Punjab, India. Unlike the Smartcards experiment, no experimental outcomes are available, so we cannot benchmark against an unbiased difference-in-means. Instead, we compare our method (which uses the authors' pretrained classifier as a remotely sensed variable) to the commonly used two-step method (which uses the authors' pretrained classifier as a surrogate outcome). We find that the commonly used two-step method may underestimate the treatment effect by 47%.

7.1 Smartcards and Village-Level Poverty

Data. We use data from Muralidharan et al. (2016, 2023). The treatment $D \in \{0,1\}$ is the early introduction of Smartcards in 2010, and the outcome $Y \in \{0,1\}$ is a measure of village-level poverty constructed from administrative data in 2012–2013. To assess robustness across different poverty definitions, we consider three outcomes: (i) "consumption" indicates whether the village is in the bottom quartile of per capita consumption; (ii) "low income" indicates whether the village has only low-income households; and (iii) "middle income" indicates whether the village has only low- and middle-income households. While (i) captures average consumption, (ii) and (iii) describe the village income distribution using categories from Indian administrative data; see Appendix F.2 for details. The remotely sensed variable $R \in \mathbb{R}^{4050}$ concatenates the nighttime luminosity and pretrained satellite embeddings described in Section 6.[16]

We analyze this experiment as having incomplete observational cases (Assumption 3(ii)) and a non-randomly selected validation sample (Section 5.1). The experimental sample $\widetilde{S} = e$ consists of treated and untreated villages, for which we observe (D,R). The validation sample $\widetilde{S} = v$ consists of buffer villages, for which we observe (D,Y,R). The observational sample $\widetilde{S} = o$

[16] Here, we use the full embedding rather than the truncated version used in the simulations.

consists of non-study villages, for which we observe (Y,R). Appendix Figure H.1 provides a map. Further details are in Appendix H.1.

Two aspects of this exercise are noteworthy. First, the experimental sample has 155 mandals while the validation sample has 136 mandals. Whereas the benchmark has "oracle" data access, using outcomes for all 291 mandals, our method has "imperfect" data access, using outcomes for only the 136 mandals in the validation sample. Second, the validation sample is non-randomly selected: rather than containing treated and untreated villages, it only contains untreated villages. As a consequence, there are no joint observations of $(D=1,Y,R)$. The PPI rectifier for the treated group cannot be implemented, rendering PPI infeasible in this setting.[17]

Plausibility of identifying assumptions. The treatment was randomized in this experiment (Assumption 1). Appendix H.1 provides evidence that stability is plausible (Assumption 2) for each of the three poverty outcomes.

With incomplete observational cases, we require no direct effects (Assumption 3(ii)): the treatment should affect the remotely sensed variable only through the outcome. Using "oracle" data access, we can visually assess this condition; the densities $f_R(r \mid \widetilde{S} \in \{e,v\}, D=0, Y=y)$ and $f_R(r \mid \widetilde{S} \in \{e,v\}, D=1, Y=y)$ should coincide for $y \in \{0,1\}$. In Figure 4, the conditional densities closely align for one of the poverty outcomes ("middle income"), and Kolmogorov–Smirnov tests cannot reject equality at the 5% level. Appendix H.1 reports the corresponding diagnostics for the other outcomes.[18]

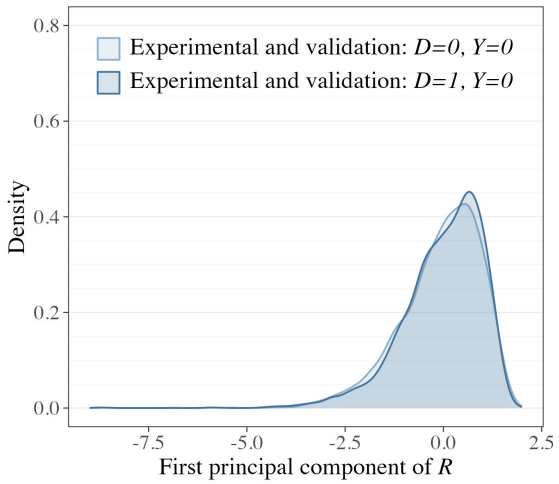
The remotely sensed variable must be relevant: its representation $H(R)$ must be correlated with outcome variation in the observational sample (Corollary 1). For the learned representation in Algorithm 1, we test whether $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ is significantly different from zero. Table 2 confirms that the remotely sensed variable is relevant for each of the three poverty outcomes.

Main results. Figure 5 summarizes our findings. For each poverty outcome, we compare confidence intervals of (i) the benchmark, a difference-in-means using all 291 mandal outcomes; and (ii) our method, using only 136 mandal outcomes but leveraging an observational sample.

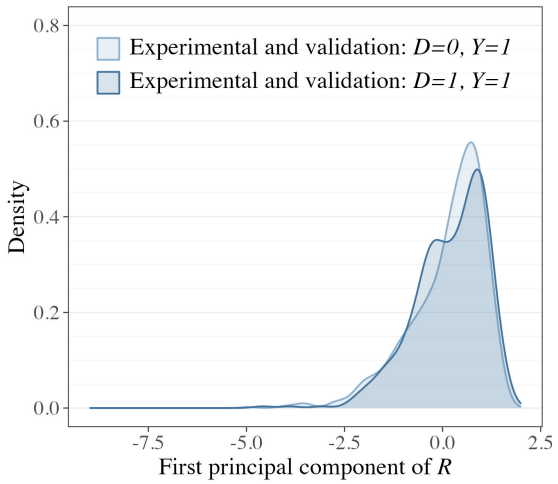
Our method yields estimates that are close to those obtained by the benchmark. For each poverty outcome, our point estimates have the same sign and similar magnitude as the benchmark, i.e. as if the economist could observe all treatments and outcomes in the Smartcards experiment. Consistent with Muralidharan et al. (2023), we find that early adoption of

[17] Formally, the term $\mathbb{E}(Y \mid D=1, \widetilde{S} \in \{v,o\})$ cannot be computed.

[18] Tests for "low income" also do not reject, while tests for "consumption" marginally reject.



(a) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=0$.



(b) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=1$.

Figure 4: No direct effects (Assumption 3(ii)) is plausible in the Smartcards experiment. We compare $f_R(R \mid \widetilde{S} \in \{e,v\}, D=0, Y=y)$ with $f_R(R \mid \widetilde{S} \in \{e,v\}, D=1, Y=y)$ for $y \in \{0,1\}$. Here, the outcome Y is "middle income." Because R is high-dimensional, we visualize the density of its standardized first principal component.

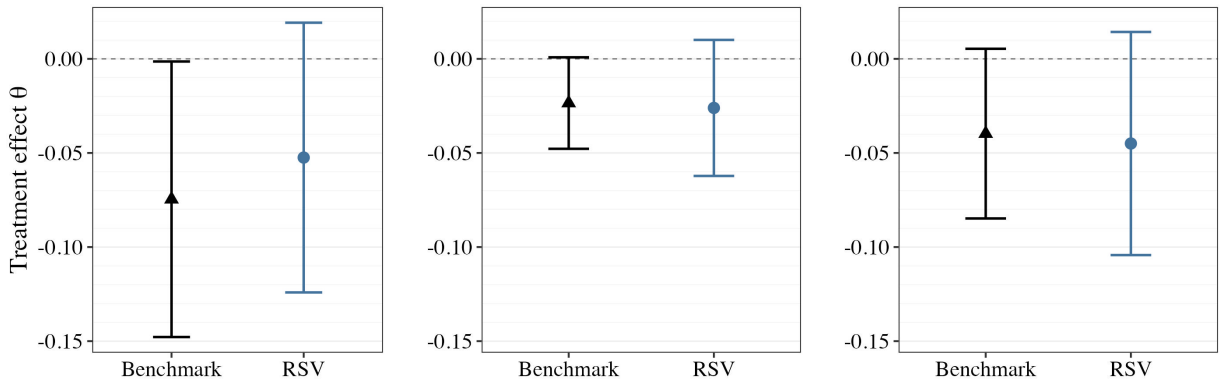
	Consumption	Low income	Middle income
Relevance, $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$	0.5000	1.0102	0.4530
	(0.0601)	(0.1156)	(0.0493)

Table 2: Satellite images are relevant to poverty outcomes. For each poverty outcome, we report $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ and its standard error using our learned representation. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

Smartcards reduces poverty. Appendix Table H.1 gives details. For each outcome, we cannot reject the null hypothesis that our method and the benchmark yield the same point estimate.

Furthermore, our method's confidence intervals are nearly the same length as the benchmark's. For the consumption outcome, our interval is essentially of the same length. For the middle-income outcome, our interval is 31% longer.

This exercise illustrates the potential for survey-cost reductions when an appropriate observational sample is available. An economist could forgo poverty surveys for a large fraction of villages in the Smartcards experiment, and instead rely on publicly available satellite images and an observational sample linking satellite images to existing poverty measurements—for



(a) Consumption. (b) Low income. (c) Middle income.

Figure 5: Our method's point estimates approximately recover the unbiased benchmark estimates. For each poverty outcome, we report 90% confidence intervals for the benchmark and for our method. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

example, from prior surveys or censuses. In this illustration, if surveying each individual in a village costs \$0.50, this choice could save \$3 million.[19]

Specification test and efficiency gains. Our model generates testable overidentifying restrictions (Remark 1): if stability and no direct effects hold, then different representations $H(R)$ and $H'(R)$ should yield the same treatment effect. We implement this test by comparing estimates from two representations: (i) the optimal representation $H(R) = H^*(R)$, combining three predictions; and (ii) the simple representation $H'(R) = \Pr(Y = 1 \mid R, \widetilde{S} \in \{o, v\})$, using one prediction. Table 3 reports the results. For each poverty outcome, we cannot reject the null of equal treatment effects, supporting the plausibility of our model.

Table 3 also confirms that the standard error using the optimal representation is meaningfully smaller than the standard error using the simple representation. For example, for the consumption outcome, the optimal representation yields a standard error that is 40% smaller. Consistent with Proposition 3, our method is valid using various representations, but the optimal representation confers efficiency. Therefore, if it is computationally feasible, researchers should consider predicting not only the outcome but also the treatment and sample indicator. The optimal representation allows for extrapolation (i.e., negative weights on villages), whereas the simple representation does not. We further interpret and compare

[19] Appendix Table F.1 summarizes the number of villages and their populations, which we use to calculate savings. This calculation is conservative; Viviano and Rudder (2020) find that phone surveys in Pakistan cost \$7 per individual, rather than \$0.50.

	Consumption	Low income	Middle income
Optimal representation	-0.0525	-0.0261	-0.0450
	(0.0436)	(0.0220)	(0.0360)
Simple representation	-0.0494	-0.0242	-0.0431
	(0.0733)	(0.0341)	(0.0542)
J statistic	0.0267	0.0600	0.0095
p value	0.8702	0.8065	0.9223

Table 3: We cannot reject the overidentifying restrictions implied by our model, and the optimal representation yields more precise inference. The optimal representation is $H^*(R)$ from Section 4.1. The simple representation is $\Pr(Y = 1 \mid R, \widetilde{S} \in \{o, v\})$. The J statistic compares treatment effect estimates based on these two representations. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

the two representations in Appendix H.1.

Remark 7 (Spillovers)**.** We compare our method to a natural benchmark: the difference-in-means estimate an economist would obtain if they could observe all treatments and outcomes in the experiment. If there are no spillovers, then we are comparing our method to an unbiased estimate of the treatment effect in Definition 1. If there are spillovers, then we are comparing to an unbiased estimate of a reasonable estimand: a difference-in-means of expected outcomes, indexing by each unit's treatment status and averaging over all other units. As discussed in Section 5.3, our method can be adapted to estimate the global average treatment effect, which is often a parameter of interest in models with spillovers. We provide this extension and apply it to the Smartcards experiment in Appendix H.1.

Remark 8 (Timing)**.** Outcomes were measured at the end of the experiment in 2013, so we study the effect on poverty in 2013. The remotely sensed variables were measured through 2020. Appendix H.1 confirms that it is valid to use remotely sensed variables measured after the outcomes. The observational sample links outcomes from 2013 with remotely sensed variables through 2020, so the estimand and method remain unchanged despite differences in timing.

7.2 Payments for Ecosystem Services and Crop Burning

Data. We use data from Jack et al. (2025). The treatment $D \in \{0,1\}$ is the offer of a PES contract, and the outcome $Y \in \{0,1\}$ indicates whether a field was *not* burned, so a positive effect means an environmental benefit.[20]

It is costly to measure whether the crop residue on a particular field has been burned, requiring a surveyor to make frequent visits to rural areas. Therefore, it is natural to turn to a remotely sensed variable R. To construct such a variable, Jack et al. (2025) link surveyor-collected measurements of crop burning to satellite-based spectral indices from PlanetScope and Sentinel-2 imagery, and then train a random forest to predict whether fields have been burned (Walker et al., 2022). The random forest outputs a continuous score at the pixel level, which is the proportion of decision trees that classify a pixel as burned; these pixel-level scores are aggregated to field-level scores. Jack et al. (2025) then construct binary classifiers at the field level by applying threshold rules to field-level scores.[21]

The authors conducted randomized spot checks, sending a surveyor to inspect certain fields, but the timing was imperfect. Because crop burning is a short-lived event, "spot checks could not be synchronized with the farmer's residue management timing...[spot checks] indicate no burning in days immediately preceding the spot check visit but provide no information about burning outside that window" (Jack et al., 2025, p. 45). For this reason, the authors caution against interpreting the spot check sample as a randomly collected validation sample.

We consider two variations. Throughout, the experimental sample ($\widetilde{S} = e$) consists of treated and untreated fields, for which we observe (D,R). In one variation, we view the spot check sample as a validation sample ($\widetilde{S} = v$) (Section 5.1) for which we observe (D,Y,R). In another variation, we view the spot check sample as an observational sample ($\widetilde{S} = o$), for which we observe (D,Y,R). In the spot check sample, there are no joint observations of $(D,Y = 1,R)$ in which D varies, so Assumption 3(i) fails and we must appeal to Assumption 3(ii).

Because of the imperfect timing of the spot checks, the outcomes from the spot check sample do not provide unbiased estimates of potential outcomes, making PPI biased (Proposi-

[20] We modify the specification of Jack et al. (2025) in two ways. While the authors distinguish between standard and up-front PES contracts, we define the treatment as whether any PES contract was offered. The authors analyze effects at the farmer level, aggregating across fields, while we analyze effects at the field level.

[21] Jack et al. (2025) construct two binary classifiers for crop burning by applying two different threshold rules to the field-level scores. We use their "max accuracy" classifier in the main text, and we report analogous results using their "balanced accuracy" classifier in Appendix H.2.

tion 2).[22] On the other hand, stability of the sensing mechanism is plausible, since it requires that the relationship between satellite images and crop burning is stable.

Plausibility of identifying assumptions. Since offers of PES contracts were randomized at the village level, experimental unconfoundedness (Assumption 1) is satisfied by design.

Unlike the Smartcards application, "oracle" data access is unavailable in this application, so we appeal to domain knowledge to justify the remaining assumptions. To justify stability (Assumption 2), we appeal to the fact that the same sensing technology was used across samples. Conditional on burning status, how crop burning manifests in the satellite-based spectral indices is determined by physical properties of the soil surface, which are plausibly the same across samples. To justify no direct effects (Assumption 3(ii)), we appeal to the specificity of the sensing mechanism: it detects particular infrared spectral bands. Other changes that PES contracts might induce, such as investments in farm equipment, should not affect the spectral indices used to detect burning.

As before, the remotely sensed variable must be relevant. We confirm that the learned representation is correlated with outcome variation: $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\} = 0.262$, with standard error 0.048.[23]

Main results. Table 4 summarizes our findings. For each variation, we compare estimates of (i) the common practice, which regresses the predicted outcome on the treatment; (ii) our method using the optimal representation; and (iii) our method using a simple representation.

Our method yields a substantially larger treatment effect. We find that offering any PES contract reduces burning by 14.1%. By contrast, the common practice estimates an effect of 7.5%, which is 47% smaller in magnitude.[24] Interpreted through our framework, the common two-step method appears to attenuate the treatment effect by 47% (Proposition 1).

Specification test and efficiency gains. As before, our model generates testable overidentifying restrictions (Remark 1). We implement this test by comparing estimates from two

[22] For example, $\tau \neq \widetilde{\tau}^{\text{PPI}}$ when the correlation between crop-burning-ever and PES contracts in the experimental sample is different than the correlation between crop-burning-in-a-time-window and PES contracts in the spot check sample.

[23] The standard error is based on 5000 bootstrap replications clustered at the village level.

[24] Jack et al. (2025, Table 2) report effects of 7.7% for up-front PES and 2.0% for standard PES, at the farmer level. Our estimate pools across the two types of PES, at the field level.

representations. Table 4 reports the results. For each variation, we cannot reject the null of equal treatment effects, supporting the plausibility of our model.

	Spot checks as validation	Spot checks as observational
Optimal representation	0.1210	0.1417
	(0.0684)	(0.0782)
Simple representation	0.1565	0.1929
	(0.0803)	(0.0989)
Common practice	0.0588	0.0752
	(0.0320)	(0.0380)
J statistic	1.3242	1.7439
p value	0.2498	0.1866

Table 4: Our method yields substantially larger treatment effects than the common practice. The common practice is from Section 3.2.1. Within our method, the optimal representation is $H^*(R)$ from Section 4.1, and the simple representation is $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$. The J statistic compares treatment effect estimates based on these two representations. Bootstrap standard errors, based on 5000 replications, are clustered at the village level.

Table 4 also confirms that the optimal representation gives a meaningfully smaller standard error than the simple representation. For example, when viewing the spot check sample as an observational sample, the optimal representation yields a standard error that is 21% smaller.

8 Recommendations for Practice

We study remotely sensed variables as *post-outcome*: variation in the economic outcome causes variation in the remotely sensed variable, not vice versa. We formalize this causal structure and derive a new formula to combine an experimental sample (containing randomized treatments and remotely sensed variables) with an observational sample (linking remotely sensed variables to outcomes).

In this setting, some alternative methods are problematic. A common practice implicitly uses the remotely sensed variable as a surrogate variable, which mediates the effect of the treatment on the outcome. Because the remotely sensed variable is post-outcome, this method

suffers from attenuation bias (Proposition 1). Prediction-powered inference (PPI) methods require an auxiliary sample that contains jointly observed treatments, outcomes, and remotely sensed variables (for both treatment values) to construct bias corrections. When such an auxiliary sample is unavailable, PPI is infeasible. Even when such an auxiliary sample is available, PPI can be biased whenever the experimental and auxiliary samples differ in baseline outcomes or treatment effects (Proposition 2). Such a distribution shift is plausible when the two samples come from different populations, time periods, or regions.

However, the core intuition underlying empirical practice is powerful: the conditional distribution of the remotely sensed variable given the outcome and treatment should be *stable* across samples. Our stability assumption formalizes the idea that the sensing mechanism—how satellite images or mobile phone data reflect economic conditions—operates the same way in different settings. Under stability, we nonparametrically identify treatment effects and precisely estimate them using remotely sensed variables.

Based on our framework, we summarize concrete recommendations for researchers conducting program evaluation with remotely sensed outcomes.

Choosing the remotely sensed variable and the observational sample. The researcher faces two key decisions: which remotely sensed variable to use, and how to construct the observational sample linking it to outcomes.

Our framework is agnostic about what the remotely sensed variable is: it may be the output of a pretrained predictor, a pretrained embedding of unstructured remote sensing inputs (e.g., imagery), or the remote sensing inputs themselves. Our main recommendation is that it should reflect the outcome through physical or technological channels. For example, satellite imagery captures crop burning through changes in soil reflectance from charred fields (Walker et al., 2022; Jack et al., 2025), deforestation through changes in color saturation (Hansen et al., 2013; Jayachandran et al., 2017), and poverty through changes in roofing materials and building footprints (Jean et al., 2016; Huang et al., 2021).

Given this choice, the observational sample should be constructed to make stability (Assumption 2) credible. Researchers should select observational units that have the same sensing mechanism (e.g., the same satellite technology) as well as similar conditions that might affect the measurement process (e.g., similar soil type and urban density). Differences in the outcome mechanism (perhaps due to timing of data collection) are permitted. The treatment need not be observed, let alone randomized, in the observational sample.

Inference with remotely sensed outcomes. Valid inference requires the remotely

sensed variable to be predictive of the outcome. More efficient inference leverages three predictions from the remotely sensed variable: predictions of the outcome Y, the treatment D, and the sample indicator S. We aggregate these three predictions into a more efficient representation of the remotely sensed variable, reducing its dimension while preserving information relevant for causal inference (Algorithm 1). Crucially, our method is robust to misspecification. In practice, more precise predictions deliver more efficient inference.

Diagnostics for remotely sensed outcomes. Our framework yields two diagnostics that should be reported when conducting program evaluation with remotely sensed outcomes.

First, just as weak instruments threaten inference in instrumental variable models, weak remotely sensed variables threaten inference in our framework. Researchers should test for weak remotely sensed variables using existing tests for weak instruments.

Second, our identifying assumptions generate testable overidentifying restrictions (Remark 1). For example, if stability and no direct effects hold, then different representations of the remotely sensed variable should yield the same treatment effect estimate. Disagreement across representations constitutes evidence against the identifying assumptions, via a J-test.

Future work. While we focus on applications to environmental and development economics—where the sensing mechanism is a physical technology that is plausibly stable—other possible applications may be in applied microeconomics. For example, arrests may be post-outcome measurements of crime; test scores may be post-outcome measurements of understanding; and biomarkers may be post-outcome measurements of latent disease.

Finally, our framework poses new questions. Future research may study how to simultaneously design treatment assignment, outcome collection, and sensor deployment to maximize statistical power, subject to budget constraints. The solution to this problem would inform not only the use of satellite images, but also the design of cheap, noisy surveys.

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Aiken, E., S. Bellue, J. E. Blumenstock, D. Karlan, and C. Udry (2025). Estimating impact with surveys versus digital traces: Evidence from randomized cash transfers in Togo. *Journal of Development Economics 175*, 103477.

Aiken, E., S. Bellue, D. Karlan, C. Udry, and J. E. Blumenstock (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature 603*(7903), 864–870.

Alix-Garcia, J. and D. L. Millimet (2023). Remotely incorrect? Accounting for nonclassical

measurement error in satellite data on deforestation. *Journal of the Association of Environmental and Resource Economists 10*(5), 1335–1367.

Allon, G., D. Chen, Z. Jiang, and D. Zhang (2023). Machine learning and prediction errors in causal inference. Technical report, The Wharton School. SSRN ID 4480696.

Andrews, D. W. (2017). Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics 199*(2), 213–220.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023). Prediction-powered inference. *Science 382*(6671), 669–674.

Angelopoulos, A. N., J. C. Duchi, and T. Zrnic (2024). Ppi++: Efficient prediction-powered inference.

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics 14*(1), 57–67.

Asher, S., T. Lunt, R. Matsuura, and P. Novosad (2021). Development research at high geographic resolution: An analysis of night-lights, firms, and poverty in India using the SHRUG open data platform. *The World Bank Economic Review 35*(4), 845–871.

Asher, S. and P. Novosad (2020). Rural roads and local economic development. *American Economic Review 110*(3), 797–823.

Assuncao, J., R. McMillan, J. Murphy, and E. Souza-Rodrigues (2023). Optimal environmental targeting in the Amazon rainforest. *The Review of Economic Studies 90*(4), 1608–1641.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2025, 09). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*, rdaf087.

Azevedo, A. A., R. Rajão, M. A. Costa, M. C. C. Stabile, M. N. Macedo, T. N. P. dos Reis, A. Alencar, B. S. Soares-Filho, and R. Pacheco (2017). Limits of brazil's forest code as a means to end illegal deforestation. *Proceedings of the National Academy of Sciences 114*(29), 7653–7658.

Balboni, C., R. Burgess, and B. A. Olken (2025). The origins and control of forest fires in the tropics. *The Review of Economic Studies*, rdaf088.

Bareinboim, E. and J. Pearl (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences 113*(27), 7345–7352.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024, February). Inference for regression with variables generated by AI or machine learning.

Bluhm, R. and M. Krause (2022). Top lights: Bright cities and their contribution to economic development. *Journal of Development Economics 157*, 102880.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science 350*(6264), 1073–1076.

Burgess, R., M. Hansen, B. A. Olken, P. Potapov, and S. Sieber (2012, 11). The political economy of deforestation in the tropics*. *The Quarterly Journal of Economics 127*(4), 1707–1754.

Burke, M., A. Driscoll, D. B. Lobell, and S. Ermon (2021). Using satellite imagery to understand and promote sustainable development. *Science 371*(6535), eabe8628.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Canay, I. A., A. Santos, and A. M. Shaikh (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica 81*(6), 2535–2559.

Carlson, J. and M. Dell (2025). A unifying framework for robust and efficient inference with unstructured data. *arXiv:2505.00282*.

Carrasco, M., J.-P. Florens, and E. Renault (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics 6*, 5633–5751.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Chen, J., D. L. Chen, and G. Lewis (2020). Mostly harmless machine learning: Learning optimal instruments in linear IV models. *arXiv preprint arXiv:2011.06158*. Accepted at NeurIPS 2020 Workshop on Machine Learning for Economic Policy.

Chen, J. J., V. Mueller, Y. Jia, and S. K.-H. Tseng (2017). Validating migration responses to flooding using satellite and vital registration data. *American Economic Review 107*(5), 441–45.

Chen, X. (2007). Chapter 76 large sample sieve estimation of semi-nonparametric models. Volume 6 of *Handbook of Econometrics*, pp. 5549–5632. Elsevier.

Chen, X., H. Hong, and D. Nekipelov (2011). Nonlinear models of measurement errors. *Journal of Economic Literature 49*(4), 901–37.

Chen, X., H. Hong, and E. Tamer (2005). Measurement error models with auxiliary data. *The Review of Economic Studies 72*(2), 343–366.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics 36*(2), 808–843.

Chen, X. and W. D. Nordhaus (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences 108*(21), 8589–8594.

Chen, X. and M. Reiss (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory 27*(3), 497–521.

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

Chernozhukov, V., W. Newey, R. Singh, and V. Syrgkanis (2025). Adversarial estimation of Riesz representers. *Journal of the American Statistical Association*.

Chernozhukov, V., W. K. Newey, and R. Singh (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika 110*(1), 257–264.

Corral, P., H. Henderson, and S. Segovia (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics 172*, 103377.

Cross, P. J. and C. F. Manski (2002). Regressions, short and long. *Econometrica 70*(1), 357–368.

Crossley, T. F., P. Levell, and S. Poupakis (2022). Regression with an imputed dependent variable. *Journal of Applied Econometrics 37*(7), 1277–1294.

Currie, J., J. Voorheis, and R. Walker (2023). What caused racial disparities in particulate exposure to fall? new evidence from the clean air act and satellite-based measures of air quality. *American Economic Review 113*(1), 71–97.

de Chaisemartin, C. and X. D'Haultfœuille (2020, September). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2025). Partially linear models under data combination. *The Review of Economic Studies 92*(1), 238–267.

Donaldson, D. and A. Storeygard (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives 30*(4), 171–98.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2024). Linear regressions with combined data. *arXiv:2412.04816*.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems 36*, 68589–68601.

Egger, D., J. Haushofer, E. Miguel, P. Niehaus, and M. Walker (2022). General equilibrium effects of cash transfers: Experimental evidence from Kenya. *Econometrica 90*(6), 2603–2643.

Fan, Y., R. Sherman, and M. Shum (2014). Identifying treatment effects under data combination. *Econometrica 82*(2), 811–822.

Fong, C. and M. Tyler (2021). Machine learning predictions as regression covariates. *Political Analysis 29*(4), 467–484.

Fowlie, M., E. Rubin, and R. Walker (2019, May). Bringing satellite-based air quality estimates down to earth. *AEA Papers and Proceedings 109*, 283–88.

Ghassami, A., C. Liu, A. Yang, D. Richardson, I. Shpitser, and E. T. Tchetgen (2022). Combining experimental and observational data for identification and estimation of long-term causal effects. *arXiv:2201.10743*.

Graham, B. S., C. C. de Xavier Pinto, and D. Egel (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *Journal of Business & Economic Statistics 34*(2), 288–301.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science 342*(6160), 850–853.

Henderson, V., A. Storeygard, and D. N. Weil (2011, May). A bright idea for measuring economic growth. *American Economic Review 101*(3), 194–99.

Henderson, V., A. Storeygard, and D. N. Weil (2012). Measuring economic growth from outer space. *American Economic Review 102*(2), 994–1028.

Horowitz, J. L. and C. F. Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica 63*(2), 281–302.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*(1), 27–61.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Huang, L. Y., S. M. Hsiang, and M. Gonzalez-Navarro (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. Working Paper w29105, National Bureau of Economic Research.

Imbens, G., N. Kallus, X. Mao, and Y. Wang (2025, 4). Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(2), 362–388.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Jack, B. K., S. Jayachandran, N. Kala, and R. Pande (2025). Money (not) to burn: Payments for ecosystem services to reduce crop residue burning. *American Economic Review: Insights 7*(1), 39–55.

Jack, B. K. and K. Walker (2023). Integrating remote sensing and randomized controlled trials: Challenges, opportunities, and practical guidance. Mimeo, Bren School of Environmental Science & Management, UC Santa Barbara.

Jayachandran, S., J. De Laat, E. F. Lambin, C. Y. Stanton, R. Audy, and N. E. Thomas (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science 357*(6348), 267–273.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*(6301), 790–794.

Ji, W., L. Lei, and T. Zrnic (2025). Predictions as surrogates: Revisiting surrogate outcomes in the age of AI.

Johannemann, J., V. Hadad, S. Athey, and S. Wager (2019). Sufficient representations for categorical variables. *arXiv preprint arXiv:1908.09874*.

Josephson, A., J. D. Michler, T. Kilic, and S. Murray (2026). The mismeasure of weather: Using earth observation data for estimation of socioeconomic outcomes. *Journal of Development Economics 178*, 103553.

Kallus, N. and X. Mao (2024, 10). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology 87*(2), 480–509.

Kluger, D. M., K. Lu, T. Zrnic, S. Wang, and S. Bates (2025). Prediction-powered inference with imputed covariates and nonuniform sampling. *arXiv preprint*.

Kress, R. (1989). *Linear Integral Equations*, Volume 82 of *Applied Mathematical Sciences*. Springer.

Lu, K., D. M. Kluger, S. Bates, and S. Wang (2025). Regression coefficient estimation from remote sensing maps. *Remote Sensing of Environment 330*, 114949.

Mackey, L., V. Syrgkanis, and I. Zadik (2018). Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pp. 3375–3383. PMLR.

Marx, B., T. M. Stoker, and T. Suri (2019). There is no free house: Ethnic patronage in a Kenyan slum. *American Economic Journal: Applied Economics 11*(4), 36–70.

Masten, M. A. and A. Poirier (2020). Inference on breakdown frontiers. *Quantitative Economics 11*(1), 41–111.

Miao, W., L. Liu, Y. Li, E. J. Tchetgen Tchetgen, and Z. Geng (2024). Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM/IMS Journal of Data Science 1*(2), 1–23.

Miao, W. and E. J. Tchetgen Tchetgen (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika 103*(2), 475–482.

Michaels, G., D. Nigmatulina, F. Rauch, T. Regan, N. Baruah, and A. Dahlstrand-Rudin (2021). Planning ahead for better neighborhoods: Long-run evidence from Tanzania. *Journal of Political Economy 129*(7), 2112–2156.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2016). Building state capacity: Evidence from biometric smartcards in India. *American Economic Review 106*(10), 2895–2929.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2023). General equilibrium effects of (improving) public employment programs: Experimental evidence from India. *Econometrica 91*(4), 1261–1295.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In G. S. Maddala, C. R. Rao, and H. D. Vinod (Eds.), *Handbook of Statistics*, Volume 11, Chapter 16, pp. 419–454. Elsevier.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Park, C., D. B. Richardson, and E. J. Tchetgen Tchetgen (2024). Single proxy control. *Biometrics 80*(2), ujae027.

Patel, D. (2024). Floods. Technical report, Department of Economics, Harvard University.

Pelletier, J., M. Korb, S. Alemu, M. B. Yonis, T. J. Lybbert, and M. Stigler (2026). Causal inference with predicted outcomes: Correcting prediction error bias in satellite-based impact evaluation. *Journal of Development Economics 179*, 103655.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine 8*(4), 431–440.

Proctor, J., T. Carleton, T. Chong, T. Fransen, S. Greenhill, J. Katz, H. Murayama, L. Sherman, J. Tseng, H. Druckenmiller, and S. Hsiang (2025, October). What can satellite imagery and machine learning measure? Working Paper 34315, National Bureau of Economic Research.

Proctor, J., T. Carleton, and S. Sum (2023). Parameter recovery using remotely sensed variables. Technical Report 30861, National Bureau of Economic Research. NBER Working Paper.

Ratledge, N., G. Cadamuro, B. De La Cuesta, M. Stigler, and M. Burke (2022). Using machine learning to assess the livelihood impact of electricity access. *Nature 611*(7936), 491–495.

Ridder, G. and R. Moffitt (2007). The econometrics of data combination. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, Chapter 75, pp. 5469–5547. Amsterdam: Elsevier.

Rolf, E., J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 12*, 4392.

Roth, J. and P. H. C. Sant'Anna (2023). When is parallel trends sensitive to functional form? *Econometrica 91*(2), 737–747.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons.

Sanford, L. C., M. Ayers, M. Gordon, and E. Stone (2025). Adversarial debiasing for unbiased parameter recovery.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica 26*(3), 393–415.

Schennach, S. M. (2020). Mismeasured and unobserved variables. In S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin (Eds.), *Handbook of Econometrics*, Volume 7, pp. 487–565. Elsevier.

Torchiana, A. L., T. Rosenbaum, P. T. Scott, and E. Souza-Rodrigues (2025). Improving estimates of transitions from satellite data: A hidden markov model approach. *The Review of Economics and Statistics 107*(2), 426–441.

Vafa, K., S. Athey, and D. M. Blei (2025). Estimating wage disparities using foundation models. *Proceedings of the National Academy of Sciences 122*(22), e2427298122.

Viviano, D. and J. Rudder (2020). Policy design in experiments with unknown interference. arXiv preprint 2011.08174.

Walker, K., B. Moscona, K. Jack, S. Jayachandran, N. Kala, R. Pande, J. Xue, and M. Burke (2022). Detecting crop burning in India using satellite data. arXiv preprint arXiv:2209.10148.

Wunder, S., S. Engel, and S. Pagiola (2008). Taking stock: A comparative analysis of payments for environmental services programs in developed and developing countries. *Ecological Economics 65*(4), 834–852.

Program Evaluation with Remotely Sensed Outcomes

Appendix Materials

Ashesh Rambachan, Rahul Singh, Davide Viviano

A Additional Model Details **4**

B Proofs for Theoretical Results in the Main Text **5**

B.1 Proofs for Section 3 . . . 5

B.1.1 Proof of Lemma 1 . . . 5

B.1.2 Proof of Theorem 1 . . . 5

B.1.3 Proof of Theorem 2 . . . 7

B.1.4 Proof of Proposition 1 . . . 7

B.1.5 Proof of Proposition 2 . . . 9

B.2 Proofs for Section 4 . . . 11

B.2.1 Proof of Proposition 3 . . . 11

B.2.2 Proof of Proposition 4 . . . 13

C Additional Theoretical Results and Discussion **16**

C.1 Bias of Alternative Approaches: Extensions . . . 16

C.1.1 Bias of Common Practice with Complete Observational Cases . . . 17

C.1.2 Bias of Common Practice with Random Sample Selection . . . 17

C.2 Quasi-Experiments with Remotely Sensed Outcomes . . . 18

C.2.1 Instrumental Variables . . . 18

C.2.2 Two-Period Difference-in-Differences . . . 20

C.3 Some Spillovers . . . 22

C.3.1 Generalized Potential Outcomes . . . 22

C.3.2 Average Global Treatment Effect . . . 24

C.4 Variables Collected at Different Times . . . 25

C.5 Partial Identification with Approximate Assumptions . . . 27

C.5.1 Approximately No Direct Effects . . . 28

C.5.2 Approximately No Direct Effects and Approximate Stability . . . 30

C.6 Continuous Outcomes . . . 31

C.6.1 Discretization . . . 32
C.6.2 Generalizing the Relevance Condition . . . 33
C.7 Estimation with Covariates . . . 35
C.7.1 Discrete Covariates and Incomplete Observational Cases . . . 35
C.7.2 Discrete Covariates and Complete Observational Cases . . . 37
C.7.3 Continuous Covariates . . . 37

D Proofs for Additional Theoretical Results 39
D.1 Bias of Alternative Approaches: Extensions . . . 39
D.1.1 Proof of Proposition C.1 . . . 39
D.1.2 Proof of Proposition C.2 . . . 41
D.2 Quasi-Experiments with Remotely Sensed Outcomes . . . 42
D.2.1 Proof of Theorem C.1 . . . 42
D.2.2 Proof of Theorem C.2 . . . 43
D.3 Some Spillovers . . . 44
D.3.1 Proof of Lemma C.1 . . . 44
D.3.2 Proof of Theorem C.3 . . . 44
D.3.3 Proof of Corollary C.3 . . . 45
D.3.4 Proof of Corollary C.4 . . . 46
D.4 Variables Collected at Different Times . . . 46
D.4.1 Proof of Theorem C.4 . . . 46
D.5 Partial Identification with Approximate Assumptions . . . 46
D.5.1 Proof of Lemma C.2 . . . 46
D.5.2 Proof of Theorem C.5 . . . 47
D.5.3 Proof of Lemma C.3 . . . 48
D.5.4 Proof of Theorem C.6 . . . 49
D.6 Continuous Outcomes . . . 50
D.6.1 Proof of Proposition C.3 . . . 50
D.6.2 Proof of Corollary C.5 . . . 51
D.6.3 Proof of Proposition C.4 . . . 52
D.7 Estimation with Covariates . . . 53
D.7.1 Proof of Proposition C.5 . . . 53
D.7.2 Proof of Proposition C.6 . . . 53

E Additional Estimation Algorithms **54**

E.1 Estimation with Covariates 54

E.2 An Algebraic Simplification 55

F Additional Simulation Details **57**

F.1 Forest Cover Details 57

F.2 Smartcards Details 58

F.3 Plausibility of Identifying Assumptions 60

G Additional Simulation Results **62**

G.1 Experimental and Observational Samples Only 62

G.2 Experimental, Observational, and Validation Samples 66

H Additional Empirical Results **70**

H.1 Smartcards and Village-Level Poverty 70

H.2 Payments for Ecosystem Services and Crop Burning 76

A Additional Model Details

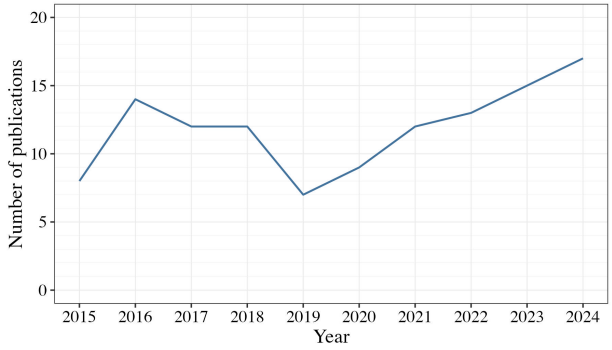


Figure A.1: Remotely sensed variables are increasingly popular in published papers.

Figure A.1 illustrates the increasing popularity of remotely sensed variables in empirical research. First, we collected papers published in the American Economic Association (AEA) journals, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economic Studies* and *Journal of Political Economy* using a keyword search of "remotely sensed variables", "mobile phone", "satellite", "machine learning", and "drones / aerial" on their websites. Next, we collected papers published in *Nature* and *Science* using a keyword search of "remotely sensed variables" on their websites. We subset to the papers with remotely sensed variables in their main empirical analysis.

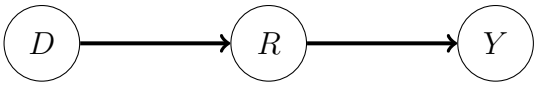


Figure A.2: Causal graph of surrogacy model.

Figure A.2 illustrates the causal graph associated with the surrogacy identifying assumptions (Prentice, 1989; Athey et al., 2025): the surrogate R fully mediates the effect of the treatment D on the outcome Y. Our identifying assumptions are the opposite, as illustrated by Figure 1.

Table A.1: Implications of the identifying assumptions.

Assumption	Quantity	Experimental	Observational	Description
1	$\Pr(Y(d)\|S,X,D)$	$\Pr(Y(d)\|S=e,X)$	$\Pr(Y(d)\|S=o,X,D)$	Differs across samples
2	$\Pr(R\|S,X,D,Y)$	$\Pr(R\|X,D,Y)$	$\Pr(R\|X,D,Y)$	Stable across samples
3(i)	$\Pr(R\|S,X,D,Y)$	$\Pr(R\|X,D,Y)$	$\Pr(R\|X,D,Y)$	Differs across treatments if complete cases
3(ii)	$\Pr(R\|S,X,D,Y)$	$\Pr(R\|X,Y)$	$\Pr(R\|X,Y)$	Stable across treatments if incomplete cases

Table A.1 summarizes the implications of our identifying assumptions. Our identifying assumptions allow the propensity score $\Pr(D=1\,|\,S,X)$ to differ across samples. In summary, the outcome and treatment mechanisms may vary across samples, but the sensing mechanism must be stable across samples.

B Proofs for Theoretical Results in the Main Text

B.1 Proofs for Section 3

B.1.1 Proof of Lemma 1

We consider the general case where (X,Y,R) may be discrete or continuous. Then, we specialize the result to the case where Y is discrete. Recall that $f_W(\cdot\,|\,...)$ is the Radon-Nikodym derivative.

By the law of total probability,

$$\begin{aligned}\delta_d^e(X,R) &:= f_R(R\,|\,S=e,X,D=d)\\ &= \int f_{R,Y}(y,R\,|\,S=e,X,D=d)\mathrm{d}y\\ &= \int f_R(R\,|\,S=e,X,D=d,Y=y) f_Y(y\,|\,S=e,X,D=d)\mathrm{d}y.\end{aligned}$$

By Assumption 1, $f_Y(y\,|\,S=e,X,D=d) = f_{Y(d)}(y\,|\,S=e,X)$. Next, notice that

$$\begin{aligned}f_R(R\,|\,S=e,X,D=d,Y=y) &= f_R(R\,|\,X,Y=y)\\ &= f_R(R\,|\,S=o,X,Y=y) =: \delta_y^o(X,R),\end{aligned}$$

where the equalities apply the contraction implied by Assumptions 2 and 3(ii): $(S,D)\perp\!\!\!\perp R|X,Y$.

Combining the previous displays, we arrive at the general result

$$\delta_d^e(X,R) = \int \delta_y^o(X,R) f_{Y(d)}(y\,|\,S=e,X)\mathrm{d}y.$$

When Y is discrete, $f_{Y(d)}(y_k\,|\,S=e,X) = \mu_k(d,X)$ and the integral becomes a sum. □

B.1.2 Proof of Theorem 1

We proceed in three steps.

1. Define the treatment weights in the experimental sample as

$$\pi_d(X,R) := \frac{\Pr(D=d,S=e\,|\,X,R)}{\Pr(D=d,S=e\,|\,X)},$$

and the outcome weights in the observational sample as

$$\gamma_y(X,R) := \frac{\Pr(Y=y,S=o\,|\,X,R)}{\Pr(Y=y,S=o\,|\,X)}.$$

By Bayes' rule,

$$f_R(R\,|\,Y=y,S=o,X) = \frac{\Pr(Y=y,S=o\,|\,X,R) f_R(R\,|\,X)}{\Pr(Y=y,S=o\,|\,X)} = \gamma_y(X,R) f_R(R|X),$$
$$f_R(R\,|\,D=d,S=e,X) = \frac{\Pr(D=d,S=e\,|\,X,R) f_R(R\,|\,X)}{\Pr(D=d,S=e\,|\,X)} = \pi_d(X,R) f_R(R\,|\,X).$$

Substituting these expressions into Lemma 1 and canceling $f_R(R\,|\,X)$ yields

$$\pi_d(X,R) = \sum_{k=1}^{K} \gamma_{y_k}(X,R) \Pr\{Y(d)=y_k\,|\,S=e,X\}.$$

2. Since $\sum_{k=1}^{K} \Pr\{Y(d)=y_k\,|\,S=e,X\}=1$, we can rewrite the previous display as

$$\pi_d(X,R) = \sum_{k=1}^{K-1} \gamma_{y_k}(X,R) \Pr\{Y(d)=y_k\,|\,S=e,X\} + \gamma_{y_K}(X,R)\left[1 - \sum_{k=1}^{K-1} \Pr\{Y(d)=y_k\,|\,S=e,X\}\right].$$

Re-arranging yields

$$\pi_d(X,R) - \gamma_{y_K}(X,R) = \sum_{k=1}^{K-1} \{\gamma_{y_k}(X,R) - \gamma_{y_K}(X,R)\} \Pr\{Y(d)=y_k\,|\,S=e,X\}.$$

We then define

$$\tilde{\gamma}(X,R) := \begin{bmatrix} \gamma_{y_1}(X,R) - \gamma_{y_K}(X,R) \\ \vdots \\ \gamma_{y_{K-1}}(X,R) - \gamma_{y_K}(X,R) \end{bmatrix}, \quad \mu(d,X) := \begin{bmatrix} \Pr\{Y(d)=y_1\,|\,S=e,X\} \\ \vdots \\ \Pr\{Y(d)=y_{K-1}\,|\,S=e,X\} \end{bmatrix}$$

and rewrite the previous display as

$$\pi_d(X,R) - \gamma_{y_K}(X,R) = \tilde{\gamma}(X,R)^\top \mu(d,X).$$

Applying iterated expectations, we have shown, for $d \in \{0,1\}$,

$$\mathbb{E}\big\{\Delta^e(d,X) - \Delta^o(X)^\top \mu(d,X)\,|\,X,R\big\} = 0 \text{ almost surely,}$$

where $\Delta^e(d,x) := \frac{1\{D=d,S=e\}}{\Pr(D=d,S=e|X)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X)}$ and

$$\Delta^o(x) := \begin{bmatrix} \frac{1\{Y=y_1,S=o\}}{\Pr(Y=y_1,S=o|X)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X)} \\ \vdots \\ \frac{1\{Y=y_{K-1},S=o\}}{\Pr(Y=y_{K-1},S=o|X)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X)} \end{bmatrix} \in \mathbb{R}^{K-1}.$$

3. We take the difference of the treated and untreated moment restrictions, thereby proving the result. □

B.1.3 Proof of Theorem 2

The proof follows the same steps as the proofs of Lemma 1 and Theorem 1.

1. As in Lemma 1, for general (X,Y,R),

$$\delta_d^e(X,R) := f_R(R \mid S=e,X,D=d) = \int f_R(R \mid S=e,X,D=d,Y=y) f_Y(y \mid S=e,X,D=d) \mathrm{d}y.$$

By Assumption 1, $f_Y(y \mid S=e,X,D=d) = f_{Y(d)}(y \mid S=e,X)$. Now, however,

$$f_R(R \mid S=e,X,D=d,Y=y) = f_R(R \mid S=o,X,D=d,Y=y) =: \delta_{y,d}^o(X,R),$$

where the first equality applies Assumption 2 and Assumption 3(i) so that the conditioning event has non zero probability.

Combining the previous displays, we arrive at the general result

$$\delta_d^e(X,R) = \int \delta_{y,d}^o(X,R) f_{Y(d)}(y \mid S=e,X) \mathrm{d}y.$$

When Y is discrete, $f_{Y(d)}(y_k \mid S=e,X) = \mu_k(d,X)$ and the integral becomes a sum.

2. As in Theorem 1, by Bayes' rule,

$$\begin{aligned}
\delta_{y,d}^o(X,R) &= \frac{\Pr(Y=y,D=d,S=o \mid X,R) f_R(R \mid X)}{\Pr(Y=y,D=d,S=o \mid X)} = \gamma_{y,d}(X,R) f_R(R \mid X) \\
\delta_d^e(X,R) &= \frac{\Pr(D=d,S=e \mid X,R) f_R(R \mid X)}{\Pr(D=d,S=e \mid X)} = \pi_d(X,R) f_R(R \mid X),
\end{aligned}$$

where

$$\pi_d(X,R) := \frac{\Pr(D=d,S=e \mid X,R)}{\Pr(D=d,S=e \mid X)}, \quad \gamma_{y,d}(X,R) = \frac{\Pr(Y=y,D=d,S=o \mid X,R)}{\Pr(Y=y,D=d,S=o \mid X)}.$$

Substituting these expressions into the previous step and canceling $f_R(R \mid X)$ yields

$$\pi_d(X,R) = \sum_{k=1}^{K} \gamma_{y_k,d}(X,R) \Pr\{Y(d)=y_k \mid S=e,X\}.$$

3. The remainder of the proof is identical to step 2 of Theorem 1, replacing γ_{y_k} with $\gamma_{y_k,d}$. □

B.1.4 Proof of Proposition 1

Under Assumption 1, 2 and 3(ii), $(S,D) \perp\!\!\!\perp R \mid Y$.

1. We prove (i) with possibly discrete outcomes. Fix $d \in \{0,1\}$. Let

$$m(R) = \mathbb{E}(Y \mid R, S=o) = \int y f_Y(y \mid R, S=o) \mathrm{d}y.$$

By Bayes' rule,

$$f_Y(Y \mid R, S=o) = \frac{f_R(R \mid Y, S=o) f_Y(Y \mid S=o)}{f_R(R \mid S=o)}$$
$$f_R(R \mid Y, D=d, S=e) = \frac{f_Y(Y \mid R, D=d, S=e) f_R(R \mid D=d, S=e)}{f_Y(Y \mid D=d, S=e)}.$$

By $(S,D) \perp\!\!\!\perp R \mid Y$,

$$f_R(R \mid Y, S=o) = f_R(R \mid Y, D=d, S=e).$$

We combine these expressions to rewrite

$$f_Y(Y \mid R, S=o) = f_Y(Y \mid R, D=d, S=e) w_d(R,Y), \quad w_d(R,Y) = \frac{f_Y(Y \mid S=o)}{f_Y(Y \mid D=d, S=e)} \frac{f_R(R \mid D=d, S=e)}{f_R(R \mid S=o)}.$$

Consequently,

$$m(R) = \int y f_Y(y \mid R, D=d, S=e) w_d(R,y) \mathrm{d}y = \mathbb{E}\{Y w_d(R,Y) \mid R, D=d, S=e\}.$$

Since $\widetilde{\mu}(d) = \mathbb{E}\{m(R) \mid D=d, S=e\}$, the result follows by iterated expectations.

2. We prove (ii) for binary outcomes, where $\mu(d) = \Pr(Y=1 \mid D=d, S=e)$. Fix $d \in \{0,1\}$. By iterated expectations and $(S,D) \perp\!\!\!\perp R \mid Y$,

$$\begin{aligned} f_R(R \mid D=d, S=e) &= \mu(d) f_R(R \mid Y=1, D=d, S=e) + \{1-\mu(d)\} f_R(R \mid Y=0, D=d, S=e) \\ &= \mu(d) f_R(R \mid Y=1, S=o) + \{1-\mu(d)\} f_R(R \mid Y=0, S=o). \end{aligned}$$

As a consequence,

$$\begin{aligned} \widetilde{\mu}(d) &= \mathbb{E}\{m(R) \mid D=d, S=e\} \\ &= \mu(d) \mathbb{E}\{m(R) \mid Y=1, S=o\} + \{1-\mu(d)\} \mathbb{E}\{m(R) \mid Y=0, S=o\}. \end{aligned}$$

Taking differences,

$$\begin{aligned} \widetilde{\tau} &:= \widetilde{\mu}(1) - \widetilde{\mu}(0) \\ &= [\mathbb{E}\{m(R) \mid Y=1, S=o\} - \mathbb{E}\{m(R) \mid Y=0, S=o\}]\{\mu(1) - \mu(0)\} \\ &= [\mathbb{E}\{m(R) \mid Y=1, S=o\} - \mathbb{E}\{m(R) \mid Y=0, S=o\}]\tau \\ &= \frac{\mathrm{Cov}\{m(R), Y \mid S=o\}}{\mathrm{Var}(Y \mid S=o)} \tau \\ &= \frac{\mathrm{Var}\{m(R) \mid S=o\}}{\mathrm{Var}(Y \mid S=o)} \tau, \end{aligned}$$

yielding $\kappa := \frac{\mathrm{Var}\{m(R)|S=o\}}{\mathrm{Var}(Y|S=o)}$.

By the law of total variance,

$$\mathrm{Var}(Y \mid S=o) = \mathrm{Var}\{m(R) \mid S=o\} + \mathbb{E}\{\mathrm{Var}(Y \mid R, S=o) \mid S=o\},$$

so $\kappa \in [0,1]$. Moreover, $\kappa = 1$ if and only if $\mathbb{E}\{\mathrm{Var}(Y \mid R, S=o) \mid S=o\} = 0$, which is equivalent to $\mathrm{Var}(Y \mid R, S=o) = 0$ almost surely.

□

B.1.5 Proof of Proposition 2

We prove each part of the proposition separately. Since $m_d(R)$ is a bounded and measurable function, the expectation is well defined. As before, the assumptions imply $(S,D) \perp\!\!\!\perp R \mid Y$.

1. Fix $d \in \{0,1\}$ and $s \in \{e,o\}$. By iterated expectations, the fact that Y is binary, and $(S,D) \perp\!\!\!\perp R \mid Y$,

$$\begin{aligned}
&\mathbb{E}\{m_d(R) \mid D=d, S=s\} \\
&= \mathbb{E}[\mathbb{E}\{m_d(R) \mid Y, D=d, S=s\} \mid D=d, S=s] \\
&= \mathbb{E}\{m_d(R) \mid Y=1, D=d, S=s\}\mathbb{E}(Y \mid D=d, S=s) + \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=s\}\{1 - \mathbb{E}(Y \mid D=d, S=s)\} \\
&= \mathbb{E}\{m_d(R) \mid Y=1\}\mathbb{E}(Y \mid D=d, S=s) + \mathbb{E}\{m_d(R) \mid Y=0\}\{1 - \mathbb{E}(Y \mid D=d, S=s)\} \\
&= \mathbb{E}\{m_d(R) \mid Y=0\} + \kappa_d \mathbb{E}(Y \mid D=d, S=s),
\end{aligned}$$

where $\kappa_d = \mathbb{E}\{m_d(R) \mid Y=1\} - \mathbb{E}\{m_d(R) \mid Y=0\}$.

The first term does not depend on s. Therefore, substitution yields

$$\begin{aligned}
\tilde{\tau}^{\mathrm{PPI}} &= \mathbb{E}\{m_1(R) \mid D=1, S=e\} - \mathbb{E}\{m_0(R) \mid D=0, S=e\} \\
&\quad + \mathbb{E}\{Y - m_1(R) \mid D=1, S=o\} - \mathbb{E}\{Y - m_0(R) \mid D=0, S=o\} \\
&= \kappa_1 \mathbb{E}(Y \mid D=1, S=e) - \kappa_0 \mathbb{E}(Y \mid D=0, S=e) \\
&\quad + \mathbb{E}(Y \mid D=1, S=o) - \kappa_1 \mathbb{E}(Y \mid D=1, S=o) - \mathbb{E}(Y \mid D=0, S=o) + \kappa_0 \mathbb{E}(Y \mid D=0, S=o) \\
&= \tilde{\tau}^o + \kappa_1\{\mathbb{E}(Y \mid D=1, S=e) - \mathbb{E}(Y \mid D=1, S=o)\} - \kappa_0\{\mathbb{E}(Y \mid D=0, S=e) - \mathbb{E}(Y \mid D=0, S=o)\}.
\end{aligned}$$

By Assumption 1(i, ii), $\mathbb{E}\{Y(d)\,|\,S=e\}=\mathbb{E}(Y\,|\,D=d,S=e)$, so

$$\begin{aligned}\tau&=\mathbb{E}(Y\,|\,D=1,S=e)-\mathbb{E}(Y\,|\,D=0,S=e)\\ \tilde{\tau}^o&=\mathbb{E}(Y\,|\,D=1,S=o)-\mathbb{E}(Y\,|\,D=0,S=o)\\ \delta&=\mathbb{E}(Y\,|\,D=0,S=e)-\mathbb{E}(Y\,|\,D=0,S=o)\\ \tau-\tilde{\tau}^o+\delta&=\mathbb{E}(Y\,|\,D=1,S=e)-\mathbb{E}(Y\,|\,D=1,S=o).\end{aligned}$$

Substituting these identities gives

$$\begin{aligned}\tilde{\tau}^{\mathrm{PPI}}&=\tilde{\tau}^o+\kappa_1(\tau-\tilde{\tau}^o+\delta)-\kappa_0\delta\\ &=\tau+(1-\kappa_1)(\tilde{\tau}^o-\tau)+(\kappa_1-\kappa_0)\delta.\end{aligned}$$

2. Fix arbitrary measurable functions m_0,m_1. We construct a relevant R with $\kappa_0=\kappa_1=0$.

Let R be discrete with support $\{r_1,r_2,r_3,r_4\}$. Construct a nonzero vector $v=(v_1,...,v_4)$ such that

$$\sum_{j=1}^4 v_j=0,\qquad \sum_{j=1}^4 m_0(r_j)v_j=0,\qquad \sum_{j=1}^4 m_1(r_j)v_j=0.$$

Such vector always exists since we have three linear constraints and four degrees of freedom.

Choose $\varepsilon>0$ small enough that $\frac{1}{4}+\varepsilon v_j$ and $\frac{1}{4}-\varepsilon v_j$ are bounded away from zero and one for all j. For $s\in\{e,o\}$ and $d\in\{0,1\}$, set

$$\Pr(R=r_j\,|\,Y=1,D=d,S=s)=\frac{1}{4}+\varepsilon v_j,\quad \Pr(R=r_j\,|\,Y=0,D=d,S=s)=\frac{1}{4}-\varepsilon v_j.$$

Then $\Pr(R=r_j\mid Y=1,D=d,S=s)\neq\Pr(R=r_j\mid Y=0,D=d,S=s)$ (since $v\neq 0$), so $R\not\perp Y\,|\,D,S$. Moreover, for each $d\in\{0,1\}$,

$$\begin{aligned}\kappa_d&=\mathbb{E}\{m_d(R)\,|\,Y=1\}-\mathbb{E}\{m_d(R)\,|\,Y=0\}\\ &=\sum_{j=1}^4 m_d(r_j)\left\{\left(\frac{1}{4}+\varepsilon v_j\right)-\left(\frac{1}{4}-\varepsilon v_j\right)\right\}\\ &=2\varepsilon\sum_{j=1}^4 m_d(r_j)v_j\\ &=0\end{aligned}$$

where the last equality uses the construction of v.

By construction, this DGP satisfies stability, coverage, and no direct effects.

□

B.2 Proofs for Section 4

B.2.1 Proof of Proposition 3

To lighten notation, define $Y=\Delta^e$, $W=\Delta^o$, $U=Y-W^\top\theta$, $Z=\widetilde{H}(R)$, and $\widehat{Z}=\widehat{H}(R)$, so that

$$Y=W^\top\theta+U,\quad \mathbb{E}(U|R)=0,\quad \mathbb{E}(UZ)=0.$$

Note that $\|W\|_{\max}\le\bar{W}$ almost surely for some constant $\bar{W}$ due to the assumption that the marginal probabilities are bounded away from zero. Moreover, $\mathbb{E}(U^2|R)\le\bar{\sigma}_U^2$ for some constant $\bar{\sigma}_U^2<\infty$ for the same reason.

While proving this result, we use standard notation for cross-fitting. Let there be L folds, each denoted by I_ℓ with $\ell\in[L]$. Each fold contains $n_\ell=n/L$ observations. The complement of I_ℓ is $I_{-\ell}$. If $i\in I_\ell$, then $\widehat{Z}_i=\widehat{H}_\ell(R_i)$ is constructed from $\widehat{H}_\ell$ estimated on the remaining folds $I_{-\ell}$.

With cross fitting, the estimator in Algorithm 1 generalizes to

$$\widehat{\theta}:=\{\mathbb{E}_n(\widehat{Z}W^\top)\}^{-1}\mathbb{E}_n(\widehat{Z}Y)=\left(\frac{1}{L}\frac{1}{n_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\widehat{Z}_iW_i^\top\right)^{-1}\left(\frac{1}{L}\frac{1}{n_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\widehat{Z}_iY_i\right).$$

With known marginal probabilities, the argument uses standard techniques, similar to Mackey et al. (2018); Chen et al. (2020). In this lighter notation,

$$n^{1/2}(\widehat{\theta}-\theta)=n^{1/2}\Big[\{\mathbb{E}_n(\widehat{Z}W^\top)\}^{-1}\mathbb{E}_n(\widehat{Z}Y)-\{\mathbb{E}_n(\widehat{Z}W^\top)\}^{-1}\mathbb{E}_n(\widehat{Z}W^\top)\theta\Big]=\{\mathbb{E}_n(\widehat{Z}W^\top)\}^{-1}n^{1/2}\mathbb{E}_n(\widehat{Z}U).$$

We derive the probability limit of $\mathbb{E}_n(\widehat{Z}W^\top)$, then the Gaussian approximation of $n^{1/2}\mathbb{E}_n(\widehat{Z}U)$.

Lemma B.1. Under Proposition 3's conditions, $\mathbb{E}_n(\widehat{Z}W^\top)\xrightarrow{p}\mathbb{E}_n(ZW^\top)$.

Proof. We express the difference as

$$\mathbb{E}_n\{(\widehat{Z}-Z)W^\top\}=\frac{1}{L}\frac{1}{n_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}(\widehat{Z}_i-Z_i)W_i^\top=\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{i\in I_\ell}(\widehat{Z}_i-Z_i)W_i^\top.$$

Focusing on the foldwise quantity, we first control $\mathbb{E}\Big\{\Big|\frac{1}{n_\ell}\sum_{i\in I_\ell}(\widehat{Z}_{ij}-Z_{ij})W_{ik}\Big||I_{-\ell}\Big\}$. We can write

$$\begin{aligned}&\mathbb{E}\left\{\frac{1}{n_\ell}\sum_{i\in I_\ell}\Big|(\widehat{Z}_{ij}-Z_{ij})W_{ik}\Big||I_{-\ell}\right\}\le\mathbb{E}\left\{\frac{\bar{W}}{n_\ell}\cdot\sum_{i\in I_\ell}\Big|\widehat{Z}_{ij}-Z_{ij}\Big||I_{-\ell}\right\}\\&=\bar{W}\mathbb{E}\Big\{\Big|\widehat{Z}_{ij}-Z_{ij}\Big||I_{-\ell}\Big\}\le\bar{W}[\mathbb{E}\{(\widehat{Z}_{ij}-Z_{ij})^2|I_{-\ell}\}]^{1/2}\le\bar{W}\mathcal{R}(\widehat{Z})^{1/2}=o_p(1).\end{aligned}$$

We use $|W_{ij}|\le\bar{W}$ almost surely, since marginal probabilities are bounded away from zero. We also write $\mathcal{R}(\widehat{Z})=\mathbb{E}(\|\widehat{Z}_i-Z_i\|^2|\,I_{-\ell})=o_p(1)$ for the mean square limit.

Finally note that for any $\epsilon>0$, $\Pr(|\frac{1}{n_\ell}\sum_{i\in I_\ell}(\widehat{Z}_{ij}-Z_{ij})W_{ik}|>\epsilon|I_{-\ell})\leq \bar{W}\mathcal{R}(\hat{Z})^{1/2}/\epsilon\to_p 0$. Therefore by the dominated convergence theorem (since $\Pr(|\frac{1}{n_\ell}\sum_{i\in I_\ell}(\widehat{Z}_{ij}-Z_{ij})W_{ik}|>\epsilon|I_{-\ell})\in[0,1]$), it follows $\mathbb{E}[\Pr(|\frac{1}{n_\ell}\sum_{i\in I_\ell}(\widehat{Z}_{ij}-Z_{ij})W_{ik}|>\epsilon|I_{-\ell})]\to 0$. Applying this to each of the $\ell\in\{1,...,L\}$ the result follows directly from the union bound (since L is finite). □

Lemma B.2. Under Proposition 3's conditions, $\mathbb{E}_n(ZW^\top)\xrightarrow{p}\mathbb{E}(ZW^\top)$.

Proof. By Chebyshev's inequality in Frobenius norm, it suffices to bound $\mathbb{V}(Z_{ij}W_{ik})\leq \mathbb{E}(Z_{ij}^2W_{ik}^2)\leq \bar{W}^2\mathbb{E}(\|Z_i\|^2)$. In summary, we use $\|W_i\|_{\max}\leq\bar{W}$ almost surely since marginal probabilities are bounded away from zero and $\mathbb{E}(\|Z_i\|^2)<\infty$ due to Assumption 4. □

Lemma B.3. Under Proposition 3's conditions, $n^{1/2}\mathbb{E}_n(\widehat{Z}U)\xrightarrow{p}n^{1/2}\mathbb{E}_n(ZU)$.

Proof. We express the difference as

$$n^{1/2}\mathbb{E}_n\{U(\widehat{Z}-Z)\}=n^{1/2}\frac{1}{L}\frac{1}{n_\ell}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}U_i(\widehat{Z}_i-Z_i)=L^{1/2}\frac{1}{L}\sum_{\ell=1}^{L}n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell}U_i(\widehat{Z}_i-Z_i).$$

Focusing on the foldwise quantity, it suffices to control

$$\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell}U_i(\widehat{Z}_{ik}-Z_{ik})\right\}^2\right]=\mathbb{E}\left(\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell}U_i(\widehat{Z}_{ik}-Z_{ik})\right\}^2|\,I_{-\ell}\right]\right).$$

Due to cross-fitting, the inner expectation is

$$\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell}U_i(\widehat{Z}_{ik}-Z_{ik})\right\}^2|\,I_{-\ell}\right]=\frac{1}{n_\ell}\mathbb{E}\left\{\sum_{i,j\in I_\ell}U_i(\widehat{Z}_{ik}-Z_{ik})U_j(\widehat{Z}_{jk}-Z_{jk})\,|\,I_{-\ell}\right\}$$
$$=\frac{1}{n_\ell}\mathbb{E}\left\{\sum_{i\in I_\ell}U_i^2(\widehat{Z}_{ik}-Z_{ik})^2\,|\,I_{-\ell}\right\}=\mathbb{E}\{U_i^2(\widehat{Z}_{ik}-Z_{ik})^2\,|\,I_{-\ell}\}=\mathbb{E}\{\mathbb{E}(U_i^2|R_i,I_{-\ell})(\widehat{Z}_{ik}-Z_{ik})^2\,|\,I_{-\ell}\}$$
$$\leq\bar{\sigma}_U^2\mathbb{E}\{(\widehat{Z}_{ik}-Z_{ik})^2\,|\,I_{-\ell}\}\leq\bar{\sigma}_U^2\mathcal{R}(\widehat{Z})=o_p(1).$$

In the first inequality, we use $\mathbb{E}(U_i^2|R_i,I_{-\ell})=\mathbb{E}(U_i^2|R_i)\leq\bar{\sigma}_U^2$, where $\bar{\sigma}_U^2<\infty$ under our assumptions. In the second inequality, we write $\mathcal{R}(\widehat{Z})=\mathbb{E}(\|\widehat{Z}_i-Z_i\|^2|\,I_{-\ell})=o_p(1)$ for the mean square limit. □

Lemma B.4. Under Proposition 3's conditions, $n^{1/2}\mathbb{E}_n(ZU)\rightsquigarrow\mathcal{N}\{0,\mathbb{E}(U^2ZZ^\top)\}$.

Proof. By Theorem 1, $\mathbb{E}(U_iZ_i)=\mathbb{E}\{\mathbb{E}(U_i|R_i)Z_i\}=0$. Moreover, $\mathbb{E}(U_i^2Z_iZ_i^\top)=\mathbb{E}\{\mathbb{E}(U_i^2|R_i)Z_iZ_i^\top\}\leq\bar{\sigma}_U^2\mathbb{E}(Z_iZ_i^\top)$ since $\mathbb{E}(U_i^2|R_i)\leq\bar{\sigma}_U^2$ under our assumptions. Here, $\mathbb{E}(Z_iZ_i^\top)$ is finite by Assumption 4, so we apply the Lindeberg-Levy central limit theorem. □

We are now ready to prove Proposition 3. Recall that

$$n^{1/2}(\widehat{\theta}-\theta)=\{\mathbb{E}_n(\widehat{Z}W^\top)\}^{-1}n^{1/2}\mathbb{E}_n(\widehat{Z}U).$$

By the continuous mapping theorem, Lemmas B.1 and B.2 imply $\mathbb{E}_n(\widehat{Z}W^\top)\xrightarrow{p}\mathbb{E}(ZW^\top)$. By Slutsky's theorem, Lemmas B.3 and B.4 imply $n^{1/2}\mathbb{E}_n(\widehat{Z}U)\rightsquigarrow\mathcal{N}\{0,\mathbb{E}(U^2ZZ^\top)\}$. Overall, by Slutsky's theorem, we conclude that $n^{1/2}(\widehat{\theta}-\theta)\rightsquigarrow\mathcal{N}[0,\{\mathbb{E}(ZW^\top)\}^{-1}\mathbb{E}(U^2ZZ^\top)\{\mathbb{E}(WZ^\top)\}^{-1}]$.

Finally, recall that for $\lambda=(y_1-y_K,\ldots,y_{K-1}-y_K)^\top$, we have $\widehat{\tau}=\lambda^\top\widehat{\theta}$ in Algorithm 1 and $\tau=\lambda^\top\theta$ by definition. Therefore, $n^{1/2}(\widehat{\tau}-\tau)=n^{1/2}\lambda^\top(\widehat{\theta}-\theta)$ and we appeal to the delta method.

Efficiency follows directly from Chamberlain (1987) and Newey (1993). Specifically, $\widehat{\theta}$ is semiparametrically efficient for θ within the class of estimators based upon the conditional moment $\mathbb{E}\{\Delta^e-(\Delta^o)^\top\theta|R\}=0$. Therefore, $\widehat{\tau}$ is semiparametrically efficient for τ, since they are known linear transformations of $\widehat{\theta}$ and θ, respectively. □

B.2.2 Proof of Proposition 4

With unknown marginal probabilities, some extra care is required.

Extending the notation from the proof of Proposition 3, we let $\widehat{Y}=\widehat{\Delta}^e$, $\widehat{W}=\widehat{\Delta}^o$, and $\widehat{U}=\widehat{Y}-\widehat{W}\theta$. If $i\in I_\ell$, then the marginal probabilities in $\widehat{Y}_i$ and $\widehat{W}_i$ are estimated on the same fold I_ℓ. Meanwhile, $\widehat{Z}_i=\widehat{H}_\ell(R_i)$ is constructed from $\widehat{H}_\ell$ estimated on the remaining folds $I_{-\ell}$.

The generalized decomposition is

$$n^{1/2}(\widehat{\theta}-\theta)=n^{1/2}\left[\{\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)\}^{-1}\mathbb{E}_n(\widehat{Z}\widehat{Y})-\{\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)\}^{-1}\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)\theta\right]=\{\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)\}^{-1}n^{1/2}\mathbb{E}_n(\widehat{Z}\widehat{U}).$$

We derive the probability limit of $\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)$, then the Gaussian approximation of $n^{1/2}\mathbb{E}_n(\widehat{Z}\widehat{U})$.

Lemma B.5. Under Proposition 4's conditions, $\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)\xrightarrow{p}\mathbb{E}_n(Z\widehat{W}^\top)$.

Proof. The argument is similar to Lemma B.1, using $|\widehat{W}_{ij}|\leq\bar{W}'$ almost surely, which follows from our assumptions. □

Lemma B.6. Under Proposition 4's conditions, $\mathbb{E}_n(Z\widehat{W}^\top)\xrightarrow{p}\mathbb{E}_n(ZW^\top)$.

Proof. We express the difference as

$$\left\|\mathbb{E}_n\{Z(\widehat{W}-W)^\top\}\right\|_F\leq\left\{\mathbb{E}_n(\|\widehat{W}-W\|^2)\right\}^{1/2}\left\{\mathbb{E}_n(\|Z\|^2)\right\}^{1/2},$$

where $||\cdot||_F$ indicates the Frobenius norm. Since $\mathbb{E}_n(\|Z\|^2)\xrightarrow{p}\mathbb{E}(\|Z\|^2)$ when $\mathbb{E}(\|Z\|^2)<\infty$ by the weak law of large numbers, it suffices to study

$$\mathbb{E}_n(\|\widehat{W}-W\|^2)=\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{i\in I_\ell}\|\widehat{W}_i-W_i\|^2=\frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_\ell(\|\widehat{W}-W\|^2),\quad\mathbb{E}_\ell(\cdot)=\frac{1}{n_\ell}\sum_{i\in I_\ell}(\cdot).$$

Unpacking the notation of $\mathbb{E}_\ell(\|\widehat{W}-W\|^2)$,

$$\widehat{W}_k - W_k = \widehat{\Delta}_k^o - \Delta_k^o = \frac{1_{Y=y_k,S=o}}{\mathbb{E}_\ell(1_{Y=y_k,S=o})} - \frac{1_{Y=y_K,S=o}}{\mathbb{E}_\ell(1_{Y=y_K,S=o})} - \left\{ \frac{1_{Y=y_k,S=o}}{\mathbb{E}(1_{Y=y_k,S=o})} - \frac{1_{Y=y_K,S=o}}{\mathbb{E}(1_{Y=y_K,S=o})} \right\}$$
$$= 1_{Y=y_k,S=o}\left\{ \frac{1}{\mathbb{E}_\ell(1_{Y=y_k,S=o})} - \frac{1}{\mathbb{E}(1_{Y=y_k,S=o})} \right\} - 1_{Y=y_K,S=o}\left\{ \frac{1}{\mathbb{E}_\ell(1_{Y=y_K,S=o})} - \frac{1}{\mathbb{E}(1_{Y=y_K,S=o})} \right\}$$
$$= 1_{Y=y_k,S=o}\left\{ \frac{\mathbb{E}(1_{Y=y_k,S=o}) - \mathbb{E}_\ell(1_{Y=y_k,S=o})}{\mathbb{E}_\ell(1_{Y=y_k,S=o})\mathbb{E}(1_{Y=y_k,S=o})} \right\} - 1_{Y=y_K,S=o}\left\{ \frac{\mathbb{E}(1_{Y=y_K,S=o}) - \mathbb{E}_\ell(1_{Y=y_K,S=o})}{\mathbb{E}_\ell(1_{Y=y_K,S=o})\mathbb{E}(1_{Y=y_K,S=o})} \right\}.$$

With population and empirical counts bounded away from zero,

$$|\widehat{W}_k - W_k| \lesssim |\mathbb{E}(1_{Y=y_k,S=o}) - \mathbb{E}_\ell(1_{Y=y_k,S=o})| + |\mathbb{E}(1_{Y=y_K,S=o}) - \mathbb{E}_\ell(1_{Y=y_K,S=o})|.$$

By Hoeffding's inequality and the union bound, with probability $1-\delta$,

$$|\mathbb{E}(1_{Y=y_k,S=o}) - \mathbb{E}_\ell(1_{Y=y_k,S=o})| \leq \left\{ \frac{\ln(2/\delta)}{2n_\ell} \right\}^{1/2}.$$

Therefore with probability $1-\delta$ for all $i \in [n]$ and $k \in [K]$ simultaneously,

$$|\widehat{W}_{ik} - W_{ik}| \lesssim 2\left\{ \frac{\ln(2K/\delta)}{2n_\ell} \right\}^{1/2} \lesssim \frac{\ln(2K/\delta)^{1/2}}{n_\ell^{1/2}}.$$

We conclude that, under this union event that holds with probability $1-\delta$,

$$\mathbb{E}_\ell(\|\widehat{W}-W\|^2) \lesssim \mathbb{E}_\ell\left\{ K \frac{\ln(2K/\delta)}{n_\ell} \right\} = K\frac{\ln(2K/\delta)}{n_\ell}.$$

For all $\delta > 0$ and fixed K, this quantity vanishes for large $n_\ell = n/L$, so $\mathbb{E}_\ell(\|\widehat{W}-W\|^2) = o_p(1)$. The continuous mapping theorem implies the desired result. □

Lemma B.7. Under Proposition 4's conditions, $\mathbb{E}_n(ZW^\top) \xrightarrow{p} \mathbb{E}(ZW^\top)$.

Proof. The argument is identical to Lemma B.2. □

Lemma B.8. Under Proposition 4's conditions, $n^{1/2}\mathbb{E}_n(\widehat{Z}\widehat{U}) \xrightarrow{p} n^{1/2}\mathbb{E}_n(Z\widehat{U})$.

Proof. The argument is similar to Lemma B.3, using up front that $|\widehat{U}_i| \leq \bar{U}'$ almost surely, which follows from our assumptions. □

Lemma B.9. Under Proposition 4's conditions, $n_\ell^{1/2}\mathbb{E}_\ell(Z\widehat{U}) \rightsquigarrow \mathcal{N}(0,V)$ where $V = v^\top \Sigma v$,

$\Sigma_{ij} = \mathrm{Cov}(C_i, C_j)$, and

$$C = \begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ \vdots \\ C_{2K+1} \\ C_{2K+2} \\ C_{2K+3} \\ C_{2K+4} \end{pmatrix} = \begin{pmatrix} 1_{D=1,S=e} \\ 1_{D=1,S=e}Z \\ 1_{D=0,S=e} \\ 1_{D=0,S=e}Z \\ 1_{Y=y_1,S=o} \\ 1_{Y=y_1,S=o}Z \\ \vdots \\ 1_{Y=y_{K-1},S=o} \\ 1_{Y=y_{K-1},S=o}Z \\ 1_{Y=y_K,S=o} \\ 1_{Y=y_K,S=o}Z \end{pmatrix}, \quad v = \begin{pmatrix} -\mathbb{E}(C_2)^\top \mathbb{E}(C_1)^{-2} \\ \mathbb{E}(C_1)^{-1} I_{K-1} \\ \mathbb{E}(C_4)^\top \mathbb{E}(C_3)^{-2} \\ -\mathbb{E}(C_3)^{-1} I_{K-1} \\ \theta_1 \mathbb{E}(C_6)^\top \mathbb{E}(C_5)^{-2} \\ -\theta_1 \mathbb{E}(C_5)^{-1} I_{K-1} \\ \vdots \\ \theta_{K-1} \mathbb{E}(C_{2K+2})^\top \mathbb{E}(C_{2K+1})^{-2} \\ -\theta_{K-1} \mathbb{E}(C_{2K+1})^{-1} I_{K-1} \\ -\left(\sum_{k=1}^{K-1} \theta_k\right) \mathbb{E}(C_{2K+4})^\top \mathbb{E}(C_{2K+3})^{-2} \\ \left(\sum_{k=1}^{K-1} \theta_k\right) \mathbb{E}(C_{2K+3})^{-1} I_{K-1} \end{pmatrix},$$

where the odd entries are scalars and the even entries are vectors in $\mathbb{R}^{K-1}$.

Proof. We unpack the definition of $\widehat{U}$:

$$\begin{aligned} \widehat{U} &= \widehat{Y} - \widehat{W}^\top \theta \\ &= \widehat{\Delta}^e - (\widehat{\Delta}^o)^\top \theta \\ &= \frac{1_{D=1,S=e}}{\mathbb{E}_\ell(1_{D=1,S=e})} - \frac{1_{D=0,S=e}}{\mathbb{E}_\ell(1_{D=0,S=e})} - \sum_{k=1}^{K-1} \left\{ \frac{1_{Y=y_k,S=o}}{\mathbb{E}_\ell(1_{Y=y_k,S=o})} - \frac{1_{Y=y_K,S=o}}{\mathbb{E}_\ell(1_{Y=y_K,S=o})} \right\} \theta_k. \end{aligned}$$

Therefore, for $h(C) = \frac{C_2}{C_1} - \frac{C_4}{C_3} - \sum_{k=1}^{K-1} \left(\frac{C_{2k+4}}{C_{2k+3}} - \frac{C_{2K+4}}{C_{2K+3}} \right) \theta_k$,

$$\begin{aligned} n_\ell^{1/2} \mathbb{E}_\ell(\widehat{U} Z) &= n_\ell^{1/2} \left[\frac{\mathbb{E}_\ell(1_{D=1,S=e}Z)}{\mathbb{E}_\ell(1_{D=1,S=e})} - \frac{\mathbb{E}_\ell(1_{D=0,S=e}Z)}{\mathbb{E}_\ell(1_{D=0,S=e})} - \sum_{k=1}^{K-1} \left\{ \frac{\mathbb{E}_\ell(1_{Y=y_k,S=o}Z)}{\mathbb{E}_\ell(1_{Y=y_k,S=o})} - \frac{\mathbb{E}_\ell(1_{Y=y_K,S=o}Z)}{\mathbb{E}_\ell(1_{Y=y_K,S=o})} \right\} \theta_k \right] \\ &= n_\ell^{1/2} h\{\mathbb{E}_\ell(C)\}. \end{aligned}$$

We make three observations. First, $\mathbb{E}(\|Z\|^2) < \infty$ by Assumption 4, so by the Lindeberg-Levy central limit theorem, $n_\ell^{1/2}\{\mathbb{E}_\ell(C) - \mathbb{E}(C)\} \rightsquigarrow \mathcal{N}(0, \Sigma)$.

Second, by the conditional moment equation,

$$\begin{aligned} h\{\mathbb{E}(C)\} &= \frac{\mathbb{E}(1_{D=1,S=e}Z)}{\mathbb{E}(1_{D=1,S=e})} - \frac{\mathbb{E}(1_{D=0,S=e}Z)}{\mathbb{E}(1_{D=0,S=e})} - \sum_{k=1}^{K-1} \left\{ \frac{\mathbb{E}(1_{Y=y_k,S=o}Z)}{\mathbb{E}(1_{Y=y_k,S=o})} - \frac{\mathbb{E}(1_{Y=y_K,S=o}Z)}{\mathbb{E}(1_{Y=y_K,S=o})} \right\} \theta_k \\ &= \mathbb{E}[\{\Delta^e - (\Delta^o)^\top \theta\} Z] = \mathbb{E}[\mathbb{E}\{\Delta^e - (\Delta^o)^\top \theta | R\} Z] = 0. \end{aligned}$$

Third, the Jacobian is

$$\nabla h(C) = \begin{pmatrix} -C_2^\top C_1^{-2} \\ C_1^{-1} I_{K-1} \\ C_4^\top C_3^{-2} \\ -C_3^{-1} I_{K-1} \\ \theta_1 C_6^\top C_5^{-2} \\ -\theta_1 C_5^{-1} I_{K-1} \\ \vdots \\ \theta_{K-1} C_{2K+2}^\top C_{2K+1}^{-2} \\ -\theta_{K-1} C_{2K+1}^{-1} I_{K-1} \\ -\left(\sum_{k=1}^{K-1} \theta_k\right) C_{2K+4}^\top C_{2K+3}^{-2} \\ \left(\sum_{k=1}^{K-1} \theta_k\right) C_{2K+3}^{-1} I_{K-1} \end{pmatrix}.$$

Therefore, by the delta method,

$$n_\ell^{1/2} \mathbb{E}_\ell(Z\widehat{U}) = n_\ell^{1/2} [h\{\mathbb{E}_\ell(C)\} - h\{\mathbb{E}(C)\}] \rightsquigarrow \mathcal{N}(0, [\nabla h\{\mathbb{E}(C)\}]^\top \Sigma [\nabla h\{\mathbb{E}(C)\}]).$$

□

We are now ready to prove Proposition 4. Recall that

$$n^{1/2}(\widehat{\theta} - \theta) = \{\mathbb{E}_n(\widehat{Z}\widehat{W}^\top)\}^{-1} n^{1/2} \mathbb{E}_n(\widehat{Z}\widehat{U}).$$

By the continuous mapping theorem, Lemmas B.5, B.6, and B.7 imply $\mathbb{E}_n(\widehat{Z}\widehat{W}^\top) \xrightarrow{p} \mathbb{E}(ZW^\top)$. By Slutsky's theorem, Lemmas B.8 and B.9 imply $n^{1/2}\mathbb{E}_n(\widehat{Z}\widehat{U}) \rightsquigarrow \mathcal{N}(0, V)$. Overall, by Slutsky's theorem, we conclude that $n^{1/2}(\widehat{\theta} - \theta) \rightsquigarrow \mathcal{N}[0, \{\mathbb{E}(ZW^\top)\}^{-1} V \{\mathbb{E}(WZ^\top)\}^{-1}]$. The rest of the argument is identical to Proposition 3.

□

C Additional Theoretical Results and Discussion

C.1 Bias of Alternative Approaches: Extensions

In Proposition 1, we characterized the bias of a common two-step method when there are incomplete observational cases (Assumption 3(ii)). In this appendix, we extend our characterization of the two-step method's bias in two ways: (i) complete observational cases, and (ii) settings where units are randomly assigned to the experimental or observational sample.

C.1.1 Bias of Common Practice with Complete Observational Cases

Consider the setting of Assumption 3(i) with complete observational cases. For simplicity, suppose there are no pretreatment covariates. The first step of the method is now to train a predictor $m_d(R)$ of the outcome Y from the remotely sensed variable R and treatment D in the observational sample. The method implicitly targets the estimand

$$\widetilde{\tau}' = \widetilde{\mu}'(1) - \widetilde{\mu}'(0) = \mathbb{E}\{m_1(R) \mid D=1, S=e\} - \mathbb{E}\{m_0(R) \mid D=0, S=e\}, \quad m_d(R) := \mathbb{E}(Y \mid R, D=d, S=o).$$

Proposition C.1 (Bias with complete observational cases)**.** Suppose Assumptions 1, 2, and 3(i) hold, with binary outcome and no covariates. Then, we have that

i. $\widetilde{\mu}'(d) = \mu(d) + \mathbb{E}[Y\{v_d(R,Y) - 1\} \mid D=d, S=e]$, where $v_d(r,y) := \frac{\Pr(Y=y \mid D=d, S=o)}{\Pr(Y=y \mid D=d, S=e)} \frac{f_R(r \mid D=d, S=e)}{f_R(r \mid D=d, S=o)}$;

ii. there exists $\kappa_d \in [0,1)$ such that $\widetilde{\mu}'(d) = \kappa_d \mu(d) + (1-\kappa_d)\mathbb{E}(Y \mid D=d, S=o)$, as long as R is an imperfect predictor of Y in the treatment arm $D=d$, i.e. as long as $\mathrm{Var}(Y \mid R, D=d, S=o) > 0$ almost surely.

Proof. See Appendix D.1.1. □

The common practice is biased towards the conditional expectation of the outcome in the observational sample.

C.1.2 Bias of Common Practice with Random Sample Selection

We now verify that the two-step method is biased, even when the researcher randomly assigns units to the experimental and observational samples. As in Proposition 1, consider the setting with binary outcome, no covariates, and incomplete observational cases (Assumption 3(ii)).

In this thought experiment, the researcher implements the following study. First, they randomize units into the experimental or observational sample. Then, they randomize experimental units to treatment or control. No observational units are treated.

Proposition C.2 (Bias with random sample selection)**.** Suppose the conditions of Proposition 1 hold. Further assume $S \perp\!\!\!\perp \{Y(0), Y(1)\}$. Then, we have that:

i. $\widetilde{\mu}(0) = \mu(0)$ and $\widetilde{\mu}(1) = \mu(1) + \mathbb{E}\left[Y\{w_1(R,Y) - 1\} \mid D=1, S=e\right]$, where $w_1(r,y) := \frac{\Pr\{Y(0)=y \mid S=e\}}{\Pr\{Y(1)=y \mid S=e\}} \frac{f_R(r \mid D=1, S=e)}{f_R(r \mid D=0, S=e)}$.

ii. there exists a scalar $\kappa \in [0,1)$ such that $\widetilde{\mu}(1) = \kappa\mu(1) + (1-\kappa)\mu(0)$, as long as R is an imperfect predictor of Y, i.e. as long as $\mathrm{Var}(Y \mid R, S=o) > 0$ almost surely.

Proof. See Appendix D.1.2. □

Even when units are randomly assigned to the experimental and observational samples, the two step method yields an estimate of the average treated potential outcome that is biased towards the average untreated potential outcome. As such, it attenuates the treatment effect.

C.2 Quasi-Experiments with Remotely Sensed Outcomes

In this section, we describe how our main identifying assumption (Assumption 2(i)) can be used to identify treatment effects in quasi-experimental samples via data combination with the observational sample. We discuss two quasi-experimental strategies: instrumental variables and difference-in-differences. To ease exposition, we discuss Example 3; we consider a binary outcome, omit pre-treatment covariates, and focus on "incomplete" observational cases (Assumption 3(ii)). The generalization to a discrete outcome with discrete covariates, as in Section 3 of the main text, is straightforward.

C.2.1 Instrumental Variables

Each unit is now characterized by the random vector

$$\{S,Z,D(0),D(1),Y(0,0),Y(0,1),Y(1,0),Y(1,1),R\},$$

where $Z \in \{0,1\}$ is a binary instrument, $D(z)$ are potential treatments and $Y(d,z)$ are potential outcomes, following Imbens and Angrist (1994). For units in the quasi-experimental sample ($S=e$), we observe (Z,D,R). For units in the observational sample ($S=o$), we observe (Y,R) since we focus on incomplete observational cases.

We modify our previous assumption about the experimental sample to accommodate the instrument. This sample may be viewed as a quasi-experiment, or as an experiment with imperfect compliance.

Assumption C.1 (Instrumental variable)**.** Suppose the following:

i. Instrument exclusion: for all $z \in \{0,1\}$ and $d \in \{0,1\}$, $Y(d,z) := Y(d)$ almost surely.

ii. SUTVA: $D = ZD(1)+(1-Z)D(0)$ and $Y = DY(1)+(1-D)Y(0)$ almost surely.

iii. Instrument randomization: $Z \perp\!\!\!\perp \{D(0),D(1),Y(0),Y(1)\} \mid S=e$.

iv. Overlap: $\Pr(Z=1 \mid S=e)$ and $P(Y=1 \mid S=o)$ are bounded away from zero and one almost surely.

v. Monotonicity: $\Pr\{D(1) \geq D(0) \,|\, S=e\} = 1$ and $\Pr\{D(1) > D(0) \,|\, S=e\} > 0$.

Under Assumption C.1, if we were to observe the outcome in the quasi-experimental sample, then the local average treatment effect (LATE) $\theta^{LATE} := \mathbb{E}\{Y(1)-Y(0) \,|\, D(1) > D(0), S=e\}$ could be identified using standard arguments. Since the outcome is unobserved in the quasi-experimental sample, we will use the observational sample as in the main text.

We next modify our stability and observational completeness assumptions to accommodate the instrument.

Assumption C.2 (Stability with an instrument)**.** Suppose Assumption 2 holds replacing D with (Z,D).

Assumption C.3 (No direct effects with an instrument)**.** Suppose Assumption 3(ii) holds replacing D with (Z,D).

Assumption C.2 remains the main assumption of our framework: $S \perp\!\!\!\perp R \,|\, Z,D,Y$. In the presence of an instrument, stability now requires that the conditional distribution of the remotely sensed variable R given (Z,D,Y) is stable across the quasi-experimental and observational samples. Analogously, Assumption C.3 becomes $Z,D \perp\!\!\!\perp R \,|\, Y$. It now requires that the instrument Z and the treatment D only affect the remotely sensed variable R via their effect on the outcome Y. Together Assumption C.2 and Assumption C.3 imply that $(S,D,Z) \perp\!\!\!\perp R \,|\, Y$, which is a testable implication as before.

These conditions allow us to identify θ^{LATE} by combining the quasi-experimental and observational samples. As notation, let $\alpha(z) = \mathbb{E}(Y \,|\, S=e, Z=z)$, and $\beta(z) := \mathbb{E}(D \,|\, S=e, Z=z)$. By standard arguments, under Assumption C.1, $\theta^{LATE} = \frac{\alpha(1)-\alpha(0)}{\beta(1)-\beta(0)}$. Of course, $\beta(0)$ and $\beta(1)$ are identified from the quasi-experimental sample since (Z,D) are observed. We will therefore identify $\alpha(0)$ and $\alpha(1)$ by combining the quasi-experimental and observational samples under Assumptions C.2 and C.3.

As a stepping stone, we first identify $\alpha(d,z) := \mathbb{E}(Y \,|\, S=e, D=d, Z=z)$.

Theorem C.1 (Identification with an instrument)**.** Suppose Assumptions C.1, C.2, and C.3 hold. Then, for any $z \in \{0,1\}$ and $d \in \{0,1\}$ satisfying $\Pr(D=d \,|\, Z=z, S=e) > 0$,

$$\mathbb{E}[\Delta^e(d,z) - \alpha(d,z)\Delta^o | R] = 0 \text{ almost surely,}$$

where $\Delta^e(d,z) := \frac{1\{D=d,Z=z,S=e\}}{\Pr(D=d,Z=z,S=e)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}$ and $\Delta^o := \frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}$.

Proof. See Appendix D.2.1. □

Corollary C.1 (Representation with an instrument)**.** Under Theorem C.1's conditions, for any $z \in \{0,1\}$, $d \in \{0,1\}$ satisfying $\Pr(D=d \mid Z=z, S=e) > 0$ and any representation $H(R)$ with $\mathbb{E}\{H(R)\Delta^o\} \neq 0$,

$$\alpha(d,z) = \frac{\mathbb{E}\{H(R)\Delta^e(d,z)\}}{\mathbb{E}\{H(R)\Delta^o\}}.$$

Theorem C.1 and Corollary C.1 immediately imply that $\alpha(0)$ and $\alpha(1)$ are identified; for $z \in \{0,1\}$, the law of iterated expectations gives $\alpha(z) = \alpha(1,z)\beta(z) + \alpha(0,z)\{1-\beta(z)\}$. Therefore, θ^{LATE} is also identified. Estimation and inference can then follow by suitably stacking moments and extending the discussion provided in Appendix C.7.2.

C.2.2 Two-Period Difference-in-Differences

Next, we consider a setting with two periods $t \in \{1,2\}$. Treated units ($D=1$) receive a treatment between period $t=1$ and $t=2$, and untreated units ($D=0$) remain untreated in both periods. Each unit is characterized by the random vector

$$\{S, D, Y_1(0), Y_1(1), Y_2(0), Y_2(1), R_1, R_2\}.$$

For units in the quasi-experimental sample ($S=e$), we observe (D, R_1, R_2), For units in the observational sample ($S=o$), we observe (Y_1, Y_2, R_1, R_2).

We again modify our previous assumption about the experimental sample, this time to accommodate the panel structure.

Assumption C.4 (Difference-in-differences)**.** Suppose the following:

i. SUTVA: For $t \in \{1,2\}$, $Y_t = DY_t(1) + (1-D)Y_t(0)$.

ii. Overlap: $\Pr(D=1 \mid S=e)$ and $\Pr(Y_t=1 \mid S=o), t \in \{1,2\}$ are bounded away from zero and one.

iii. Parallel trends: $\mathbb{E}\{Y_2(0) - Y_1(0) \mid S=e, D=1\} = \mathbb{E}\{Y_2(0) - Y_1(0) \mid S=e, D=0\}$.

iv. No anticipation: $\Pr\{Y_1(0) = Y_1(1) \mid S=e, D=1\} = 1$.

Under Assumption C.4, if we were to observe the outcomes in the quasi-experimental sample, then the average treatment effect on the treated $\theta^{ATT} := \mathbb{E}\{Y_2(1) - Y_2(0) \mid S=e, D=1\}$ could be identified using standard arguments. Assumption C.4 is akin to a parallel trends-type assumption on the cumulative distribution functions for untreated potential outcomes (Roth and Sant'Anna, 2023).

We modify our stability and observational completeness assumptions to accommodate the panel setting.

Assumption C.5 (Stability for difference-in-differences)**.** Suppose Assumption 2 holds for each time period $t \in \{1,2\}$.

Assumption C.6 (No direct effects for difference-in-differences)**.** Suppose Assumption 3(ii) holds for each time period $t \in \{1,2\}$.

In other words, we impose the same assumptions as in the main text, period by period. The main conditions are $S \perp\!\!\!\perp R_t \,|\, D, Y_t$ and $D \perp\!\!\!\perp R_t \,|\, Y_t$. Together, Assumption C.5 and Assumption C.6 imply that, for each $t \in \{1,2\}$, $(S,D) \perp\!\!\!\perp R_t \,|\, Y_t$, which is a testable implication as before.

These conditions allow us to identify θ^{ATT} by combining the quasi-experimental and observational samples. As notation, let $\alpha_t(d) = \mathbb{E}(Y_t \,|\, S=e, D=d)$. By standard arguments, under Assumption C.4, $\theta^{ATT} = \{\alpha_2(1) - \alpha_1(1)\} - \{\alpha_2(0) - \alpha_1(0)\}$. We will identify $\alpha_t(d)$ for $t \in \{1,2\}$ and $d \in \{0,1\}$ by combining the quasi-experimental and observational samples, under Assumption C.5 and Assumption C.6.

Theorem C.2 (Identification for difference-in-differences)**.** Suppose Assumptions C.4, C.5 and C.6 hold. Then, for any $t \in \{1,2\}$ and $d \in \{0,1\}$,

$$\mathbb{E}\{\Delta_t^e(d) - \Delta_t^o \alpha_t(d) \,|\, R_t\} = 0 \text{ almost surely,}$$

where $\Delta_t^e(d) := \frac{1\{D=d, S=e\}}{\Pr(D=d, S=e)} - \frac{1\{Y_t=0, S=o\}}{\Pr(Y_t=0, S=o)}$ and $\Delta_t^o := \frac{1\{Y_t=1, S=o\}}{\Pr(Y_t=1, S=o)} - \frac{1\{Y_t=0, S=o\}}{\Pr(Y_t=0, S=o)}$.

Proof. See Appendix D.2.2. □

Corollary C.2. Under Theorem C.2's conditions, for any $t \in \{1,2\}$, $d \in \{0,1\}$ and representation $H_t(R_t)$ with $\mathbb{E}\{H_t(R_t)\Delta_t^o\} \neq 0$,

$$\alpha_t(d) = \frac{\mathbb{E}\{H_t(R_t)\Delta_t^e(d)\}}{\mathbb{E}\{H_t(R_t)\Delta_t^o\}}.$$

Theorem C.2 and Corollary C.2 imply that $\alpha_t(d)$ are identified. Therefore, θ^{ATT} is also identified. Estimation and inference can again follow by suitably stacking moments and extending the discussion provided in Appendix C.7.2.

Our two-period difference-in-differences identification result extends to a setting with multiple periods and a common treatment adoption date. For each period, we apply the same argument to identify $\alpha_t(d) = \mathbb{E}(Y_t \,|\, S=e, D=d)$ from the remotely sensed variable R_t, and

then form the desired multi-period contrast by differencing these identified means across time periods and groups. In a staggered-adoption design (e.g., cohorts G with $D_t = 1\{t \geq G\}$), the same period-by-period argument identifies the means $\mathbb{E}(Y_t \mid S = e, D_t = d)$. We can then substitute these identified means into staggered difference-in-differences aggregations, such as de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2021), built from the corresponding period-by-period contrasts.

C.3 Some Spillovers

C.3.1 Generalized Potential Outcomes

With spillover effects, we write the potential outcome for unit i as $Y_i(d_i, \mathbf{d}_{-i})$. The potential outcome is indexed by the unit of interest's treatment assignment $d_i \in \{0,1\}$, as well as the vector of other units' treatment assignments $\mathbf{d}_{-i} \in \{0,1\}^{n-1}$.

From this potential outcome, we define the direct potential outcome by taking the expectation over other units' treatment assignments: $Y_i^{\text{dir}}(d_i) = \mathbb{E}_{\mathbf{D}_{-i}}\{Y_i(d_i, \mathbf{D}_{-i}) | D_i = d_i, S_i = e\}$. Note that it remains random due to unobserved heterogeneity.[25] The direct potential outcome may be viewed as design-based, since the expectation is over the design-induced distribution of $\mathbf{D}_{-i}$ conditional on $D_i = d$ in the experiment.

From direct potential outcomes, we define the average direct treatment effect. We relax the assumption of identical distribution across observations that is maintained throughout the main text. As such, it is a sample average direct treatment effect, though we omit "sample" for brevity.

Definition C.1 (Average direct treatment effect)**.** The average direct treatment effect in the experimental sample is $\theta_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\{Y_i^{\text{dir}}(1) - Y_i^{\text{dir}}(0) | S_i = e\}$.

The parameter θ_n may be viewed as an average direct effect, because it involves counterfactuals for unit i under a "direct" intervention on unit i's treatment assignment.

We now generalize Assumption 1 for the setting with spillovers. For clarity of exposition, we focus on the setting of Example 3.

Assumption C.7 (Spillovers)**.** Suppose the following:

i. Spillovers: If $D_i = d_i$ and $\mathbf{D}_{-i} = \mathbf{d}_{-i}$ then $Y_i = Y_i(d_i, \mathbf{d}_{-i})$.

[25] Formally, we may denote unobserved heterogeneity by η_i and use the nonseparable model notation $Y_i = Y_i(D_i, \mathbf{D}_{-i}, \eta_i)$. Then the potential outcome is $Y_i(d_i, \mathbf{d}_{-i}, \eta_i)$ and the direct potential outcome is $Y_i^{\text{dir}}(d_i, \eta_i) = \int Y_i(d_i, \mathbf{d}_{-i}, \eta_i)\}\text{dPr}(\mathbf{d}_{-i} | D_i = d_i, S_i = e, \eta_i)$.

ii. Randomization: $D_i \perp\!\!\!\perp \{Y_i(d_i, \mathbf{d}_{-i})\} \mid S_i = e$.[26]

iii. Overlap: $\Pr(D_i = 1 \mid S_i = e)$ is bounded away from zero and one almost surely.

Lemma C.1. Under Assumption C.7, the average direct treatment effect equals a difference of outcome means. Formally, $\theta_n = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}(Y_i|D_i=1,S_i=e) - \mathbb{E}(Y_i|D_i=0,S_i=e)\}$.

Proof. See Appendix D.3.1. □

If the outcomes were perfectly observed in the experimental sample, then θ_n would be identified by Lemma C.1.

Our benchmark estimator in the semi-synthetic exercise is the empirical analogue to this expression. It is the benchmark that an economist would obtain if they could fully observe the treatments and outcomes in the experiment. By Lemma C.1, the benchmark is an unbiased estimator of a reasonable estimand, even in the presence of spillovers.

As before, the crux of our problem is that the outcomes are not perfectly observed in the experimental sample. Instead, we have access to an auxiliary, observational sample. Our method recovers θ_n under appropriate modifications of our identifying assumptions. We now extend Assumptions 2 and 3(ii) accordingly.

Assumption C.8 (Stability with spillovers)**.** Suppose Assumption 2 holds for each unit $i \in \{1,...,n\}$.

Assumption C.9 (No direct effects with spillovers)**.** Suppose Assumption 3(ii) holds for each unit $i \in \{1,...,n\}$.

Most importantly, we modify stability (Assumption 2). In the presence of spillovers, stability requires $S_i \perp\!\!\!\perp R_i \mid (Y_i, D_i, X_i)$: conditional on a unit's *own* outcome, treatment, and covariates, the distribution of the unit's remotely sensed variable R_i would be the same had it belonged to the experimental or observational samples. In particular, the unit's remotely sensed variable distribution depends on its own outcome, but does not depend on other units' outcomes.

In our semi-synthetic exercise, unit i is a village. Treatments are randomized at the mandal level. A mandal is a sub-district containing villages, roughly comparable to a U.S. county. Substantively, this modified stability assumption means that the satellite image of a village depends only on that village's poverty level and on the treatment status of the mandal in which the village lies. It does not depend on the poverty levels of other villages, nor on the

[26] In nonseparable model notation, $D_i \perp\!\!\!\perp \eta_i | S_i = e$.

treatment statuses of other mandals. We think of this as a plausible approximation; intuitively the satellite image of a village should only depend on variables observed in that village.

In summary, our method extends as long as spillovers are in treatment effects, not in the distribution of the remotely sensed variable. We formalize this claim in what follows.

Theorem C.3 (Identification with spillovers)**.** Suppose Assumptions C.7, C.8, and C.9 hold. Then

$$\theta_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{E}(\Delta_i^e|R_i)}{\mathbb{E}(\Delta_i^o|R_i)} \quad \text{almost surely,}$$

where $\Delta_i^e = \frac{1(D_i=1,S_i=e)}{\Pr(D_i=1,S_i=e)} - \frac{1(D_i=0,S_i=e)}{\Pr(D_i=0,S_i=e)}$ and $\Delta_i^o = \frac{1(Y_i=1,S_i=o)}{\Pr(Y_i=1,S_i=o)} - \frac{1(Y_i=0,S_i=o)}{\Pr(Y_i=0,S_i=o)}$.

Proof. See Appendix D.3.2. □

C.3.2 Average Global Treatment Effect

Theorem C.3 identifies the average direct treatment effect, which is a reasonable estimand. We now ask: when does this reasonable estimand coincide with the standard estimand for models with spillovers, namely the average global treatment effect? We describe two plausible and empirically relevant scenarios. In these two scenarios, our method estimates the average global treatment effect in the presence of spillovers.

Definition C.2 (Average global treatment effect)**.** The average global treatment effect in the experimental sample is $\tilde{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{Y_i(1,\mathbf{1}_{n-1}) - Y_i(0,\mathbf{0}_{n-1})|S_i = e\}$, where $\mathbf{1}_{n-1}$ and $\mathbf{0}_{n-1}$ are vectors of repeated entries in $\mathbb{R}^{n-1}$.

The parameter $\tilde{\theta}_n$ may be viewed as an average global effect, because it involves counterfactuals for unit i under a "global" intervention on all units' treatment assignments.

Corollary C.3 (Spillovers within but not across mandals)**.** Suppose the assumptions of Theorem C.3 hold. Suppose that each unit i is a village, and that treatment assignment is randomized at the mandal level, as in the real experiment we study. Now, further assume that spillovers only occur within mandals. Then $\tilde{\theta}_n = \theta_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}(\Delta_i^e|R_i)}{\mathbb{E}(\Delta_i^o|R_i)}$.

Proof. See Appendix D.3.3. □

For simplicity, Corollary C.3 assumes that there are no spillovers across mandals. In fact, our argument will go through as long as spillovers across mandals are asymptotically negligible.

Next, we consider an alternative restriction on spillovers: that they only occur within a fixed geographic radius. This alternative restriction is also widely adopted in the empirical

literature. For example, Muralidharan et al. (2023) posit a radius of 20 km. After appropriately subsetting villages, our method once again estimates the average global treatment effect.

Corollary C.4 (Spillovers within a geographic radius)**.** Suppose the assumptions of Theorem C.3 hold. Suppose that each unit i is a village, and that treatment assignment is randomized at the mandal level, as in the real experiment we study. Now, further assume that spillovers only occur within a geographic radius of 20 km. Then $\tilde{\theta}_n$ coincides with θ_n after dropping from the experimental sample (both in the definition of $\tilde{\theta}_n$ and θ_n) any village that is within 20 km of another village with the opposite treatment assignment.

Proof. See Appendix D.3.4. □

By assuming no spillovers beyond a 20 km radius, our method estimates the average global treatment effect from a subset of the sample. On the one hand, this assumption is closer to that in the empirical literature. On the other hand, it comes at the cost of a smaller effective sample size.

C.4 Variables Collected at Different Times

In the main text, we avoided time indexing and viewed the remotely sensed variable as a post-outcome variable. In real data, such as the Smartcard illustration, the data have time indices. In particular, the remotely sensed variable and the outcome may be measured at different times. Moreover, the treatment variable may be defined as early adoption, raising the question of whether our method applies to such a setting. We now confirm that it does, and clarify the interpretation of our reported estimate.

To ease exposition, we discuss Example 3; we consider a binary outcome, omit pre-treatment covariates, and focus on "incomplete" observational cases (Assumption 3(ii)). The generalization to a discrete outcome with discrete covariates, as in Section 3 of the main text, is straightforward.

Suppose each unit is independent and identically distributed. Suppose early adoption of the treatment is randomly assigned at time t, the outcome is collected at time $t' \geq t$, and the remotely sensed variable is collected at time $t'' \geq t'$. In particular, only the treated experimental units receive treatment at time t, while all remaining units receive treatment at time t'. Overall, each unit is now characterized by the random vector

$$\{S, D_t, Y_{t'}(0), Y_{t'}(1), R_{t''}\}$$

where $Y_{t'}(d_t)$ are potential outcomes. For units in the experimental sample ($S=e$), we observe $(D_t, R_{t''})$. For units in the observational sample ($S=o$), we observe $(Y_{t'}, R_{t''})$. In this setting, $D_{t'}=1$ and $D_{t''}=1$ for all units, so we omit them.

The Smartcards illustration takes this form. Simplifying some of the details in Appendix F.2, only treated units in the experimental sample received Smartcards at time $t=2010$. Untreated units in the experimental sample received Smartcards by the time the outcomes were collected in $t'=2013$. All remaining units, i.e. the observational sample, received Smartcards in $t'=2013$ as well. Satellite images were collected at the later time $t''=2019$.

The parameter of interest is defined from the time t' potential outcomes. In the Smartcards illustration, it is the effect of 2010 Smartcards adoption on 2013 poverty levels.

Definition C.3 (Average time-specific treatment effect)**.** The average time-specific treatment effect in the experimental sample is $\theta_{t'} = \mathbb{E}\{Y_{t'}(1) - Y_{t'}(0) | S=e\}$.

We modify our assumption about the experimental sample to reflect the time variation in the measurement of the outcome and remotely sensed variable.

Assumption C.10 (Experiment with time variation)**.** Suppose the following:

i. SUTVA: $Y_{t'} = D_t Y_{t'}(1) + (1-D_t) Y_{t'}(0)$ almost surely.

ii. Randomization: $D_t \perp\!\!\!\perp \{Y_{t'}(0), Y_{t'}(1)\} \,|\, S=e$.

iii. Overlap: $\Pr(D_t=1 \,|\, S=e)$ is bounded away from zero and one almost surely.

iv. Ultimate adoption: $D_{t'}=1$ and $D_{t''}=1$ almost surely.

Next, we modify our assumptions on stability and on the observational sample to reflect the time variation in the measurement of the outcome and remotely sensed variable.

Assumption C.11 (Stability with time variation)**.** Suppose Assumption 2 holds replacing (D, Y, R) with $(D_t, Y_{t'}, R_{t''})$.

Assumption C.11 remains the main assumption of our framework: $S \perp\!\!\!\perp R_{t''} | D_t, Y_{t'}$. In an experiment with time variation, stability now requires that the conditional distribution of the remotely sensed variable $R_{t''}$ given $(D_t, Y_{t'})$ is stable across the experimental and observational samples. Importantly, the remotely sensed variable may be collected at a later time than the outcome. Figure H.2 provides empirical evidence supporting this assumption.

Assumption C.12 (No direct effects with time variation)**.** Suppose Assumption 3(ii) holds replacing (D,Y,R) with $(D_t,Y_{t'},R_{t''})$.

Assumption C.12 becomes $D_t \perp\!\!\!\perp R_{t''}|Y_{t'}$. Given the time variation in the measurement of the outcome and the remotely sensed variable, this amounts to a Markov property: the effect of the earlier treatment D_t on the later remotely sensed variable $R_{t''}$ must operate only though its effect on the intermediate outcome $Y_{t'}$. Importantly, $R_{t''}$ may depend on later outcomes $Y_{t''}$, but any effect from the earlier treatment D_t must be via the intermediate outcome $Y_{t'}$.

Concretely, consider the following example: 2010 Smartcards may affect 2013 consumption; 2013 consumption may affect 2019 consumption; and 2013 and 2019 consumption may affect 2019 satellite images. However, in this example, 2010 Smartcards cannot affect 2019 consumption in any channel besides 2013 consumption. This is one possible example that satisfies the Markov restriction. Figure H.3 provides empirical evidence supporting this assumption.

Theorem C.4 (Identification with time variation)**.** Suppose Assumptions C.10, C.11, and C.12 hold. Then

$$\theta_{t'} = \frac{\mathbb{E}(\Delta_t^e|R_{t''})}{\mathbb{E}(\Delta_{t'}^o|R_{t''})} \quad \text{almost surely,}$$

where $\Delta_t^e = \frac{1(D_t=1,S=e)}{\Pr(D_t=1,S=e)} - \frac{1(D_t=0,S=e)}{\Pr(D_t=0,S=e)}$ and $\Delta_{t'}^o = \frac{1(Y_{t'}=1,S=o)}{\Pr(Y_{t'}=1,S=o)} - \frac{1(Y_{t'}=0,S=o)}{\Pr(Y_{t'}=0,S=o)}$.

Proof. See Appendix D.4.1. □

In summary, our results directly apply when D_t is an early adoption treatment, and when $Y_{t'}$ and $R_{t''}$ are an outcome and a remotely sensed variable collected later on. In the main text, we report the effect of 2010 Smartcards adoption on 2013 consumption. With incomplete cases, there is an implicit Markov restriction that researchers must asses. However, with complete cases, this additional restriction may be relaxed.

C.5 Partial Identification with Approximate Assumptions

Our point identification results in Section 3 rely on two key conditions beyond experimental unconfoundedness: stability of the remotely sensed variable (Assumption 2) and, with incomplete observational cases, no direct effects of the treatment on the remotely sensed variable (Assumption 3(ii)). Sometimes researchers may believe these assumptions hold only approximately. For example, the treatment may slightly affect the remotely sensed variable through channels other than the outcome, or the sensing mechanism may vary modestly

across samples. In this section, we develop partial identification results that allow researchers to assess how sensitive their conclusions are to small departures from our assumptions.

We proceed in two steps. First, in Section C.5.1, we maintain stability but relax the no-direct-effects condition, bounding the magnitude of any direct effect. The direct effect bias enters additively. Second, in Section C.5.2, we also relax stability. The stability violation enters multiplicatively.

For simplicity, we focus on the special case of Example 3: a binary outcome, no covariates, and incomplete observational cases. Our results extend pointwise in x for discrete covariates. Recall that the treatment effect of interest is

$$\tau = \mu(1) - \mu(0), \quad \mu(d) = \mathbb{E}\{Y(d) | S = e\}.$$

Recall that, in this special case, the experimental and observational variation are

$$\Delta^e := \frac{1\{D=1, S=e\}}{\Pr(D=1, S=e)} - \frac{1\{D=0, S=e\}}{\Pr(D=0, S=e)}, \quad \Delta^o = \frac{1\{Y=1, S=o\}}{\Pr(Y=1, S=o)} - \frac{1\{Y=0, S=o\}}{\Pr(Y=0, S=o)}.$$

Moreover, no treatment variation is observed in the observational sample: $\Pr(D=0 \,|\, S=o) = 1$.

C.5.1 Approximately No Direct Effects

We first consider violations of Assumption 3(ii), allowing the treatment D to have a direct effect on the remotely sensed variable R. For any measurable function H with $\mathbb{E}\{\Delta^o H(R)\} \neq 0$, define the quantity

$$\widetilde{\tau}(H) := \frac{\mathbb{E}\{\Delta^e H(R)\}}{\mathbb{E}\{\Delta^o H(R)\}}$$

When Assumption 3(ii) (no direct effects) holds, Section 3 showed that $\widetilde{\tau}(H) = \tau$. When direct effects are present, $\widetilde{\tau}(H)$ remains identified but may differ from τ. We bound this discrepancy.

For the chosen representation H, define the outcome-specific direct-effect bias as

$$b_y(H) := \mathbb{E}\{H(R) \,|\, S=e, D=1, Y=y\} - \mathbb{E}\{H(R) \,|\, S=e, D=0, Y=y\}.$$

This quantity measures how much the treatment shifts the representation $H(R)$, while holding the outcome fixed. Under no-direct-effects, $b_y(H) = 0$. Define the overall direct-effect bias as

$$b(H) = \{1 - \mu(1)\} b_0(H) + \mu(1) b_1(H).$$

Lemma C.2 (Bias due to direct effects)**.** Suppose Assumptions 1 and 2 hold, and $\Pr(D = 0 \,|\, S = o) = 1$. Fix H with $\mathbb{E}\{\Delta^o H(R)\} \neq 0$. Then,

$$\tau = \widetilde{\tau}(H) - \frac{b(H)}{\mathbb{E}\{\Delta^o H(R)\}}.$$

Proof. See Appendix D.5.1. □

Lemma C.2 shows that bias is additive. The magnitude of the bias depends on the size of direct effect $b(H)$ and the relevance of the remotely sensed variable $\mathbb{E}\{\Delta^o H(R)\}$. If the researcher is willing to bound $b(H)$, then we can derive an identified set for τ.

Theorem C.5 (Partial identification with incomplete observational cases and bounded direct effects)**.** Suppose the conditions of Lemma C.2 hold. Further suppose the direct effects are bounded: $|b_y(H)| \leq \bar{b}(H)$ for $y \in \{0,1\}$. Then

$$\tau \in \left[\widetilde{\tau}(H) - \frac{\bar{b}(H)}{|\mathbb{E}\{\Delta^o H(R)\}|}, \widetilde{\tau}(H) + \frac{\bar{b}(H)}{|\mathbb{E}\{\Delta^o H(R)\}|}\right] \cap [-\mu(0), 1-\mu(0)].$$

Moreover, the set is sharp with respect to the reduced-form moments $\mathbb{E}\{\Delta^e H(R)\}$ and $\mathbb{E}\{\Delta^o H(R)\}$.

Proof. See Appendix D.5.2. □

There are several aspects of Theorem C.5 worth emphasizing. First, the width of the interval scales with the ratio $\frac{\bar{b}(H)}{|\mathbb{E}\{\Delta^o H(R)\}|}$, i.e. the size of the direct effect over the relevance of the remotely sensed variable. The bound is most informative when the scope for direct effects is limited (small $\bar{b}(H)$) and the remotely sensed variable is highly predictive of the outcome (large $|\mathbb{E}\{\Delta^o H(R)\}|$). In the environmental application of Examples 1, the treatment (e.g., a PES contract) is unlikely to directly affect the satellite image, so $\bar{b}(H)$ may be plausibly small.

Second, the feasibility constraint $\tau \in [-\mu(0), 1-\mu(0)]$ is operational because $\mu(0)$ is identified under Assumptions 1 and 2 alone, without requiring the no-direct-effects condition. By the proof of Lemma C.2, stability, and $\Pr(D=0|S=o)=1$,

$$\begin{aligned}
\mathbb{E}\{H(R)\,|\,S=e,D=0\} &= \mu(0)\mathbb{E}\{H(R)\,|\,S=e,D=0,Y=1\} + \{1-\mu(0)\}\mathbb{E}\{H(R)\,|\,S=e,D=0,Y=0\} \\
&= \mu(0)\mathbb{E}\{H(R)\,|\,S=o,D=0,Y=1\} + \{1-\mu(0)\}\mathbb{E}\{H(R)\,|\,S=o,D=0,Y=0\} \\
&= \mu(0)\mathbb{E}\{H(R)\,|\,S=o,Y=1\} + \{1-\mu(0)\}\mathbb{E}\{H(R)\,|\,S=o,Y=0\}
\end{aligned}$$

so $\mu(0)$ is identified.

Third, because the bound $\bar{b}(H)$ and the relevance $|\mathbb{E}\{\Delta^o H(R)\}|$ both depend on the choice of representation H, the researcher has a meaningful degree of freedom: they can select H to load on features of R for which direct effects are plausibly small while maintaining relevance. For example, in satellite imagery application, the researcher might construct H from spectral bands that are specific to the outcome of interest—such as near-infrared

reflectance for vegetation cover, or thermal signatures for crop burning—while avoiding those that could be affected by the treatment through other channels.

Remark C.1 (Breakpoint analysis.)**.** Researchers can use Theorem C.5 to conduct a breakpoint analysis: for each H, report the smallest value $\bar{b}(H)$ at which the identified set includes zero (or, more generally, at which a conclusion of interest is overturned) (Masten and Poirier, 2020). This value $\bar{b}(H)^{\text{bp}} = |\widetilde{\tau}(H)| \cdot |\mathbb{E}\{\Delta^o H(R)\}|$ is the smallest direct-effect (measured on the scale of the representation) required to explain away the estimated treatment effect. Large breakpoints indicate that the conclusion is robust to substantial departures from the no-direct-effects assumption.

Remark C.2 (Multiple moments.)**.** If we have a class $\mathcal{H}$ and, for each $H \in \mathcal{H}$, a bound $\bar{b}(H)$ such that $|b_y(H)| \leq \bar{b}(H)$ for $y \in \{0,1\}$, then intersecting the single-H bounds tightens identification:

$$\tau \in \left(\bigcap_{H \in \mathcal{H}: \mathbb{E}\{\Delta^o H(R)\} \neq 0} \left[\widetilde{\tau}(H) - \frac{\bar{b}(H)}{|\mathbb{E}\{\Delta^o H(R)\}|}, \widetilde{\tau}(H) + \frac{\bar{b}(H)}{|\mathbb{E}\{\Delta^o H(R)\}|} \right] \right) \cap [-\mu(0), 1-\mu(0)].$$

This intersection can also be viewed as a test for the stability restriction: if it is empty, then no DGP can satisfy the maintained assumptions.

C.5.2 Approximately No Direct Effects and Approximate Stability

We now relax both identifying assumptions simultaneously. We allow direct effects of D on R (relaxing Assumption 3(ii)) and instability across samples (relaxing Assumption 2).

For the chosen representation H, we introduce the outcome-specific stability violation as

$$s_y(H) := \mathbb{E}\{H(R) \mid S = e, D = 0, Y = y\} - \mathbb{E}\{H(R) \mid S = o, D = 0, Y = y\}.$$

This quantity measures how much the distribution of $H(R)$ differs across the experimental and observational samples, conditional upon $D = 0$ and $Y = y$. Under stability (Assumption 2), $s_y(H) = 0$. Define the overall stability violation

$$s(H) = s_1(H) - s_0(H).$$

Lemma C.3 (Bias due to direct effects and stability violations)**.** Suppose Assumption 1 holds and $\Pr(D = 0 \mid S = o) = 1$. Fix H with $\mathbb{E}\{\Delta^o H(R)\} + s(H) \neq 0$. Then

$$\tau = \frac{\mathbb{E}\{\Delta^e H(R)\} - b(H)}{\mathbb{E}\{\Delta^o H(R)\} + s(H)}.$$

Proof. See Appendix D.5.3. □

Lemma C.3 clarifies how the two violations interact. The direct-effect bias enters additively, just as in Lemma C.2. The stability violation enters multiplicatively, distorting the relevance of the remotely sensed variable from $\mathbb{E}\{\Delta^o H(R)\}$ to $\mathbb{E}\{\Delta^o H(R)\}+s(H)$ for some $s(H)$. If the researcher is willing to bound both the direct effect and the stability violation, then Lemma C.3 delivers an identified set for τ.

Theorem C.6 (Partial identification with incomplete observational cases, bounded direct effects, and bounded stability violations)**.** Suppose the conditions of Lemma C.3 hold. Further suppose the direct effects and stability violations are bounded: $|b_y(H)| \leq \bar{b}(H)$ and $|s_y(H)| \leq \bar{s}(H)$ for $y \in \{0,1\}$. Then

$$\tau \in \left\{ \frac{\mathbb{E}\{\Delta^e H(R)\} - b}{\mathbb{E}\{\Delta^o H(R)\} + s} : |b| \leq \bar{b}(H), |s| \leq 2\bar{s}(H), \mathbb{E}\{\Delta^o H(R)\} + s \neq 0 \right\}.$$

Moreover, if $|\mathbb{E}\{\Delta^o H(R)\}| > 2\bar{s}(H)$, then the identified set is contained within the interval

$$\tau \in [\widetilde{\tau}(H) - \kappa(H), \widetilde{\tau}(H) + \kappa(H)], \quad \kappa(H) := \frac{\bar{b}(H) + 2\bar{s}(H)|\widetilde{\tau}(H)|}{|\mathbb{E}\{\Delta^o H(R)\}| - 2\bar{s}(H)}.$$

Proof. See Appendix D.5.4. □

As before, the choice of representation H determines the width of these bounds. Stability is most plausible for representations that depend on the sensing technology rather than on sample-specific factors, e.g spectral indices derived from a common satellite platform. Choosing H to load on technologically stable features can simultaneously reduce the plausible value of $\bar{s}(H)$ and maintain a strong relevance $|\mathbb{E}\{\Delta^o H(R)\}|$, tightening the bounds in Theorem C.6.

C.6 Continuous Outcomes

We extend our results to outcomes that are continuous and bounded. Our practical suggestion is to discretize the support of a continuous outcome into bins, then to apply our method for discrete outcomes. Under a continuity condition on how the remotely sensed variable captures the underlying outcome, we bound the bias incurred from this discretization. Finally, we clarify how our relevance condition generalizes to a standard "completeness" condition when outcomes are continuous.

For clarity of exposition, we set aside covariates and focus on incomplete observational cases (Assumption 3(ii)). Our results directly extend to settings with covariates and with complete observational cases (Assumption 3(i)). Recall that $f_W(\cdot \mid ...)$ is our symbol for the Radon-Nikodym derivative.

C.6.1 Discretization

To begin, we place weak regularity conditions upon $f_{Y(d)}(y \mid S=e)$ (i.e., the conditional density of the potential outcome $Y(d) \in \mathbb{R}$ given the sample indicator $S=e$).

Assumption C.13 (Bounded outcomes)**.** Suppose that the support $\mathcal{Y}$ is uniformly bounded above and below.

Assumption C.13 states that the outcomes are uniformly bounded, so their support can be discretized into a finite number of bins $K < \infty$.

The general technique is to transform a continuous outcome $Y \in \mathcal{Y}$ into a discrete approximation $Y_\varepsilon \in \mathcal{Y}_\varepsilon$. Specifically, we discretize the continuous support $\mathcal{Y}$ into a grid $\mathcal{Y}_\varepsilon = \{y_1, \cdots, y_K\}$, where each grid value is the center of a bin of radius ε. As notation, let $B_\varepsilon(y)$ define an ℓ_∞-ball of radius $\varepsilon > 0$ around the value y, i.e. $B_\varepsilon(y) = \{y' \in \mathcal{Y} : \|y - y'\|_\infty \leq \varepsilon\}$. Define the discretized parameter $\theta_\varepsilon \in \mathbb{R}^{K-1}$. The kth element of this vector is the average effect of the treatment on a grid value: $\Pr\{Y(1) \in B_\varepsilon(y_k) | S=e\} - \Pr\{Y(0) \in B_\varepsilon(y_k) | S=e\}$.

We will show that the bias from discretizing the outcome is small as long as the remotely sensed variable smoothly reflects the outcome of interest. In other words, the remote sensing is continuous.

Assumption C.14 (Continuous sensing)**.** Suppose that for all $y, y' \in \mathcal{Y}$,

$$\sup_{r \in \mathcal{R}} |f_R(r \mid S=o, Y=y) - f_R(r \mid S=o, Y=y')| \leq \omega(|y - y'|)$$

for a continuous, non-negative function $\omega(t)$ with $\omega(0) = 0$. Let $\inf_{r \in \mathcal{R}} f_R(r) \geq \underline{\ell} > 0$.

Assumption C.14 states that if the outcome values y and y' are close, then the densities of the remotely sensed variable that they induce are close. Under this plausible assumption, we first characterize the moment restriction then analyze the bias of the treatment effect.

Proposition C.3 (Error from discretization)**.** Suppose Assumptions 1, 2, and 3(ii) hold. Suppose that $X = \varnothing$, and that Y has a continuous support $\mathcal{Y} \subset \mathbb{R}$. In addition, suppose Assumptions C.13 and C.14 hold. Then,

$$\left| \mathbb{E}\{\Delta^e - (\Delta^o_\varepsilon)^\top \theta_\varepsilon | R\} \right| \leq \frac{2\omega(2\varepsilon)}{\underline{\ell}} \text{ almost surely, where}$$

$$\Delta^e = \frac{1\{D=1, S=e\}}{\Pr(D=1, S=e)} - \frac{1\{D=0, S=e\}}{\Pr(D=0, S=e)}, \quad \Delta^o_{\varepsilon,k} = \frac{1\{Y \in B_\varepsilon(y_k), S=o\}}{\Pr\{Y \in B_\varepsilon(y_k), S=o\}} - \frac{1\{Y \in B_\varepsilon(y_K), S=o\}}{\Pr\{Y \in B_\varepsilon(y_K), S=o\}}.$$

Proof. See Appendix D.6.1. □

Proposition C.3 extends our conditional moment restriction from discrete outcomes to continuous outcomes via a discretization technique. The discretized parameter θ_ε approximately satisfies the conditional moment restriction, up to a small error that depends on the number of bins K and the smoothness of the sensing mechanism $\omega(\cdot)$.

Corollary C.5 (Discretization bias)**.** Suppose the conditions of Proposition C.3 hold. Consider any measurable function $H:\mathcal{R}\to\mathbb{R}^J$ with $J=K-1$, $\mathbb{E}\{\|H(R)\|\}<\infty$, and $\mathbb{E}\{H(R)(\Delta^o_\varepsilon)^\top\}$ invertible. Define the discretized average treatment effect

$$\widetilde{\tau}_\varepsilon=\lambda_\varepsilon^\top\mathbb{E}[\{H(R)(\Delta^o_\varepsilon)^\top\}]^{-1}\mathbb{E}\{H(R)\Delta^e\}$$

where $\lambda_\varepsilon=(y_1-y_K,\ldots,y_{K-1}-y_K)^\top$ for the grid $\mathcal{Y}_\varepsilon=\{y_1,\cdots,y_K\}$. Then,

$$|\widetilde{\tau}_\varepsilon-\tau|\leq 2\varepsilon+\frac{2\omega(2\varepsilon)}{\underline{\ell}}\cdot\|\lambda_\varepsilon\|\cdot\|[\mathbb{E}\{H(R)(\Delta^o_\varepsilon)^\top\}]^{-1}\|_{\mathrm{op}}\mathbb{E}\{\|H(R)\|\}.$$

Proof. See Appendix D.6.2. □

Corollary C.5 further shows that we can bound the bias of the discretized average treatment effect $\widetilde{\tau}_\varepsilon$ implied by the average effect of the treatment on grid values θ_ε.

In summary, under a plausible auxiliary assumption, we can bound the bias incurred by applying our estimation algorithm (Algorithm E.1) to discretized outcomes. When the bin size ε is fixed, one can easily extend our inference guarantees to accommodate this bias, e.g. with bias-aware confidence intervals.

A further extension is to allow the bin size to vanish: $\varepsilon\to 0$. This choice alters the estimator and asymptotics. Intuitively, it is like using local smoothing with a vanishing bandwidth. On the one hand, the first term in the bound converges to zero. On the other hand, the second term in the bound diverges, due to the factor $\|[\mathbb{E}\{H(R)(\Delta^o_\varepsilon)^\top\}]^{-1}\|_{\mathrm{op}}$. Choosing an ε that balances the two terms leads to an overall slower convergence rate for the treatment effect. More subtly, this choice changes the meaning of the relevance condition, which we now make explicit.

C.6.2 Generalizing the Relevance Condition

With binary outcomes, the relevance condition was $\mathbb{E}\{H(R)\Delta^o\}>0$. With discrete outcomes, it was invertibility of $\mathbb{E}\{H(R)(\Delta^o)^\top\}$. With discretized outcomes, it became invertibility of $\mathbb{E}\{H(R)(\Delta^o_\varepsilon)^\top\}$. We now articulate the appropriate generalization of this relevance condition when outcomes are viewed as continuous. It is a familiar condition called "completeness" from the literature on nonparametric inverse problems.

In Appendix B.1, we defined the treatment weights $\pi_d(R)$ and the outcome weights $\gamma_y(R)$:

$$\pi_d(R) = \frac{\Pr(D=d, S=e \mid R)}{\Pr(D=d, S=e)}, \quad \gamma_y(R) = \frac{f_{Y,S}(y,o|R)}{f_{Y,S}(y,o)}.$$

The treatment weights are identified from the experimental sample, while the outcome weights are identified from the observational sample.

As before, we combine the treatment weights and outcome weights to identify the causal parameter. With continuous outcomes, we obtain a conditional moment equation that exactly generalizes our conditional moment equation with discrete outcomes. Specifically, we will verify that Theorem 1 generalizes to

$$\pi_d(R) = \int_{y \in \mathcal{Y}} \gamma_y(R) f_{Y(d)}(y|S=e) \mathrm{d}y.$$

To connect this equation to the literature on nonparametric inverse problems, it helps to define the operator T which acts on the counterfactual density $f_{Y(d)}(y|S=e)$ and returns the treatment weights $\pi_d(R)$. The operator may be viewed as a way of encoding the outcome weights. In this notation, our conditional moment equation can be expressed succinctly:

$$\pi_d = T f_{Y(d)}, \quad (Tf)(R) = \int \gamma_y(R) f(y) \mathrm{d}y.$$

It is a classic inverse problem: to isolate $f_{Y(d)}(y|S=e)$, we must invert T. For such inverse problems, regularity conditions are well known (Kress, 1989).

A solution to the equation exists if π_d is in the range T. For a solution to exist, the treatment weights cannot be too correlated with variation in R that is only weakly generated by the outcome weights. This standard assumption is known as Picard's criterion (Kress, 1989).

Assumption C.15 (Picard's criterion)**.** Suppose that T is compact and has a singular value decomposition with left singular functions $\{\psi_j(r)\}_{j \geq 1}$, positive singular values $(\sigma_j)_{j \geq 1}$, and right singular functions $\{\phi_j(y)\}_{j \geq 1}$. Suppose that π_d is in the range T: $\sum_{j \geq 1} \frac{[\mathbb{E}\{\pi_d(R)\psi_j(R)\}]^2}{\sigma_j^2} < \infty$ and $\pi_d \in \text{null}(T^*)^\perp$ (i.e., the orthogonal complement to the null space of the adjoint of T).

A solution to the equation is unique if T is a one-to-one mapping. In other words, its null space is simply the zero function. Intuitively, no perturbation of the counterfactual density can leave the distribution of the remotely sensed variable unchanged. This standard assumption is called "completeness" (Newey and Powell, 2003). While it is often untestable (Canay et al., 2013), examples that satisfy this condition are well known (Andrews, 2017).

Assumption C.16 (Completeness)**.** If a function h satisfies $\int \gamma_y(R) h(y) \mathrm{d}y = 0$ almost surely, then $h(y) = 0$ almost everywhere.

When the outcome has finite support, completeness reduces to linear independence of the outcome weights for different outcome values. After projecting the outcome weights onto a sufficiently rich finite representation $H(R)$, we recover our relevance condition from the main text: invertibility of $\mathbb{E}\{H(R)(\Delta^o)^\top\}$.

Using this generalized relevance condition for the remotely sensed variable, we identify the counterfactual density.

Proposition C.4 (Identification without discretization)**.** Suppose Assumptions 1, 2, 3(ii), C.15, and C.16 hold. Suppose that $X=\varnothing$, and that Y has a continuous support $\mathcal{Y}\subset\mathbb{R}$. Then, for each $d\in\{0,1\}$, the counterfactual density $f_{Y(d)}(y|S=e)$ is identified as the unique solution to the conditional moment equation

$$\pi_d(R)=\int_{y\in\mathcal{Y}}\gamma_y(R)f_{Y(d)}(y|S=e)\mathrm{d}y.$$

Proof. See Appendix D.6.3. □

With continuous outcomes, the remotely sensed variable estimation problem is closely related to nonparametric instrumental variable regression and deconvolution, for which several nonparametric estimators are available in the literature; see e.g. Carrasco et al. (2007) for a review. The rate of estimation for $f_{Y(d)}(y\,|\,S=e)$ will depend on the rates of estimation for $\pi_d(R)$ and T, as throttled by the ill-posedness of inverting the operator T (e.g. Chen and Reiss, 2011). Articulating the technical details of such an estimation strategy is possible and left to future work.

C.7 Estimation with Covariates

In this section, we study estimation with discrete or continuous covariates.

C.7.1 Discrete Covariates and Incomplete Observational Cases

To begin, suppose we have incomplete observational cases (Assumption 3(ii)), and that the covariates X are discrete with finite support. Estimation and inference are straightforward: we apply Algorithm 1 within each covariate stratum and then average over the covariates.

Even in the presence of covariates, results in Chamberlain (1987) and Newey (1993) imply that representation that achieves efficiency, within a class of models satisfying Theorem 1, is

$$H^*(X,R)=\frac{\mathbb{E}\{\Delta^o(X)\,|\,X,R\}}{\sigma^2(\theta,X,R)},\quad \sigma^2(\theta,X,R)=\mathbb{E}[\{\Delta^e(X)-\Delta^o(X)^\top\theta(X)\}^2\,|\,X,R].$$

Here $H^*(X,R) \in \mathbb{R}^{K-1}$, $\mathbb{E}\{\Delta^o(X) \mid X,R\} \in \mathbb{R}^{K-1}$, and $\sigma^2(\theta,X,R) \in \mathbb{R}$. Its k-th component is the scalar $H_k^*(X,R) = \frac{\mathbb{E}\{\Delta_k^o(X)|X,R\}}{\sigma^2(\theta,X,R)}$, where $\Delta_k^o(X) = \frac{1\{Y=y_k,S=o\}}{\Pr(Y=y_k,S=o|X)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o|X)}$.

Like in the main text, the algorithm has three steps: divide the sample into TRAIN and TEST folds; learn the representation on TRAIN; and apply the learned representation to estimate the ATE on TEST. Algorithm E.1 provides details with sample splitting, though our analysis below allows for cross-fitting with any fixed number of folds. To analyze the properties of Algorithm E.1, we require only that the predictions, and hence the representation estimator, have some probability limit; they may be misspecified.

Assumption C.17 (Limit)**.** For each $x \in \mathcal{X}$, the learned representation has some mean square limit: $\mathbb{E}_R\left\{\|\widehat{H}(x,R) - \widetilde{H}(x,R)\|^2 | X = x\right\} = o_p(1)$ for some limit $\widetilde{H}(x,R) \in \mathbb{R}^{K-1}$ with $\mathbb{E}\{\|\widetilde{H}(x,R)\|^2 | X = x\} < \infty$. The limit is correlated with outcome variation: $\mathbb{E}\{\widetilde{H}(x,R)\Delta^o(x)^\top | X = x\}$ is non-singular for each $x \in \mathcal{X}$, and its smallest singular value is bounded away from zero.

The following inference guarantees generalize our results in Section 4 to include discrete covariates. Let $n_x = \sum_{i=1}^n 1_{X_i = x}$ denote the number of observations with cell $X_i = x$. We refer to $\Pr(X = x \mid S = e)$ as the marginal cell probabilities. We refer to $\Pr(D = d, S = e \mid X = x)$ and $\Pr(Y = y_k, S = o \mid X = x)$ as the conditional cell probabilities.

Proposition C.5 (Inference with known conditional cell probabilities)**.** Suppose Theorem 1's conditions and Assumption C.17 hold. Suppose the conditional cell probabilities are known and bounded away from zero. Then for each $x \in \mathcal{X}$, and for $\lambda = (y_1 - y_K, \ldots, y_{K-1} - y_K)^\top$,

$$(n_x)^{1/2}\{\widehat{\tau}(x) - \tau(x)\} \rightsquigarrow \mathcal{N}\{0, \lambda^\top A(x) B(x) A(x)^\top \lambda\}, \text{ where}$$

$A(x) = \mathbb{E}\{\widetilde{H}(x,R)\Delta^o(x)^\top \mid X = x\}^{-1}$ and $B(x) = \mathbb{E}[\{\Delta^e(x) - \Delta^o(x)^\top \theta(x)\}^2 \widetilde{H}(x,R)\widetilde{H}(x,R)^\top \mid X = x]$. Moreover, if $\widetilde{H}(x,R) = H^*(x,R)$, then $\widehat{\tau}(x) = \lambda^\top \widehat{\theta}(x)$ is semiparametrically efficient for $\tau(x) = \lambda^\top \theta(x)$ satisfying the conditional moment restriction $\mathbb{E}[\Delta^e(x) - \{\Delta^o(x)\}^\top \theta(x) | X = x, R] = 0$ with known conditional cell probabilities.

Proof. See Appendix D.7.1. □

Proposition C.6 (Inference with estimated conditional cell probabilities)**.** Suppose Theorem 1's conditions and Assumption C.17 hold. If the conditional cell probabilities and their estimators are bounded away from zero, then

$$(n_x)^{1/2}\{\widehat{\tau}(x) - \tau(x)\} \rightsquigarrow \mathcal{N}\{0, \lambda^\top A(x) V(x) A(x)^\top \lambda\},$$

where $V(x)$ is the conditional generalization of V in Lemma B.9.

Proof. See Appendix D.7.2. □

These $n_x^{-1/2}$ inference guarantees for the conditional average treatment effect (CATE) imply $n^{-1/2}$ inference guarantees for ATE by averaging over cells.

C.7.2 Discrete Covariates and Complete Observational Cases

Next, suppose we have complete observational cases (Assumption 3(i)), and that the covariates X are discrete with finite support. In this setting, the observational sample contains (X,D,Y,R), and we allow direct effects of treatment on the remotely sensed variable.

For each treatment arm $d \in \{0,1\}$ and covariate cell $x \in \mathcal{X}$, define the counterfactual outcome probabilities $\mu_k(d,x) = \Pr\{Y(d) = y_k \mid S = e, X = x\}$. Collect them into the vector $\mu(d,x) = \{\mu_1(d,x),\ldots,\mu_{K-1}(d,x)\}^\top \in \mathbb{R}^{K-1}$. Finally define $\mu_K(d,x) = 1 - \sum_{k=1}^{K-1} \mu_k(d,x)$. The CATE is then $\tau(x) = \sum_{k=1}^{K} y_k\{\mu_k(1,x) - \mu_k(0,x)\}$. The ATE is $\tau = \mathbb{E}\{\tau(X) \mid S = e\}$.

Once again, results in Chamberlain (1987) and Newey (1993) pin down the efficient representation. Within the class of models satisfying Theorem 2, the representation to use for $\mu(d,x)$ is

$$H_d^*(X,R) = \frac{\mathbb{E}\{\widetilde{\Delta}^o(d,X) \mid X,R\}}{\sigma_d^2(\mu,X,R)}, \quad \sigma_d^2(\mu,X,R) = \mathbb{E}[\{\widetilde{\Delta}^e(d,X) - \widetilde{\Delta}^o(d,X)^\top \mu(d,X)\}^2 \mid X,R].$$

Here $H_d^*(X,R) \in \mathbb{R}^{K-1}$, $\mathbb{E}\{\widetilde{\Delta}^o(d,X) \mid X,R\} \in \mathbb{R}^{K-1}$, and $\sigma_d^2(\mu,X,R) \in \mathbb{R}$.

So far, we have derived an optimal representation for each treatment arm d. It turns out that stacking the representations is the efficient choice for CATE. So see why, note that $1_{D=1}1_{D=0} = 0$, so mechanically the multiplication of $\widetilde{\Delta}^e(1,X)$ and $\widetilde{\Delta}^o(1,X)$ with $\widetilde{\Delta}^e(0,X)$ and $\widetilde{\Delta}^o(0,X)$ yields zero diagonals.

Having derived the efficient representation, our algorithms and guarantees for inference directly extend from the setting with incomplete observational cases to the setting with complete observational cases.

C.7.3 Continuous Covariates

As a final extension, Algorithm E.2 modifies our inferential procedure to accommodate continuous covariates. We focus on low-dimensional covariates with $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$. For simplicity, we focus on incomplete observational cases (Assumption 3(ii)).

Our identification results already apply to continuous covariates as written. The conditional moment equation from Theorem 1 is

$$\mathbb{E}\{\Delta^e(X) - \Delta^o(X)^\top \theta(X) \mid X,R\} = 0,$$

where $\theta_k(x) = \Pr\{Y(1) = y_k \mid S = e, X = x\} - \Pr\{Y(0) = y_k \mid S = e, X = x\}$. Corollary 1 gives

$$\theta(X) = \mathbb{E}\{H(X,R)\Delta^o(X)^\top \mid X\}^{-1}\mathbb{E}\{H(X,R)\Delta^e(X) \mid X\}.$$

ATE is then identified as $\tau = \mathbb{E}\{\lambda^\top \theta(X)\}$ for $\lambda = (y_1 - y_K, \ldots, y_{K-1} - y_K)^\top$.

With continuous covariates, the key modification to the algorithm is that the conditional expectations $\mathbb{E}\{H(X,R)\Delta^o(X)^\top \mid X\}$ and $\mathbb{E}\{H(X,R)\Delta^e(X) \mid X\}$ should be estimated as smooth functions of X using nonparametric methods, e.g., kernel smoothing, local polynomials, or series. With this modification, estimation proceeds in the same manner: divide the sample into TRAIN and TEST folds, learn the representation on TRAIN, then use the learned representation to estimate the ATE on TEST.

Valid inference would require standard smoothness conditions on these conditional expectations as functions of X, as well as undersmoothing or bias correction. The overall analysis with continuous covariates would largely follow our analysis with discrete covariates, combined with standard arguments for Z-estimation with nonparametric first stages (e.g., Newey, 1997; Ai and Chen, 2003; Chen, 2007).

As in the main text, no rate condition would be required on the estimated representation due to the infinite order Neyman orthogonality of the conditional moment restriction and sample splitting. The key technical insight—that the moment condition permits misspecified representation learning—continues to hold.

Series estimator. As a concrete example, we describe a series estimator.

Consider any representation H. Write $H_i = H(X_i, R_i) \in \mathbb{R}^{K-1}$. Construct the vector $C_i(H) \in \mathbb{R}^{K(K+2)}$ as the transpose of

$$\left(1_{D_i=1,S_i=e},\ 1_{D_i=1,S_i=e}H_i^\top,\ 1_{D_i=0,S_i=e},\ 1_{D_i=0,S_i=e}H_i^\top,\ 1_{Y_i=y_1,S_i=o},\ 1_{Y_i=y_1,S_i=o}H_i^\top,\ \cdots,\ 1_{Y_i=y_K,S_i=o},\ 1_{Y_i=y_K,S_i=o}H_i^\top\right)^\top$$

where the odd entries are scalars and the even entries are vectors in $\mathbb{R}^{K-1}$. The expectations of these quantities, conditional upon covariates, pin down $\mathbb{E}\{H(X,R)\Delta^o(X)^\top \mid X\}$ and $\mathbb{E}\{H(X,R)\Delta^e(X) \mid X\}$. Therefore, the conditional expectations pin down the CATE and ATE.

Consider a basis b and denote $B_i = b(X_i) \in \mathbb{R}^J$. Define the series estimator as the vector-valued regression

$$\widehat{g}_{\text{TRAIN}}(x;H) = b(x)^\top \widehat{\beta}_{\text{TRAIN}}(H), \quad \widehat{\beta}_{\text{TRAIN}}(H) = \operatorname*{argmin}_\beta \sum_{i \in \text{TRAIN}} \|C_i(H) - B_i^\top \beta\|^2.$$

This vector valued regression is the essential ingredient to estimate $\mathbb{E}\{H(X,R)\Delta^o(X)^\top \mid X\}$ and $\mathbb{E}\{H(X,R)\Delta^e(X) \mid X\}$. An undersmoothed series is one where J grows more quickly than the optimal rate for estimation.

Properties. The series estimator will be valid under standard series regularity conditions. First, each component of the vector-valued regression should be Hölder smooth over a compact support, so the approximation error vanishes as J increases. Second, the covariance matrix $\mathbb{E}_{\text{TRAIN}}(BB^\top)$ converges in probability to $\mathbb{E}(BB^\top)$, e.g. when $J^2 \ll n$, so that the sampling error vanishes as J increases. Finally, the objects $\mathbb{E}(BB^\top)$ and $\mathbb{E}\{\widetilde{H}(X,R)\Delta^o(X)^\top \mid X\}$ are well conditioned with singular values bounded away from zero.

Under such conditions, standard arguments yield pointwise asymptotic normality of $\widehat{\theta}(X)$ after either undersmoothing (i.e., choosing J large enough that the series bias is negligible) or bias correction (e.g., Newey, 1997; Chen, 2007). Since τ is smooth functional of $\theta(X)$, $\widehat{\tau}$ will be $n^{-1/2}$ asymptotically normal, provided that the series dimension J satisfies restrictions ensuring small bias.

D Proofs for Additional Theoretical Results

D.1 Bias of Alternative Approaches: Extensions

D.1.1 Proof of Proposition C.1

We prove each result.

1. We prove (i) with possibly discrete outcomes. Fix $d \in \{0,1\}$. Let

$$m_d(R) = \mathbb{E}(Y \mid R, D=d, S=o) = \int y f_Y(y \mid R, D=d, S=o) \mathrm{d}y.$$

By Bayes' rule,

$$f_Y(Y \mid R, D=d, S=o) = \frac{f_R(R \mid Y, D=d, S=o) f_Y(Y \mid D=d, S=o)}{f_R(R \mid D=d, S=o)}$$
$$f_R(R \mid Y, D=d, S=e) = \frac{f_Y(Y \mid R, D=d, S=e) f_R(R \mid D=d, S=e)}{f_Y(Y \mid D=d, S=e)}.$$

By stability,

$$f_R(R \mid Y, D=d, S=o) = f_R(R \mid Y, D=d, S=e).$$

We combine these expressions to rewrite

$$f_Y(Y \mid R, D=d, S=o) = f_Y(Y \mid R, D=d, S=e) v_d(R,Y)$$
$$v_d(R,Y) = \frac{f_Y(Y \mid D=d, S=o)}{f_Y(Y \mid D=d, S=e)} \frac{f_R(R \mid D=d, S=e)}{f_R(R \mid D=d, S=o)}.$$

Consequently,

$$m_d(R) = \int y f_Y(y \mid R, D=d, S=e) v_d(R,y) \mathrm{d}y = \mathbb{E}\{Y v_d(R,Y) \mid R, D=d, S=e\}.$$

Since $\widetilde{\mu}'(d) = \mathbb{E}\{m_d(R) \mid D=d, S=e\}$, the result follows by iterated expectations.

2. We prove (ii) for binary outcomes, where $\mu(d) = \Pr(Y=1 \mid D=d, S=e)$. Fix $d \in \{0,1\}$. By iterated expectations and stability,

$$\begin{aligned} f_R(R \mid D=d, S=e) &= \mu(d) f_R(R \mid Y=1, D=d, S=e) + \{1-\mu(d)\} f_R(R \mid Y=0, D=d, S=e) \\ &= \mu(d) f_R(R \mid Y=1, D=d, S=o) + \{1-\mu(d)\} f_R(R \mid Y=0, D=d, S=o). \end{aligned}$$

As a consequence,

$$\begin{aligned} \widetilde{\mu}'(d) &= \mathbb{E}\{m_d(R) \mid D=d, S=e\} \\ &= \mu(d) \mathbb{E}\{m_d(R) \mid Y=1, D=d, S=o\} + \{1-\mu(d)\} \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\} \\ &= \mu(d) \kappa_d + \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\}, \end{aligned}$$

where within the first term,

$$\begin{aligned} \kappa_d &:= \mathbb{E}\{m_d(R) \mid Y=1, D=d, S=o\} - \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\} \\ &= \frac{\mathrm{Cov}\{m_d(R), Y \mid D=d, S=o\}}{\mathrm{Var}(Y \mid D=d, S=o)} \\ &= \frac{\mathrm{Var}\{m_d(R) \mid D=d, S=o\}}{\mathrm{Var}(Y \mid D=d, S=o)}. \end{aligned}$$

By the law of total variance,

$$\mathrm{Var}(Y \mid D=d, S=o) = \mathrm{Var}\{m_d(R) \mid D=d, S=o\} + \mathbb{E}\{\mathrm{Var}(Y \mid R, D=d, S=o) \mid D=d, S=o\},$$

so $\kappa_d \in [0,1]$. Moreover, $\kappa_d = 1$ if and only if $\mathbb{E}\{\mathrm{Var}(Y \mid R, D=d, S=o) \mid D=d, S=o\} = 0$, which is equivalent to $\mathrm{Var}(Y \mid R, D=d, S=o) = 0$ almost surely.

What remains is to show

$$\mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\} = (1-\kappa_d) \mathbb{E}(Y \mid D=d, S=o).$$

By iterated expectations,

$$
\begin{aligned}
p_d &:= \mathbb{E}(Y \mid D=d, S=o) \\
&= \mathbb{E}\{\mathbb{E}(Y \mid R, D=d, S=o) \mid D=d, S=o\} \\
&= \mathbb{E}\{m_d(R) \mid D=d, S=o\} \\
&= \mathbb{E}\{m_d(R) \mid Y=1, D=d, S=o\} p_d + \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\}(1-p_d) \\
&= [\kappa_d + \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\}] p_d + \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\}(1-p_d) \\
&= \kappa_d p_d + \mathbb{E}\{m_d(R) \mid Y=0, D=d, S=o\}.
\end{aligned}
$$

□

D.1.2 Proof of Proposition C.2

We prove each result.

1. Recall that Proposition 1(i) generally established

$$
\widetilde{\mu}(d) = \mu(d) + \mathbb{E}[Y\{w_d(R,Y)-1\} \mid D=d, S=e], \quad w_d(r,y) := \frac{f_Y(y \mid S=o)}{f_Y(y \mid D=d, S=e)} \frac{f_R(r \mid D=d, S=e)}{f_R(r \mid S=o)}.
$$

Because $S \perp\!\!\!\perp \{Y(0), Y(1)\}$ and we are in the setting with incomplete observational cases,

$$
f_Y(y \mid S=o) = f_{Y(0)}(y \mid S=o) = f_{Y(0)}(y \mid S=e) = f_Y(y \mid D=0, S=e).
$$

Next, we prove that $f_R(r \mid S=o) = f_R(r \mid D=0, S=e)$. First, we write

$$
\begin{aligned}
f_R(R \mid S=o) &= \int_y f_R(R \mid Y=y, S=o) f_Y(y \mid S=o) \mathrm{d}y \\
f_R(R \mid D=0, S=e) &= \int_y f_R(R \mid Y=y, D=0, S=e) f_Y(y \mid D=0, S=e) \mathrm{d}y.
\end{aligned}
$$

By Assumptions 2 and 3(ii), $(S,D) \perp\!\!\!\perp R \mid Y$, hence $f_R(R \mid Y, S=o) = f_R(R \mid Y, D=0, S=e)$. Above, we proved that $f_Y(y \mid S=o) = f_Y(y \mid D=0, S=e)$.

In summary, we have shown

$$
\begin{aligned}
w_0(r,y) &= \frac{f_Y(y \mid S=o)}{f_Y(y \mid D=0, S=e)} \frac{f_R(r \mid D=0, S=e)}{f_R(r \mid S=o)} = 1 \\
w_1(r,y) &= \frac{f_Y(y \mid S=o)}{f_Y(y \mid D=1, S=e)} \frac{f_R(r \mid D=1, S=e)}{f_R(r \mid S=o)} = \frac{f_Y(y \mid D=0, S=e)}{f_Y(y \mid D=1, S=e)} \frac{f_R(r \mid D=1, S=e)}{f_R(r \mid D=0, S=e)}.
\end{aligned}
$$

2. Recall that Proposition 1(ii) established that $\widetilde{\tau} = \kappa\tau$ for $\kappa \in [0,1)$. We have shown $\widetilde{\mu}(0) = \mu(0)$. Therefore

$$
\widetilde{\mu}(1) - \mu(0) = \widetilde{\mu}(1) - \widetilde{\mu}(0) = \widetilde{\tau} = \kappa\tau = \kappa\{\mu(1) - \mu(0)\}.
$$

□

D.2 Quasi-Experiments with Remotely Sensed Outcomes

D.2.1 Proof of Theorem C.1

1. By the law of total probability,

$$\begin{aligned}\delta^e_{d,z}(R) &:= f_R(R \mid S=e, D=d, Z=z) \\ &= \int f_{R,Y}(R, y \mid S=e, D=d, Z=z)\mathrm{d}y \\ &= \int f_R(R \mid S=e, D=d, Z=z, Y=y) f_Y(y \mid S=e, D=d, Z=z)\mathrm{d}y.\end{aligned}$$

Next, notice that

$$\begin{aligned}f_R(R \mid S=e, D=d, Z=z, Y=y) &= f_R(R \mid Y=y) \\ &= f_R(R \mid S=o, Y=y) \\ &=: \delta^o_y(R),\end{aligned}$$

where the equalities apply the contraction implied by Assumptions C.2 and C.3: $(S, D, Z) \perp\!\!\!\perp R \mid Y$.

Combining the previous displays, we arrive at the general result

$$\delta^e_{d,z}(R) = \int \delta^o_y(R) f_Y(y \mid S=e, D=d, Z=z)\mathrm{d}y.$$

When Y is binary,

$$\begin{aligned}f_Y(1 \mid S=e, D=d, Z=z) &= \mathbb{E}(Y \mid S=e, D=d, Z=z) = \alpha(d,z) \\ f_Y(0 \mid S=e, D=d, Z=z) &= 1 - \mathbb{E}(Y \mid S=e, D=d, Z=z) = 1 - \alpha(d,z).\end{aligned}$$

Therefore, the general result specializes to

$$\delta^e_{d,z}(R) = \delta^o_1(R)\alpha(d,z) + \delta^o_0(R)\{1-\alpha(d,z)\} = \delta^o_0(R) + \{\delta^o_1(R) - \delta^o_0(R)\}\alpha(d,z).$$

2. We apply Bayes' rule to rewrite

$$\begin{aligned}\delta^o_y(R) = f_R(R \mid S=o, Y=y) &= \frac{\Pr(Y=y, S=o \mid R) f_R(R)}{\Pr(Y=y, S=o)}, \\ \delta^e_{d,z}(R) = f_R(R \mid S=e, D=d, Z=z) &= \frac{\Pr(D=d, Z=z, S=e \mid R) f_R(R)}{\Pr(D=d, Z=z, S=e)}.\end{aligned}$$

Substituting these expressions into the previous step and canceling $f_R(R)$ yields

$$\mathbb{E}\{\Delta^e(d,z) - \Delta^o \alpha(d,z) \mid R\} = 0.$$

□

D.2.2 Proof of Theorem C.2

1. By the law of total probability,

$$\begin{aligned}\delta^e_{d,t}(R_t) &:= f_{R_t}(R_t \mid S=e, D=d) \\ &= \int f_{R_t,Y_t}(R_t, y \mid S=e, D=d)\mathrm{d}y \\ &= \int f_{R_t}(R_t \mid S=e, D=d, Y_t=y) f_{Y_t}(y \mid S=e, D=d)\mathrm{d}y.\end{aligned}$$

Next, notice that

$$\begin{aligned}f_{R_t}(R_t \mid S=e, D=d, Y_t=y) &= f_{R_t}(R_t \mid Y_t=y) \\ &= f_{R_t}(R_t \mid S=o, Y_t=y) \\ &=: \delta^o_{y,t}(R_t),\end{aligned}$$

where the equalities apply the contraction implied by Assumptions C.5 and C.6: $(S,D) \perp\!\!\!\perp R_t \mid Y_t$.

Combining the previous displays, we arrive at the general result

$$\delta^e_{d,t}(R_t) = \int \delta^o_{y,t}(R_t) f_{Y_t}(y \mid S=e, D=d)\mathrm{d}y.$$

When Y is binary,

$$\begin{aligned}f_{Y_t}(1 \mid S=e, D=d) &= \mathbb{E}(Y_t \mid S=e, D=d) = \alpha_t(d) \\ f_{Y_t}(0 \mid S=e, D=d) &= 1 - \mathbb{E}(Y_t \mid S=e, D=d) = 1 - \alpha_t(d).\end{aligned}$$

Therefore, the general result specializes to

$$\delta^e_{d,t}(R_t) = \delta^o_{1,t}(R_t)\alpha_t(d) + \delta^o_{0,t}(R_t)\{1-\alpha_t(d)\} = \delta^o_{0,t}(R_t) + \{\delta^o_{1,t}(R_t) - \delta^o_{0,t}(R_t)\}\alpha_t(d).$$

2. We apply Bayes' rule to rewrite

$$\begin{aligned}\delta^o_{y,t}(R_t) &= f_{R_t}(R_t \mid Y_t=y, S=o) = \frac{\Pr(Y_t=y, S=o \mid R_t) f_{R_t}(R_t)}{\Pr(Y_t=y, S=o)}, \\ \delta^e_{d,t}(R_t) &= f_{R_t}(R_t \mid D=d, S=e) = \frac{\Pr(D=d, S=e \mid R_t) f_{R_t}(R_t)}{\Pr(D=d, S=e)}.\end{aligned}$$

Substituting these expressions into the previous step and canceling $f_{R_t}(R_t)$ yields

$$\mathbb{E}\{\Delta^e_t(d) - \Delta^o_t \alpha_t(d) \mid R_t\} = 0.$$

□

D.3 Some Spillovers

D.3.1 Proof of Lemma C.1

Using nonseparable model notation with unobserved heterogeneity η_i, as defined in Footnote 25,

$$\begin{aligned}\mathbb{E}(Y_i|D_i=d,S_i=e) &= \mathbb{E}\{Y_i(D_i,\mathbf{D}_{-i},\eta_i)|D_i=d,S_i=e\}\\ &= \mathbb{E}[\mathbb{E}\{Y_i(D_i,\mathbf{D}_{-i},\eta_i)|D_i=d,S_i=e,\eta_i)\}|D_i=d,S_i=e]\\ &= \mathbb{E}\{Y_i^{\text{dir}}(d,\eta_i)|D_i=d,S_i=e\}\\ &= \mathbb{E}\{Y_i^{\text{dir}}(d,\eta_i)|S_i=e\}\end{aligned}$$

where the final line appeals to randomization.

D.3.2 Proof of Theorem C.3

The argument mirrors the proof of Lemma 1 and Theorem 1.

1. By the law of total probability,

$$\begin{aligned}\delta_{d,i}^e(R_i) &:= f_{R_i}(R_i\,|\,S_i=e,D_i=d)\\ &= \int f_{R_i,Y_i}(R_i,y\,|\,S_i=e,D_i=d)\mathrm{d}y\\ &= \int f_{R_i}(R_i\,|\,S_i=e,D_i=d,Y_i=y)f_{Y_i}(y\,|\,S_i=e,D_i=d)\mathrm{d}y.\end{aligned}$$

Next, notice that

$$\begin{aligned}f_{R_i}(R_i\,|\,S_i=e,D_i=d,Y_i=y) &= f_{R_i}(R_i\,|\,Y_i=y)\\ &= f_{R_i}(R_i\,|\,S_i=o,Y_i=y)\\ &=: \delta_{y,i}^o(R_i),\end{aligned}$$

where the equalities apply the contraction implied by Assumptions C.8 and C.9: $(S_i,D_i)\perp\!\!\!\perp R_i\,|\,Y_i$.

Combining the previous displays, we arrive at the general result

$$\delta_{d,i}^e(R_i) = \int \delta_{y,i}^o(R_i)f_{Y_i}(y\,|\,S_i=e,D_i=d)\mathrm{d}y.$$

When Y is binary,

$$\begin{aligned}f_{Y_i}(1\,|\,S_i=e,D_i=d) &= \mathbb{E}(Y_i\,|\,S_i=e,D_i=d) =: \mu_i(d)\\ f_{Y_i}(0\,|\,S_i=e,D_i=d) &= 1-\mathbb{E}(Y_i\,|\,S_i=e,D_i=d) = 1-\mu_i(d).\end{aligned}$$

By the proof of Lemma C.1, $\mu_i(d) = \mathbb{E}\{Y_i^{\text{dir}}(d)|S_i = e\}$. Therefore, the general result specializes to

$$\delta^e_{d,i}(R_i) = \delta^o_{1,i}(R_i)\mu_i(d) + \delta^o_{0,i}(R_i)\{1-\mu_i(d)\} = \delta^o_{0,i}(R_i) + \{\delta^o_{1,i}(R_i) - \delta^o_{0,i}(R_i)\}\mathbb{E}\{Y_i^{\text{dir}}(d)|S_i = e\}.$$

2. We apply Bayes' rule to rewrite

$$\delta^o_{y,i}(R_i) = f_{R_i}(R_i \,|\, S_i = o, Y_i = y) = \frac{\Pr(Y_i = y, S_i = o \,|\, R_i) f_{R_i}(R_i)}{\Pr(Y_i = y, S_i = o)},$$
$$\delta^e_{d,i}(R_i) = f_{R_i}(R_i \,|\, S_i = e, D_i = d) = \frac{\Pr(D_i = d, S_i = e \,|\, R_i) f_{R_i}(R_i)}{\Pr(D_i = d, S_i = e)}.$$

Substituting these expressions into the previous step and canceling $f_{R_i}(R_i)$ gives

$$\mathbb{E}(\Delta^e_i|R_i) = \mathbb{E}(\Delta^o_i|R_i)\theta_i, \quad \theta_i = \mathbb{E}\{Y_i^{\text{dir}}(1) - Y_i^{\text{dir}}(0)|S_i = e\}.$$

We conclude that $\theta_n = \frac{1}{n}\sum_{i=1}^n \theta_i = \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}(\Delta^e_i|R_i)}{\mathbb{E}(\Delta^o_i|R_i)}$.

D.3.3 Proof of Corollary C.3

Because spillovers only occur within mandals, the potential outcome of village i only depends on the treatment assignments of other villages within its mandal; it does not depend on treatment assignments of other villages in other mandals. Because treatment assignment is at the mandal level, the treatment of villages i matches the treatment of all other villages in its mandal. These statements imply that $\mathbb{E}\{Y_i^{\text{dir}}(1)|S_i = e\} = \mathbb{E}\{Y_i(1, \mathbf{1}_{n-1})|S_i = e\}$. The same is true for $d_i = 0$. Therefore, $\theta_n = \tilde{\theta}_n$ and we are done.

We formally prove this statement for village $i = 1$ and $d_1 = 1$. Suppose that, among villages $\{1,...,n\}$, the initial m belong to the same mandal. As argued in the proof of Lemma C.1, with nonseparable model notation defined in Footnote 25,

$$\begin{aligned}
\mathbb{E}\{Y_1^{\text{dir}}(1,\eta_1)|S_1 = e\} &= \mathbb{E}\{Y_1(1, D_2, ..., D_n, \eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1, D_2, ..., D_m, \eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1, \mathbf{1}_{m-1}, \eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1, \mathbf{1}_{n-1}, \eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1, \mathbf{1}_{n-1}, \eta_1)|S_1 = e\}
\end{aligned}$$

by spillovers only within mandals, mandal-level treatment assignment, spillovers only within mandals, and randomization.

D.3.4 Proof of Corollary C.4

The argument is similar to the proof of Corollary C.3. We prove the key equality.

As before, consider village $i=1$ and $d_1=1$. Suppose that, among villages $\{1,...,n\}$, the initial m are within a 20 km radius. Each of these m villages must have been treated due to the subetting rule.

As argued in the proof of Lemma C.1, with nonseparable model notation defined in Footnote 25,

$$\begin{aligned}\mathbb{E}\{Y_1^{\text{dir}}(1,\eta_1)|S_1=e\} &= \mathbb{E}\{Y_1(1,D_2,...,D_n,\eta_1)|D_1=1,S_1=e\} \\ &= \mathbb{E}\{Y_1(1,D_2,...,D_m,\eta_1)|D_1=1,S_1=e\} \\ &= \mathbb{E}\{Y_1(1,\mathbf{1}_{m-1},\eta_1)|D_1=1,S_1=e\} \\ &= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|D_1=1,S_1=e\} \\ &= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|S_1=e\}\end{aligned}$$

by spillovers only within the radius, the subsetting rule, spillovers only within the radius, and randomization.

D.4 Variables Collected at Different Times

D.4.1 Proof of Theorem C.4

The argument is identical to the proof of Lemma 1 and Theorem 1, replacing (D,Y,R) with $(D_t,Y_{t'},R_{t''})$.

D.5 Partial Identification with Approximate Assumptions

D.5.1 Proof of Lemma C.2

We proceed in steps.

1. First, we reduce the numerator $\mathbb{E}\{\Delta^e H(R)\}$ and denominator $\mathbb{E}\{\Delta^o H(R)\}$ to differences in conditional means. By the definition of Δ^e,

$$\mathbb{E}\{\Delta^e H(R)\} = \mathbb{E}\{H(R)\,|\,S=e,D=1\} - \mathbb{E}\{H(R)\,|\,S=e,D=0\}.$$

By the definition of Δ^o and $\Pr(D=0\,|\,S=o)=1$,

$$\mathbb{E}\{\Delta^o H(R)\} = \mathbb{E}\{H(R)\,|\,S=o,D=0,Y=1\} - \mathbb{E}\{H(R)\,|\,S=o,D=0,Y=0\}.$$

2. Next, we analyze the numerator. For each $d \in \{0,1\}$,

$$\begin{aligned}
&\mathbb{E}\{H(R) \mid S=e, D=d\} \\
&= \mathbb{E}\{H(R) \mid S=e, D=d, Y=1\} \Pr(Y=1 \mid S=e, D=d) + \mathbb{E}\{H(R) \mid S=e, D=d, Y=0\} \Pr(Y=0 \mid S=e, D=d) \\
&= \mu(d) \mathbb{E}\{H(R) \mid S=e, D=d, Y=1\} + \{1-\mu(d)\} \mathbb{E}\{H(R) \mid S=e, D=d, Y=0\} \\
&= \mu(d) m_{d,1} + \{1-\mu(d)\} m_{d,0}.
\end{aligned}$$

by the law of iterated expectations, Assumption 1, and the notation $m_{d,y} = \mathbb{E}\{H(R) \mid S = e, D=d, Y=y\}$.

Taking differences across treatment arms and using $m_{1,y} = m_{0,y} + b_y(H)$,

$$\begin{aligned}
\mathbb{E}\{\Delta^e H(R)\} &= \mathbb{E}\{H(R) \mid S=e, D=1\} - \mathbb{E}\{H(R) \mid S=e, D=0\} \\
&= \mu(1) m_{1,1} + \{1-\mu(1)\} m_{1,0} - \mu(0) m_{0,1} - \{1-\mu(0)\} m_{0,0} \\
&= \mu(1)\{m_{0,1} + b_1(H)\} + \{1-\mu(1)\}\{m_{0,0} + b_0(H)\} - \mu(0) m_{0,1} - \{1-\mu(0)\} m_{0,0} \\
&= \{\mu(1) - \mu(0)\}(m_{0,1} - m_{0,0}) + \mu(1) b_1(H) + \{1-\mu(1)\} b_0(H) \\
&= \tau(m_{0,1} - m_{0,0}) + b(H).
\end{aligned}$$

3. Finally, we analyze the denominator. By Assumption 2,

$$m_{0,y} = \mathbb{E}\{H(R) \mid S=e, D=0, Y=y\} = \mathbb{E}\{H(R) \mid S=o, D=0, Y=y\}.$$

Therefore

$$\mathbb{E}\{\Delta^o H(R)\} = \mathbb{E}\{H(R) \mid S=o, D=0, Y=1\} - \mathbb{E}\{H(R) \mid S=o, D=0, Y=0\} = m_{0,1} - m_{0,0}.$$

Collecting results,

$$\mathbb{E}\{\Delta^e H(R)\} = \tau \mathbb{E}\{\Delta^o H(R)\} + b(H).$$

Dividing by $\mathbb{E}\{\Delta^o H(R)\}$ gives the desired conclusion.

D.5.2 Proof of Theorem C.5

We prove each result.

1. First, we prove the set is a valid outer bound. By Lemma C.2,

$$\widetilde{\tau}(H) - \tau = \frac{b(H)}{\mathbb{E}\{\Delta^o H(R)\}}.$$

By construction, $b(H)$ is a convex combination of $b_1(H)$ and $b_0(H)$, so the assumed bound implies $|b(H)| \leq \bar{b}(H)$. Therefore,

$$|\widetilde{\tau}(H) - \tau| \leq \frac{\bar{b}(H)}{|\mathbb{E}\{\Delta^o H(R)\}|}.$$

This gives the main interval.

The second interval is due to the binary outcomes. Since $\tau = \mu(1) - \mu(0)$ and $\mu(1) \in [0,1]$, we must have $-\mu(0) \leq \tau \leq 1 - \mu(0)$.

2. Next, we prove the set is sharp with respect to the reduced form moments. For clarity, let $\mathbb{E}_{\text{obs}}(\cdot)$ denote the moments under the observed data distribution and $\mathbb{E}_\star(\cdot)$ the moments under a constructed DGP.

For any candidate $\tau_\star$ in the set, define the required aggregate bias

$$b_\star := \mathbb{E}_{\text{obs}}\{\Delta^e H(R)\} - \tau_\star \mathbb{E}_{\text{obs}}\{\Delta^o H(R)\}.$$

Because $\tau_\star$ lies in the interval, $|b_\star| \leq \bar{b}(H)$. Now, construct reduced-form moments satisfying

$$\mathbb{E}_\star\{\Delta^o H(R)\} = \mathbb{E}_{\text{obs}}\{\Delta^o H(R)\}, \quad \tau = \tau_\star, \quad b(H) = b_\star$$

Then Lemma C.2 implies

$$\mathbb{E}_\star\{\Delta^e H(R)\} = \tau_\star \mathbb{E}_\star\{\Delta^o H(R)\} + b_\star.$$

Substituting the definitions gives

$$\begin{aligned}\mathbb{E}_\star\{\Delta^e H(R)\} &= \tau_\star \mathbb{E}_{\text{obs}}\{\Delta^o H(R)\} + [\mathbb{E}_{\text{obs}}\{\Delta^e H(R)\} - \tau_\star \mathbb{E}_{\text{obs}}\{\Delta^o H(R)\}] \\ &= \mathbb{E}_{\text{obs}}\{\Delta^e H(R)\}.\end{aligned}$$

Thus the same observed reduced-form moments are reproduced while the treatment effect equals $\tau_\star$. Since $\tau_\star$ was arbitrary, the bound is sharp with respect to the reduced-form moments.

D.5.3 Proof of Lemma C.3

We proceed in steps.

1. As in the proof of Lemma C.2, we rewrite the numerator and denominator as

$$\begin{aligned}\mathbb{E}\{\Delta^e H(R)\} &= \mathbb{E}\{H(R) \mid S = e, D = 1\} - \mathbb{E}\{H(R) \mid S = e, D = 0\} \\ \mathbb{E}\{\Delta^o H(R)\} &= \mathbb{E}\{H(R) \mid S = o, D = 0, Y = 1\} - \mathbb{E}\{H(R) \mid S = o, D = 0, Y = 0\}.\end{aligned}$$

2. Next, we analyze the numerator. As in the proof of Lemma C.2,

$$\mathbb{E}\{\Delta^e H(R)\} = \tau(m_{0,1} - m_{0,0}) + b(H),$$

where $m_{d,y} = \mathbb{E}\{H(R) \mid S = e, D = d, Y = y\}$.

3. Finally, we analyze the denominator. Using

$$m_{0,y} = \mathbb{E}\{H(R) \mid S = o, D = 0, Y = y\} + s_y(H),$$

we have that

$$\begin{aligned} m_{0,1} - m_{0,0} &= \mathbb{E}\{H(R) \mid S = o, D = 0, Y = 1\} + s_1(H) - \mathbb{E}\{H(R) \mid S = o, D = 0, Y = 0\} - s_0(H) \\ &= \mathbb{E}\{\Delta^o H(R)\} + s(H). \end{aligned}$$

Collecting results,

$$\mathbb{E}\{\Delta^e H(R)\} = \tau[\mathbb{E}\{\Delta^o H(R)\} + s(H)] + b(H).$$

Rearranging gives the desired conclusion.

D.5.4 Proof of Theorem C.6

For the first claim, Lemma C.3 gives

$$\tau = \frac{\mathbb{E}\{\Delta^e H(R)\} - b(H)}{\mathbb{E}\{\Delta^o H(R)\} + s(H)}.$$

Since $b(H)$ is a convex combination of $b_0(H)$ and $b_1(H)$, $|b(H)| \leq \bar{b}(H)$. By the triangle inequality, $|s(H)| \leq 2\bar{s}(H)$.

For the second claim, suppose $|\mathbb{E}\{\Delta^o H(R)\}| > 2\bar{s}(H)$. The reverse triangle inequality gives

$$\begin{aligned} |\mathbb{E}\{\Delta^o H(R)\} + s(H)| &\geq |\mathbb{E}\{\Delta^o H(R)\}| - |s(H)| \\ &\geq |\mathbb{E}\{\Delta^o H(R)\}| - 2\bar{s}(H) \\ &> 0. \end{aligned}$$

Substituting $\mathbb{E}\{\Delta^e H(R)\} = \widetilde{\tau}(H)\mathbb{E}\{\Delta^o H(R)\}$ into the expression for τ, we obtain

$$\tau - \widetilde{\tau}(H) = -\frac{b(H) + \widetilde{\tau}(H)s(H)}{\mathbb{E}\{\Delta^o H(R)\} + s(H)}.$$

Therefore, by the previous bound,

$$|\tau - \widetilde{\tau}(H)| \leq \frac{|b(H)| + |\widetilde{\tau}(H)||s(H)|}{|\mathbb{E}\{\Delta^o H(R)\} + s(H)|} \leq \frac{\bar{b}(H) + 2\bar{s}(H)|\widetilde{\tau}(H)|}{|\mathbb{E}\{\Delta^o H(R)\}| - 2\bar{s}(H)} = \kappa(H).$$

D.6 Continuous Outcomes

D.6.1 Proof of Proposition C.3

Abbreviate each bin as $B_k := B_\varepsilon(y_k)$. We proceed in steps.

1. By the proof of Lemma 1,

$$f_R(r|S=e,D=d) = \int_{y\in\mathcal{Y}} f_R(r|S=o,Y=y) f_{Y(d)}(y|S=e)\mathrm{d}y.$$

Partitioning $\mathcal{Y}$ into $\{B_k\}_{k=1}^K$ gives

$$f_R(r\,|\,S=e,D=d) = \sum_{k=1}^K \int_{y\in B_k} f_R(r\,|\,S=o,Y=y) f_{Y(d)}(y|S=e)\mathrm{d}y.$$

2. Fix k and $y\in B_k$. Assumptions 3(ii) and C.14 imply

$$\sup_{r\in\mathcal{R}} |f_R(r\,|\,S=o,Y=y) - f_R(r\,|\,S=o,Y\in B_k)| \le \omega(2\varepsilon).$$

Therefore, adding and subtracting then applying the triangle inequality gives

$$\begin{aligned}
&f_R(r\,|\,S=e,D=d)\\
&= \sum_{k=1}^K \int_{y\in B_k} \{f_R(r\,|\,S=o,Y=y) - f_R(r\,|\,S=o,Y\in B_k) + f_R(r\,|\,S=o,Y\in B_k)\} f_{Y(d)}(y|S=e)\mathrm{d}y\\
&= \mathrm{err}_\varepsilon(r,d) + \sum_{k=1}^K \int_{y\in B_k} f_R(r\,|\,S=o,Y\in B_k) f_{Y(d)}(y|S=e)\mathrm{d}y\\
&= \mathrm{err}_\varepsilon(r,d) + \sum_{k=1}^K f_R(r\,|\,S=o,Y\in B_k)\Pr\{Y(d)\in B_k|S=e\},
\end{aligned}$$

where, uniformly over $d\in\{0,1\}$ and $r\in\mathcal{R}$,

$$|\mathrm{err}_\varepsilon(r,d)| \le \omega(2\varepsilon).$$

3. Subtracting this result across $d\in\{1,0\}$ and appealing to the definition of θ_ε gives

$$\begin{aligned}
&f_R(r\,|\,S=e,D=1) - f_R(r\,|\,S=e,D=0)\\
&= \mathrm{err}_\varepsilon(r,1) - \mathrm{err}_\varepsilon(r,0) + \sum_{k=1}^K f_R(r\,|\,S=o,Y\in B_k)[\Pr\{Y(1)\in B_k|S=e\} - \Pr\{Y(0)\in B_k|S=e\}]\\
&= \mathrm{err}_\varepsilon(r) + \sum_{k=1}^K f_R(r\,|\,S=o,Y\in B_k)\theta_{\varepsilon,k},
\end{aligned}$$

where, uniformly over $r\in\mathcal{R}$,

$$|\mathrm{err}_\varepsilon(r)| \le |\mathrm{err}_\varepsilon(r,1)| + |\mathrm{err}_\varepsilon(r,0)| \le 2\omega(2\varepsilon).$$

4. By Bayes' rule,

$$f_R(r \mid S=e, D=d) = \mathbb{E}\left[\frac{1\{D=d, S=e\}}{\Pr(D=d, S=e)} \,\middle|\, R=r\right] f_R(r)$$
$$f_R(r \mid S=o, Y \in B_k) = \mathbb{E}\left[\frac{1\{Y \in B_k, S=o\}}{\Pr(Y \in B_k, S=o)} \,\middle|\, R=r\right] f_R(r).$$

By the proof of Theorem 1, we conclude that

$$\mathbb{E}\left\{\Delta^e - (\Delta_\varepsilon^o)^\top \theta_\varepsilon \mid R=r\right\} = \frac{\mathrm{err}_\varepsilon(r)}{f_R(r)},$$

proving the desired result since Assumption C.14 implies $|f_R(r)| \geq \underline{\ell}$.

D.6.2 Proof of Corollary C.5

To prove this result, it is useful to introduce notation for the discretized outcome and discretized potential outcomes. Define $Y_\varepsilon = \sum_{k=1}^K y_k 1\{Y \in B_\varepsilon(y_k)\}$. Analogously define $Y_\varepsilon(d) = \sum_{k=1}^K y_k 1\{Y(d) \in B_\varepsilon(y_k)\}$ for $d \in \{0,1\}$. By construction, $|Y - Y_\varepsilon| \leq \varepsilon$ almost surely and $|Y(d) - Y_\varepsilon(d)| \leq \varepsilon$ almost surely.

We proceed in steps.

1. For any $d \in \{0,1\}$, observe

$$|\mathbb{E}\{Y(d) \mid S=e\} - \mathbb{E}\{Y_\varepsilon(d) \mid S=e\}| \leq \mathbb{E}\{|Y(d) - Y_\varepsilon(d)| \mid S=e\} \leq \varepsilon.$$

Defining $\tau_\varepsilon = \mathbb{E}\{Y_\varepsilon(1) - Y_\varepsilon(0) \mid S=e\}$, it follows that

$$|\tau - \tau_\varepsilon| = |[\mathbb{E}\{Y(1) \mid S=e\} - \mathbb{E}\{Y_\varepsilon(1) \mid S=e\}] - [\mathbb{E}\{Y(0) \mid S=e\} - \mathbb{E}\{Y_\varepsilon(0) \mid S=e\}]| \leq 2\varepsilon.$$

2. Defining $U_\varepsilon = \Delta^e - (\Delta_\varepsilon^o)^\top \theta_\varepsilon$, Proposition C.3 establishes

$$|\mathbb{E}(U_\varepsilon \mid R)| \leq \frac{2\omega(2\varepsilon)}{\underline{\ell}}.$$

Therefore by the law of iterated expectations,

$$\begin{aligned}
\|\mathbb{E}\{H(R)U_\varepsilon\}\| &= \|\mathbb{E}\{H(R)\mathbb{E}(U_\varepsilon|R)\}\| \\
&\leq \mathbb{E}\{\|H(R)\mathbb{E}(U_\varepsilon|R)\|\} \\
&\leq \mathbb{E}\{|\mathbb{E}(U_\varepsilon|R)| \cdot \|H(R)\|\} \\
&\leq \frac{2\omega(2\varepsilon)}{\underline{\ell}} \mathbb{E}\{\|H(R)\|\}.
\end{aligned}$$

3. As notation, let $\widetilde{\theta}_\varepsilon = [\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}]^{-1}\mathbb{E}\{H(R)\Delta^e\}$. By definition of $\widetilde{\theta}_\varepsilon$,

$$\mathbb{E}\{H(R)\Delta^e\} = \mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}\widetilde{\theta}_\varepsilon.$$

By definition of U_ε,

$$\mathbb{E}\{H(R)\Delta^e\} = \mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}\theta_\varepsilon + \mathbb{E}\{H(R)U_\varepsilon\}.$$

Combining these expressions,

$$\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}\widetilde{\theta}_\varepsilon = \mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}\theta_\varepsilon + \mathbb{E}\{H(R)U_\varepsilon\}.$$

Since $\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}$ is invertible, it therefore follows that

$$\widetilde{\theta}_\varepsilon - \theta_\varepsilon = [\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}]^{-1}\mathbb{E}\{H(R)U_\varepsilon\}.$$

We conclude that

$$\begin{aligned}\|\widetilde{\theta}_\varepsilon - \theta_\varepsilon\| &\leq \|[\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}]^{-1}\|_{\text{op}}\|\mathbb{E}\{H(R)U_\varepsilon\}\| \\ &\leq \|[\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}]^{-1}\|_{\text{op}}\frac{2\omega(2\varepsilon)}{\underline{\ell}}\mathbb{E}\{\|H(R)\|\}.\end{aligned}$$

4. We translate this bound for the discretized parameter into a bound for the discretized average treatment effect. By the Cauchy-Schwarz inequality,

$$\begin{aligned}|\widetilde{\tau}_\varepsilon - \tau_\varepsilon| &= |\lambda_\varepsilon^\top(\widetilde{\theta}_\varepsilon - \theta_\varepsilon)| \\ &\leq \|\lambda_\varepsilon\| \cdot \|\widetilde{\theta}_\varepsilon - \theta_\varepsilon\| \\ &\leq \|\lambda_\varepsilon\| \cdot \|[\mathbb{E}\{H(R)(\Delta_\varepsilon^o)^\top\}]^{-1}\|_{\text{op}} \cdot \frac{2\omega(2\varepsilon)}{\underline{\ell}}\mathbb{E}\{\|H(R)\|\}\end{aligned}$$

5. Finally, we appeal to the triangle inequality: $|\widetilde{\tau}_\varepsilon - \tau| \leq |\widetilde{\tau}_\varepsilon - \tau_\varepsilon| + |\tau_\varepsilon - \tau|$. □

D.6.3 Proof of Proposition C.4

The derivation of the conditional moment equation is identical to the proof of Lemma 1 and step one in the proof of Theorem 1, using the generalized outcome weights instead of the previously defined outcome weights.

By standard arguments in Kress (1989), Assumption C.15 guarantees existence of a solution to this equation. Assumption C.16 guarantees uniqueness of the solution to this equation. □

D.7 Estimation with Covariates

D.7.1 Proof of Proposition C.5

The proof follows the argument of Proposition 3 within each covariate cell $x \in \mathcal{X}$. □

D.7.2 Proof of Proposition C.6

The proof follows the argument of Proposition 4 within each covariate cell $x \in \mathcal{X}$. □

E Additional Estimation Algorithms

E.1 Estimation with Covariates

Algorithm E.1 Estimation with Discrete Outcomes and Discrete Covariates

Require: Data $\{S_i, X_i, 1_{S_i=e} D_i, 1_{S_i=o} Y_i, R_i\}_{i=1}^n$ with $Y \in \mathcal{Y} = \{y_1,...,y_K\}$ and $X \in \mathcal{X} = \{x_1,...,x_{|\mathcal{X}|}\}$.

Ensure: $\widehat{\tau}$ and a confidence interval.

Randomly split indices into a TRAIN fold and a TEST fold.

for all $x \in \mathcal{X}$ **do**

Step 1: Learn a representation on TRAIN.

Compute $\mathbb{E}_{\text{TRAIN}}\left(1_{D_i=d,S_i=e} \,|\, X_i = x\right)$ for $d \in \{0,1\}$ and $\mathbb{E}_{\text{TRAIN}}\left(1_{Y_i=y_k,S_i=o} \,|\, X_i = x\right)$ for $k = 1,...,K$.

Train $\text{PRED}_Y(x,\cdot) : \mathcal{R} \to [0,1]^K$ on $\{(R_i, Y_i) : i \in \text{TRAIN}, X_i = x, S_i = o\}$ to estimate $\Pr(Y = y_k \,|\, S = o, X = x, R)$.

Train $\text{PRED}_D(x,\cdot) : \mathcal{R} \to [0,1]$ on $\{(R_i, D_i) : i \in \text{TRAIN}, X_i = x, S_i = e\}$ to estimate $\Pr(D = 1 \,|\, S = e, X = x, R)$.

Train $\text{PRED}_S(x,\cdot) : \mathcal{R} \to [0,1]$ on $\{(R_i, S_i) : i \in \text{TRAIN}, X_i = x\}$ to estimate $\Pr(S = e \,|\, X = x, R)$.

Define, for generic (x, R),

$$\widehat{\mathbb{E}}\{\Delta^e(x) \,|\, x, R\} = \left\{ \frac{\text{PRED}_D(x,R)}{\mathbb{E}_{\text{TRAIN}}(1_{D=1,S=e} \,|\, X = x)} - \frac{1 - \text{PRED}_D(x,R)}{\mathbb{E}_{\text{TRAIN}}(1_{D=0,S=e} \,|\, X = x)} \right\} \text{PRED}_S(x,R),$$

and for each $k = 1,...,K-1$,

$$\widehat{\mathbb{E}}\{\Delta^o_k(x) \,|\, x, R\} = \left\{ \frac{\text{PRED}_{Y,k}(x,R)}{\mathbb{E}_{\text{TRAIN}}(1_{Y=y_k,S=o} \,|\, X = x)} - \frac{\text{PRED}_{Y,K}(x,R)}{\mathbb{E}_{\text{TRAIN}}(1_{Y=y_K,S=o} \,|\, X = x)} \right\} \{1 - \text{PRED}_S(x,R)\}.$$

Stack $\widehat{\mathbb{E}}\{\Delta^o(x) \,|\, x, R\} = (\widehat{\mathbb{E}}\{\Delta^o_1(x) \,|\, x, R\},...,\widehat{\mathbb{E}}\{\Delta^o_{K-1}(x) \,|\, x, R\})^\top$.

Compute

$$\widehat{\theta}_{\text{INIT}}(x) = \arg\min_{\theta \in \mathbb{R}^{K-1}} \sum_{i \in \text{TRAIN}: X_i = x} \left[\widehat{\mathbb{E}}\{\Delta^e(x) \,|\, x, R_i\} - \widehat{\mathbb{E}}\{\Delta^o(x) \,|\, x, R_i\}^\top \theta \right]^2.$$

Compute $\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}}, x, R)$ using Lemma E.1 by substituting in the predictors above.

Set $\widehat{H}(x,R) = \frac{\widehat{\mathbb{E}}\{\Delta^o(x) | x,R\}}{\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}}, x, R)} \in \mathbb{R}^{K-1}$.

Step 2: Estimate $\widehat{\theta}$ on TEST.

Compute $\mathbb{E}_{\text{TEST}}\left(1_{D_i=d,S_i=e} \,|\, X_i = x\right)$ for $d \in \{0,1\}$ and $\mathbb{E}_{\text{TEST}}\left(1_{Y_i=y_k,S_i=o} \,|\, X_i = x\right)$ for $k = 1,...,K$.

for all $i \in \text{TEST}$ with $X_i = x$ **do**

Set

$$\widehat{\Delta}^e_i(x) = \frac{1_{D_i=1,S_i=e}}{\mathbb{E}_{\text{TEST}}(1_{D=1,S=e} \,|\, X = x)} - \frac{1_{D_i=0,S_i=e}}{\mathbb{E}_{\text{TEST}}(1_{D=0,S=e} \,|\, X = x)}.$$

for $k = 1,...,K-1$ **do**

Set

$$\widehat{\Delta}^o_{ik}(x) = \frac{1_{Y_i=y_k,S_i=o}}{\mathbb{E}_{\text{TEST}}(1_{Y=y_k,S=o} \,|\, X = x)} - \frac{1_{Y_i=y_K,S_i=o}}{\mathbb{E}_{\text{TEST}}(1_{Y=y_K,S=o} \,|\, X = x)}.$$

end for

Stack $\widehat{\Delta}^o_i(x) = \{\widehat{\Delta}^o_{i1}(x),...,\widehat{\Delta}^o_{i(K-1)}(x)\}^\top$.

end for

Compute

$$\widehat{\theta}(x) = \left\{ \sum_{i \in \text{TEST}: X_i = x} \widehat{H}(x,R_i) \widehat{\Delta}^o_i(x)^\top \right\}^{-1} \left\{ \sum_{i \in \text{TEST}: X_i = x} \widehat{H}(x,R_i) \widehat{\Delta}^e_i(x) \right\}.$$

end for

Step 3: Compute the ATE and standard error.

Compute $\widehat{\theta} = \mathbb{E}_{\text{TEST}}\{\widehat{\theta}(X) \,|\, S = e\}$, i.e the empirical mean over $i \in \text{TEST}$ with $S_i = e$.

Compute $\widehat{\tau} = \sum_{k=1}^{K-1} (y_k - y_K) \widehat{\theta}_k$.

Bootstrap the standard error $\widehat{v}$ of $\widehat{\tau}$, holding $\widehat{H}$ fixed.

Return: the confidence interval $\widehat{\tau} \pm z_{1-\alpha/2}\, \widehat{v}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

Algorithm E.2 Estimation with Discrete Outcomes and Continuous Covariates

Require: Data $\{S_i, X_i, 1_{S_i=e} D_i, 1_{S_i=o} Y_i, R_i\}_{i=1}^n$ with $Y \in \mathcal{Y} = \{y_1, \ldots, y_K\}$ and $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$.

Require: Basis $b : \mathcal{X} \to \mathbb{R}^J$, an initial representation $H_{\text{INIT}}(X,R) \in \mathbb{R}^{K-1}$, and the link function $\psi(\cdot)$ such that $\theta(x) = \psi\{g(x)\}$.

Notation. Consider any representation H and basis b. Write $H_i = H(X_i, R_i) \in \mathbb{R}^{K-1}$ and $B_i = b(X_i) \in \mathbb{R}^J$. Construct the vector $C_i(H) \in \mathbb{R}^{K(K+2)}$ as the transpose of

$$\left(1_{D_i=1,S_i=e} \quad 1_{D_i=1,S_i=e} H_i^\top \quad 1_{D_i=0,S_i=e} \quad 1_{D_i=0,S_i=e} H_i^\top \quad 1_{Y_i=y_1,S_i=o} \quad 1_{Y_i=y_1,S_i=o} H_i^\top \quad \cdots \quad 1_{Y_i=y_K,S_i=o} \quad 1_{Y_i=y_K,S_i=o} H_i^\top\right)$$

where the odd entries are scalars and the even entries are vectors in $\mathbb{R}^{K-1}$. Define the series estimator

$$\widehat{g}_{\text{TRAIN}}(x;H) = b(x)^\top \widehat{\beta}_{\text{TRAIN}}(H), \quad \widehat{\beta}_{\text{TRAIN}}(H) = \operatorname*{argmin}_\beta \sum_{i \in \text{TRAIN}} \|C_i(H) - B_i^\top \beta\|^2,$$

and likewise define $\widehat{g}_{\text{TEST}}(x;H)$

1: Randomly split indices into a TRAIN fold and a TEST fold.

Step 1: Learn a representation on TRAIN.

2: Train $\text{PRED}_Y(\cdot,\cdot) : \mathcal{X} \times \mathcal{R} \to [0,1]^K$ on $\{(X_i,R_i,Y_i) : i \in \text{TRAIN}, S_i = o\}$ to estimate $\Pr(Y = y_k \mid S = o, X, R)$.

3: Train $\text{PRED}_D(\cdot,\cdot) : \mathcal{X} \times \mathcal{R} \to [0,1]$ on $\{(X_i,R_i,D_i) : i \in \text{TRAIN}, S_i = e\}$ to estimate $\Pr(D = 1 \mid S = e, X, R)$.

4: Train $\text{PRED}_S(\cdot,\cdot) : \mathcal{X} \times \mathcal{R} \to [0,1]$ on $\{(X_i,R_i,S_i) : i \in \text{TRAIN}\}$ to estimate $\Pr(S = e \mid X, R)$.

5: Compute $\widehat{g}_{\text{TRAIN}}(x) = \widehat{g}_{\text{TRAIN}}(x; H_{\text{INIT}})$.

6: Let $\widehat{\mathbb{E}}_{\text{TRAIN}}(1_{D=d,S=e} \mid X)$ and $\widehat{\mathbb{E}}_{\text{TRAIN}}(1_{Y=y_k,S=o} \mid X)$ denote the corresponding components of $\widehat{g}_{\text{TRAIN}}(X)$.

7: Define, for generic (X,R),

$$\widehat{\mathbb{E}}\{\Delta^e(X) \mid X, R\} = \left\{ \frac{\text{PRED}_D(X,R)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1_{D=1,S=e} \mid X)} - \frac{1 - \text{PRED}_D(X,R)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1_{D=0,S=e} \mid X)} \right\} \text{PRED}_S(X,R),$$

and for each $k = 1, \ldots, K-1$,

$$\widehat{\mathbb{E}}\{\Delta_k^o(x) \mid X, R\} = \left\{ \frac{\text{PRED}_{Y,k}(X,R)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1_{Y=y_k,S=o} \mid X)} - \frac{\text{PRED}_{Y,K}(X,R)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1_{Y=y_K,S=o} \mid X)} \right\} \{1 - \text{PRED}_S(X,R)\}.$$

8: Stack $\widehat{\mathbb{E}}\{\Delta^o(X) \mid X, R\} = (\widehat{\mathbb{E}}\{\Delta_1^o(X) \mid X, R\}, \ldots, \widehat{\mathbb{E}}\{\Delta_{K-1}^o(X) \mid X, R\})^\top$.

9: Set $\widehat{\theta}_{\text{INIT}}(x) = \psi\{\widehat{g}_{\text{TRAIN}}(x)\}$.

10: Compute $\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}}, X, R)$ using Lemma E.1 by substituting in the predictors above.

11: Set $\widehat{H}(X,R) = \frac{\widehat{\mathbb{E}}\{\Delta^o(x) \mid X, R\}}{\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}}, X, R)} \in \mathbb{R}^{K-1}$.

Step 2: Estimate θ on TEST.

12: Compute $\widehat{g}_{\text{TEST}}(x) = \widehat{g}_{\text{TEST}}(x; \widehat{H})$.

13: Set $\widehat{\theta}(x) = \psi\{\widehat{g}_{\text{TEST}}(x)\}$.

Step 3: Compute the ATE and standard error.

14: Compute $\widehat{\tau}(x) = \sum_{k=1}^{K-1} (y_k - y_K) \widehat{\theta}_k(x)$.

15: Compute $\widehat{\tau} = \mathbb{E}_{\text{TEST}}\{\widehat{\tau}(X) \mid S = e\}$, i.e. the empirical mean over $i \in \text{TEST}$ with $S_i = e$.

16: Bootstrap the standard error of $\widehat{\tau}$ on TEST, holding $\widehat{H}$ fixed.

17: **Return:** the confidence interval $\widehat{\tau} \pm z_{1-\alpha/2}\, \widehat{v}$ for $z_{1-\alpha/2}$ the $(1-\alpha/2)$ quantile of a standard normal distribution.

E.2 An Algebraic Simplification

The following lemma provides algebra that simplifies the denominator of the optimal representation.

Lemma E.1 (An algebraic simplification)**.** By construction,

$$\{\Delta^e(X)-\Delta^o(X)^\top\theta(X)\}^2=\frac{1_{D=1,S=e}}{\Pr(D=1,S=e\,|\,X)^2}+\frac{1_{D=0,S=e}}{\Pr(D=0,S=e\,|\,X)^2}$$
$$+\sum_{k=1}^{K-1}\theta_k^2(X)\left\{\frac{1_{Y=y_k,S=o}}{\Pr(Y=y_k,S=o\,|\,X)^2}+\frac{1_{Y=y_K,S=o}}{\Pr(Y=y_K,S=o\,|\,X)^2}\right\}$$
$$+\sum_{j\neq k}\theta_j(X)\theta_k(X)\frac{1_{Y=y_K,S=o}}{\Pr(Y=y_K,S=o\,|\,X)^2}.$$

Proof. To lighten notation, we suppress the arguments of Δ^e, Δ^o, and θ_j. Observe that

$$\{\Delta^e-(\Delta^o)^\top\theta\}^2=(\Delta^e)^2-2\{(\Delta^o)^\top\theta\}\Delta^e+\{(\Delta^o)^\top\theta\}^2=(\Delta^e)^2+\{(\Delta^o)^\top\theta\}^2,$$

since the events in Δ^e are exclusive of the events in Δ^o. Specifically, $1_{S=e}1_{S=o}=0$.

Consider the former term. Since $1_{D=1}1_{D=0}=0$, we have by a similar logic that

$$(\Delta^e)^2=\frac{1_{D=1,S=e}}{\Pr(D=1,S=e\,|\,X)^2}+\frac{1_{D=0,S=e}}{\Pr(D=0,S=e\,|\,X)^2}.$$

Consider the latter term. Developing the square,

$$\{(\Delta^o)^\top\theta\}^2=\left(\sum_{k=1}^{K-1}\Delta_k^o\theta_k\right)^2=\sum_{j,k}\theta_j\theta_k\Delta_j^o\Delta_k^o.$$

When $j=k<K$, the term becomes $\theta_k^2(\Delta_k^o)^2$. Since $1_{Y=y_k}1_{Y=y_K}=0$, we have

$$(\Delta_k^o)^2=\frac{1_{Y=y_k,S=o}}{\Pr(Y=y_k,S=o\,|\,X)^2}+\frac{1_{Y=y_K,S=o}}{\Pr(Y=y_K,S=o\,|\,X)^2}.$$

When $j\neq k$ and $j,k<K$, the term becomes $\theta_j\theta_k\Delta_j^o\Delta_k^o$. Since $1_{Y=y_j}1_{Y=y_k}=0$, $1_{Y=y_j}1_{Y=y_K}=0$, and $1_{Y=y_k}1_{Y=y_K}=0$, we have

$$\Delta_j^o\Delta_k^o=\frac{1_{Y=y_K,S=o}}{\Pr(Y=y_K,S=o\,|\,X)^2}.$$

□

F Additional Simulation Details

F.1 Forest Cover Details

We utilize forest cover measurements from Hansen et al. (2013) for Uganda, which are derived from Landsat satellite imagery at roughly 30-meter resolution. In the Hansen et al. (2013) product, tree cover is defined as the fraction of area within a pixel containing vegetation taller than 5 meters. We aggregate these pixel-level measurements to a common spatial grid and then construct the outcome used in our simulations.

The unit of analysis is a $0.01^\circ \times 0.01^\circ$ grid cell (approximately 1 km $\times$ 1 km) covering Uganda. All variables entering the simulations are aggregated to, or spatially aligned with, this grid. For each grid cell, we compute an average of the underlying 30-meter tree-cover pixels to obtain a grid cell-level tree-cover share. We then define a binary outcome $Y \in \{0,1\}$ indicating whether at least 80% of the grid cell is forested. As noted in the main text, we use this high-forest threshold to align with conservation policies and program rules that are commonly expressed in terms of forest-cover retention requirements, such as mandates under Brazil's Forest Code (Azevedo et al., 2017) and eligibility criteria in payments for ecosystem services programs (Wunder et al., 2008).

The geographic samples used in our simulations are partitioned using the randomized study conducted by Jayachandran et al. (2017) in Uganda. We define the experimental sample as the set of grid cells falling within the geographic boundaries of the original study area. We define observational samples using surrounding geographic bands around the experimental region (0–2 km, 0–5 km, and 0–10 km). Appendix Figure F.1 provides a map of these regions. In the main text, we focus on the 0–2 km band.

For each 0.01° grid cell, we obtain a 4,000-dimensional vector of MOSAIKS satellite embeddings (Rolf et al., 2021), which are derived from publicly available satellite imagery and are distinct from the Landsat imagery underlying the Hansen et al. (2013) forest-cover measurements. We merge these embeddings to the grid cell-level forest-cover outcome by geographic coordinates.

We then train a random forest to predict Y from the MOSAIKS embeddings using grid cells from the rest of Uganda, excluding both the experimental region and the surrounding observational bands. The resulting predictor achieves high predictive accuracy, with an area under the curve (AUC) of 0.967 on held-out grid cells. We define the remotely sensed variable R as the scalar probability output of this trained random forest.

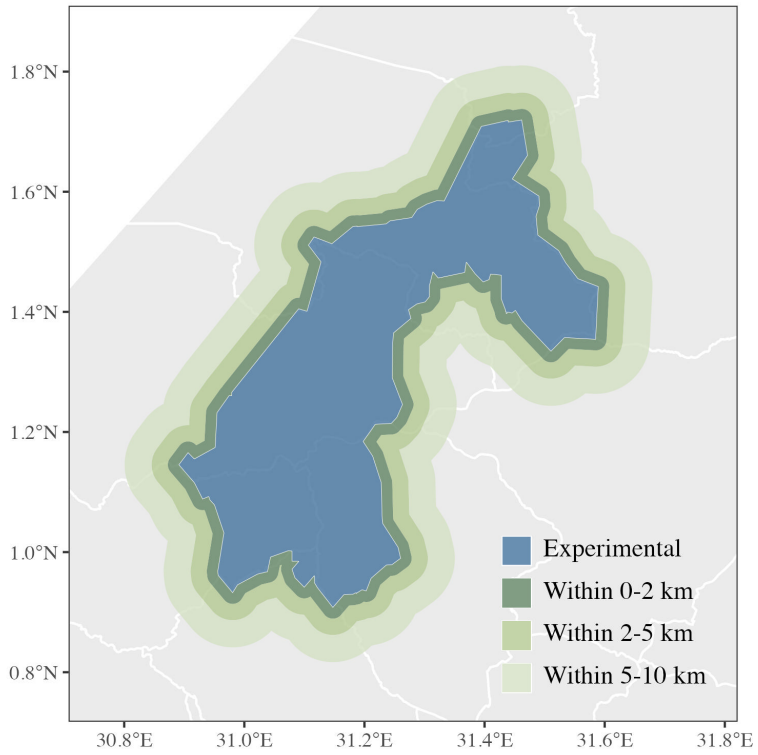


Figure F.1: We illustrate the experimental and observational samples in the Uganda forest cover data. The experimental sample consists of the grid cells covered by the payments for ecosystem services experiment (Jayachandran et al., 2017). The observational sample consists of grid cells in surrounding geographic bands. In the main text, we focus on the 0–2 kilometer geographic band.

F.2 Smartcards Details

Randomization took place at the level of the mandal, i.e. subdistrict, of Andhra Pradesh, India. Muralidharan et al. (2023) partition 396 mandals into the following subgroups:

- treated mandals (111), randomly assigned to receive Smartcards in 2010;
- buffer mandals (136), randomly assigned to receive Smartcards in 2011;
- untreated mandals (44), randomly assigned to receive Smartcards in 2012;
- non-study mandals (105), which were excluded from the experiment.

We study villages within mandals as the units of analysis, where villages are defined by Asher et al. (2021). For each village, our treatment D indicates whether the village received Smartcards in 2010. We interpret villages within the treated mandals as treated experimental units; villages within the buffer and untreated mandals as untreated experimental units; and

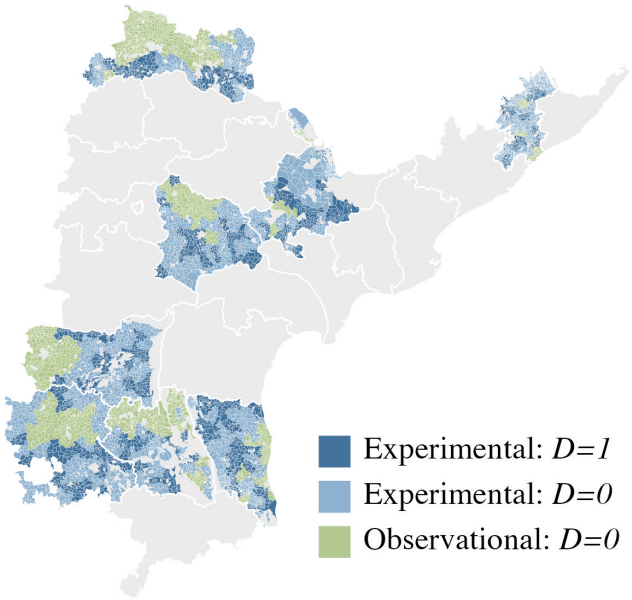


Figure F.2: We illustrate the experimental and observational samples in the Smartcards simulation (Muralidharan et al., 2023).

villages within the non-study mandals as observational units with missing treatment. Finally, we drop villages with fewer than 100 individuals. This removes about 2% of villages.

Appendix Figure F.2 illustrates our village classification. Appendix Table F.1 summarizes village characteristics. The causal parameter is the effect of early adoption (2010 Smartcards) for villages in the experiment.

Table F.1: Summary statistics for the Smartcards experiment.

Sample	Non-Study Mandals	Untreated Mandals	Buffer Mandals	Treated Mandals
Smartcards Rollout	N/A	2012	2011	2010
Number of villages	2257	852	2929	2274
Average population	1985	2255	2105	2116
Average fraction female	0.491	0.492	0.496	0.497

For each village, we collect poverty measurements to serve as the outcome. Following Asher and Novosad (2020), the data sources are the 2012-2013 Socio-Economic and Caste Census (SECC), and the 2013 Indian Economic Census. We consider three outcomes which we construct directly from these sources. The consumption outcome variable indicates whether a village's per capita consumption is in the bottom quartile. We consider two additional outcome variables based on income: does a village have only low income households, i.e. no earner making above 5,000 rupees; and does a village only low and middle income households, i.e.

no earner making above 10,000 rupees. The definitions of low and middle income households are from the SECC.

For each village, we extract satellite images to serve as the remotely sensed variable. First, we extract coordinates for the perimeter of the village (Asher et al., 2021). Then, we extract luminosity measures, a vector in $\mathbb{R}^{50}$, which includes the minimum, maximum, mean, and sum of night light within each polygon, along with the total number of pixels in the polygon, from 2012 to 2020 (Asher et al., 2021). Finally, we extract satellite images from 2019, summarized as a high-dimensional, pre-trained embedding vector in $\mathbb{R}^{4000}$ (Rolf et al., 2021). The concatenation of these objects is our remotely sensed variable R.

F.3 Plausibility of Identifying Assumptions

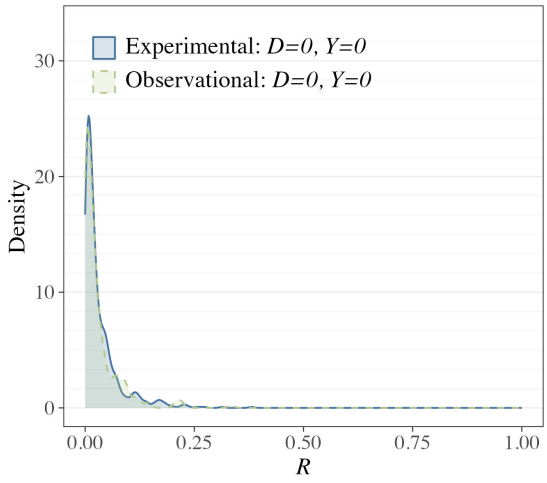
We assess whether stability (Assumption 2) is plausible in each empirical setting. Here, stability requires that four conditional distributions of the remotely sensed variable should be the same across samples: $f_R(r \mid S=e,D=d,Y=y) = f_R(r \mid S=o,D=d,Y=y)$, for each (d,y) stratum (or in the absence of direct effects, only for the $(d=0,y=0),(d=0,y=1)$ stratum). Using "oracle" data access, we can visually assess this condition. Specifically, we consider the outcomes of untreated units in the experimental and observational samples.[27] This allows us to directly evaluate the conditions for $d=0$ and $y \in \{0,1\}$ by subsetting and visualizing densities.

Figure F.3 presents the visual evidence. In the forest cover setting (Panel A), we compare the conditional distributions of R for the experimental sample (grid cells in the region studied by Jayachandran et al. (2017)) and for the observational sample (the two kilometer geographic band surrounding it). The conditional distributions reasonably align. In the Smartcards setting (Panel B), R is high-dimensional, so we plot the conditional distributions of its first principal component, comparing experimental and observational villages. The conditional distributions almost coincide.

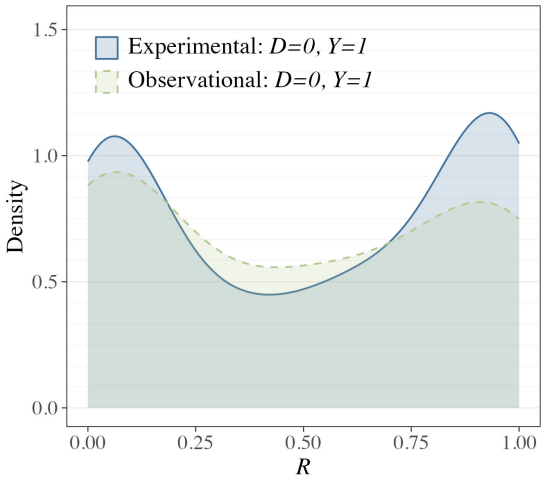
To complement the visual evidence, we conduct Kolmogorov-Smirnov (KS) tests for equality of the conditional distributions across samples. In all comparisons, we cannot reject the null of equal distributions at the 5% level. Taken together, visual and statistical evidence supports the plausibility of our main identifying assumption in these two empirical settings.

[27] In the forest cover data, neither sample involves a treatment. In the Smartcards data, we observe outcomes for untreated experimental villages and for observational villages (which are all untreated).

Panel A: Forest Cover in Uganda

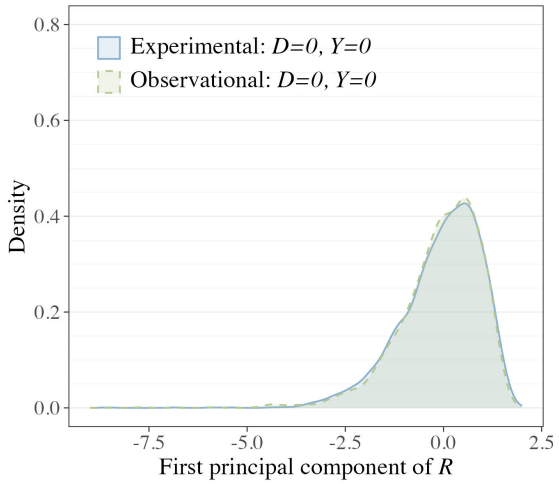


(a) Densities of $R \mid S,D=0,Y=0$.

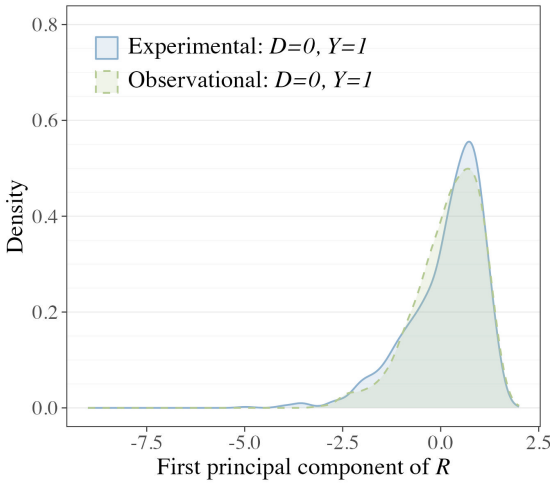


(b) Densities of $R \mid S,D=0,Y=1$.

Panel B: Smartcards Experiment in India



(c) Densities of $R \mid S,D=0,Y=0$.



(d) Densities of $R \mid S,D=0,Y=1$.

Figure F.3: Our stability assumption (Assumption 2(i)) is plausible in two empirical settings. We compare $f_R(R \mid S=e,D=d,Y=y)$ with $f_R(R \mid S=o,D=d,Y=y)$ for $d=0$ and $y \in \{0,1\}$. Because the remotely sensed variable is high-dimensional in the Smartcards experiment, we visualize the density of its standardized first principal component.

G Additional Simulation Results

In Section 6, we reported calibrated simulations for the Uganda forest cover data and the Smartcards experiment in the two sample setting (i.e., experimental and observational samples only). In this section, we provide more details, then extend our calibration simulations to incorporate a validation sample (Section 5.1). The extension facilitates comparison with prediction powered inference (PPI).

G.1 Experimental and Observational Samples Only

Simulation design. We observe (D,R) when $S=e$ and (R,Y) when $S=o$. Details are given in Section 6.1.

Implementation details. We compare two methods.

The two-step method trains a predictor of Y from R in the observational sample, then regresses predicted outcomes on D in the experimental sample. Its estimand is $\widetilde{\tau} \neq \tau$ (Section 3.2.1). In the forest cover design, we predict Y from R using logistic regression. In the Smartcards design, we predict Y from R using probability random forests. Specifically, we use the `R` package `ranger` with 100 trees and the package default options for all other tuning parameters.

For our method, we use cross-fitting with two folds. Its estimand is τ. In the forest cover design, we estimate the efficient representation by combining logistic regressions of $\widehat{\Pr}(Y=1 \mid S=o,R)$, $\widehat{\Pr}(D=1 \mid S=e,R)$, and $\widehat{\Pr}(S=e \mid R)$. In the Smartcards design, we do so by combining probability random forests. Specifically, we use the `R` package `ranger` with 100 trees and the package default options.

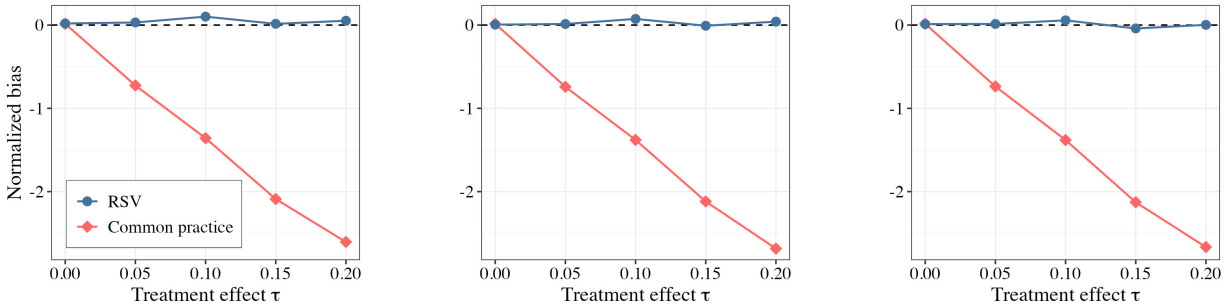
For computational tractability, we truncate the remotely sensed in the Smartcards design from $\mathbb{R}^{4050}$ to $\mathbb{R}^{1050}$. In particular, we use the initial 1,000 features of the pre-trained MOSAIKS embedding and the nighttime luminosity measures.

PPI methods are infeasible in this setting because we do not have joint observations of $(D=0,Y,R)$ and $(D=1,Y,R)$.

Additional results. We further compare our method and the two-step method in simulations, varying the baseline outcome probabilities and the sample sizes in the simulation designs. For these variations on the simulation designs, we find the same results as in the main text.

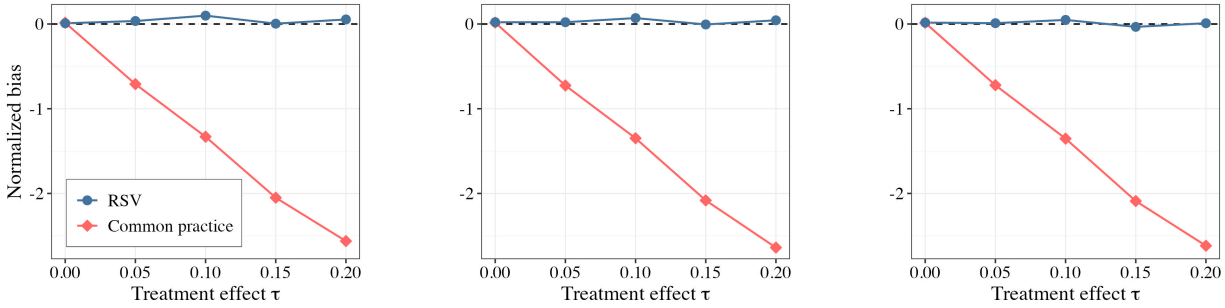
For the forest cover design, Figure G.1 reports the normalized bias and Figure G.2 reports the coverage of each method. The different panels vary the choice of α_o, by calibrating its value to different geographic bands around the experimental sample. The different plots vary the choice of n_o.

Panel A: 0–5 Kilometer Geographic Band



(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 2000$.

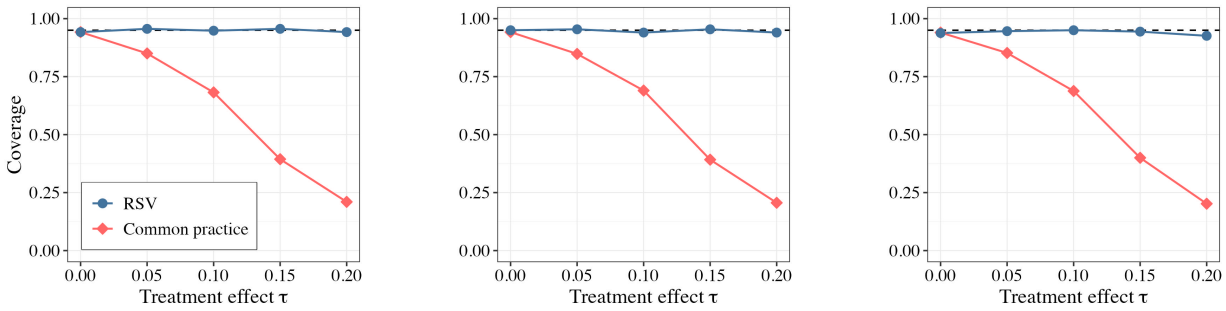
Panel B: 0–10 Kilometer Geographic Band



(d) $n_o = 500$. (e) $n_o = 1000$. (f) $n_o = 2000$.

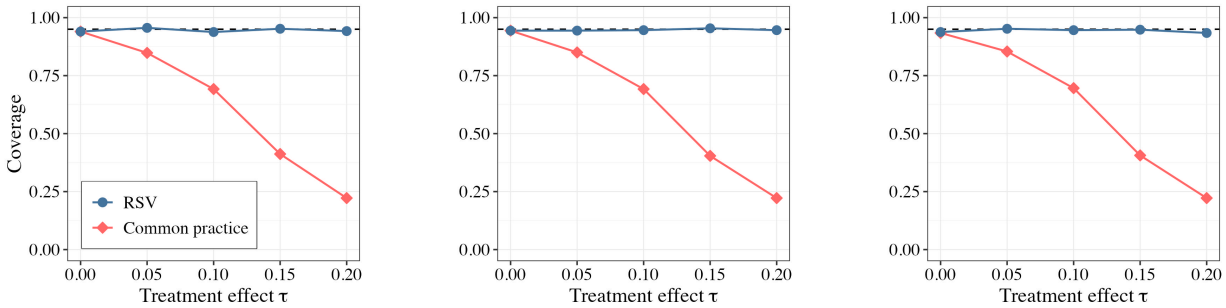
Figure G.1: In the forest cover design, our method outperforms the two-step method in terms of normalized average bias. The different panels vary α_o. The different plots vary n_o. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

Panel A: 0–5 Kilometer Geographic Band



(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 2000$.

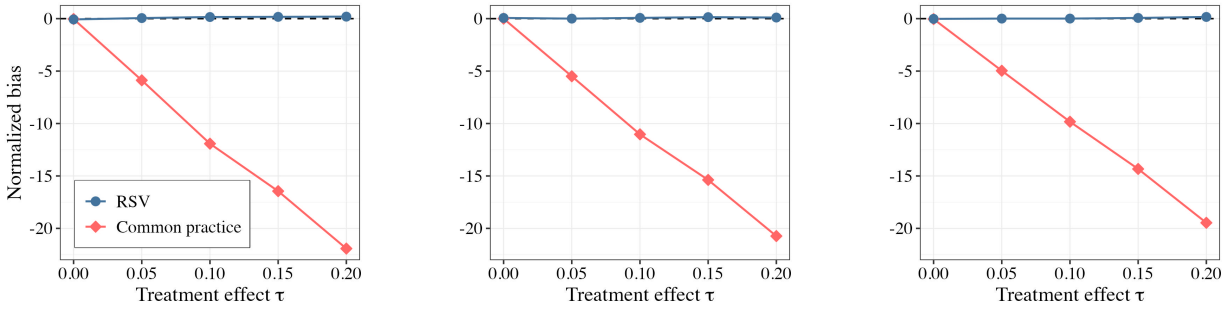
Panel B: 0–10 Kilometer Geographic Band



(d) $n_o = 500$. (e) $n_o = 1000$. (f) $n_o = 2000$.

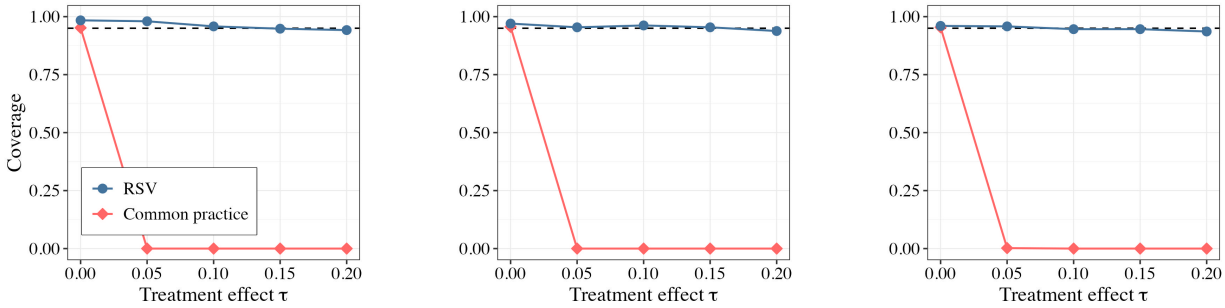
Figure G.2: In the forest cover design, our method outperforms the two-step method in terms of coverage. The different panels vary α_o. The different plots vary n_o. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

For the Smartcards design, Figure G.3 reports the normalized bias and Figure G.4 reports the coverage of each method. The different plots vary the choice of n_o.



(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 1500$.

Figure G.3: In the Smartcards design, our method outperforms the two-step method in terms of normalized average bias. The different plots vary n_o. For each value of the synthetic treatment effect τ, we conduct 500 simulations.



(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 1500$.

Figure G.4: In the Smartcards design, our method outperforms the two-step method in terms of coverage. The different plots vary n_o. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

G.2 Experimental, Observational, and Validation Samples

To compare our method with PPI methods, we create a new simulation design with a randomly selected validation sample, as discussed in Section 5.1. PPI methods correct the errors of generic machine learning outputs, so we focus on the forest cover design in which the remotely sensed variable R is the scalar output of the pretrained random forest.

Simulation design. We observe (D,R) when $\widetilde{S}=e$, (R,Y) when $\widetilde{S}=o$, and now (D,Y,R) when $\widetilde{S}=v$. The simulation design is as described in Section 6.1, except we now retain outcomes for a random subset of experimental units. Specifically, we fix the total number of experimental units at $n_e+n_v=1000$ and the observational sample size at $n_o=1000$. We vary the number of validation units: $n_v\in\{100,250,500\}$. These choices correspond to scenarios in which outcomes are randomly collected for 10%, 25%, or 50% of experimental units while relying on the remotely sensed outcome for the remainder. For the n_v randomly selected validation units, we have (D,Y,R); for the remaining $n_e=1000-n_v$ experimental units, we have only (D,R) as before.

Implementation details. We compare four methods.

Our method uses all three samples: $\widetilde{S}\in\{e,v,o\}$. Its estimand is τ. We estimate the efficient representation by combining logistic regressions of $\widehat{\Pr}(Y=1\mid\widetilde{S}\in\{o,v\},R)$, $\widehat{\Pr}(D=1\,|\,\widetilde{S}\in\{e,v\},R)$, and $\widehat{\Pr}(\widetilde{S}\in\{e,v\}\,|\,R)$. As before, we use cross-fitting with two folds. Our approach is unbiased and efficient.

One variation of PPI uses all three samples: $\widetilde{S}\in\{e,v,o\}$. We call it PPIO. Its estimand is

$$\widetilde{\tau}^{\text{PPIO}}=\mathbb{E}(R\,|\,D=1,\widetilde{S}=e)-\mathbb{E}(R\,|\,D=0,\widetilde{S}=e)+\mathbb{E}(Y-R\,|\,D=1,\widetilde{S}=v)-\mathbb{E}(Y-R\,|\,D=0,\widetilde{S}\in\{v,o\}),$$

which does not equal τ because the rectifier for the untreated arm uses $S=o$. Therefore, PPIO is biased.

Another variation uses only the $\widetilde{S}=e$ and $\widetilde{S}=v$ samples. We call it PPIV. Its estimand is

$$\widetilde{\tau}^{\text{PPIV}}=\mathbb{E}(R\,|\,D=1,\widetilde{S}=e)-\mathbb{E}(R\,|\,D=0,\widetilde{S}=e)+\mathbb{E}(Y-R\,|\,D=1,\widetilde{S}=v)-\mathbb{E}(Y-R\,|\,D=0,\widetilde{S}=v),$$

which equals τ because $\widetilde{S}=v$ is randomly selected in this simulation design. PPIV is unbiased but inefficient because it discards $\widetilde{S}=o$.

Finally, we consider a benchmark regression using only the $\widetilde{S}=v$ sample. Its estimand is τ. Using this small subset of units, it is a least squares regression of Y on D, ignoring R. It is unbiased but inefficient because it discards $\widetilde{S}=e$ and $\widetilde{S}=o$.

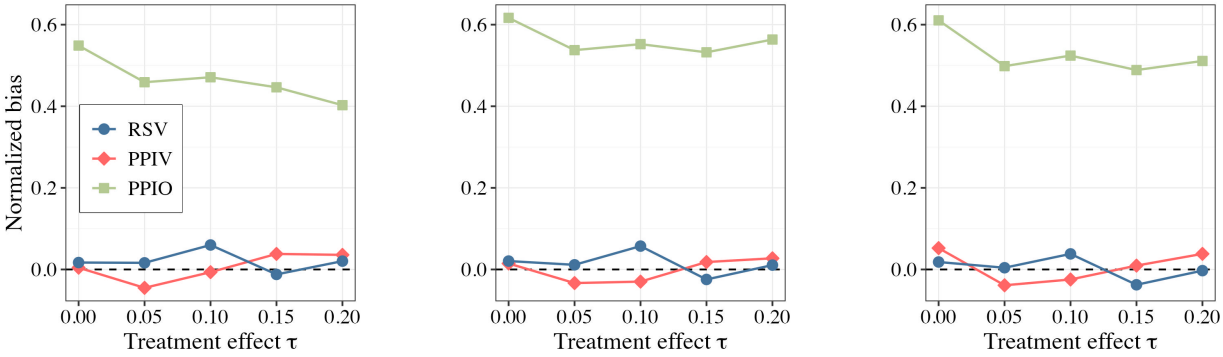
Additional results. Two findings emerge, pertaining to bias and variance.

Our method, PPIV, and the benchmark regression are unbiased, but PPIO is biased. Figure G.5 visualizes the normalized bias of the four methods in simulations. As before, the different panels vary the choice of α_o, by calibrating its value to different geographic bands around the experimental sample. The different plots vary the choice of n_v. Since $\alpha_e \neq \alpha_o$, PPIO displays substantial bias, confirming Proposition 2.

Focusing on the unbiased methods, we find that our method is more efficient than PPIV, which is more efficient than the benchmark regression. Intuitively, our method uses all three samples, PPIV uses two samples, and the benchmark regression uses one sample. Figure G.6 visualizes the percent reduction of standard errors. As before, the different panels vary the choice of α_o, by calibrating its value to different geographic bands around the experimental sample. The different plots vary the choice of n_v. Our method yields standard errors that can be 40% smaller than PPIV, and 60% smaller than the benchmark.

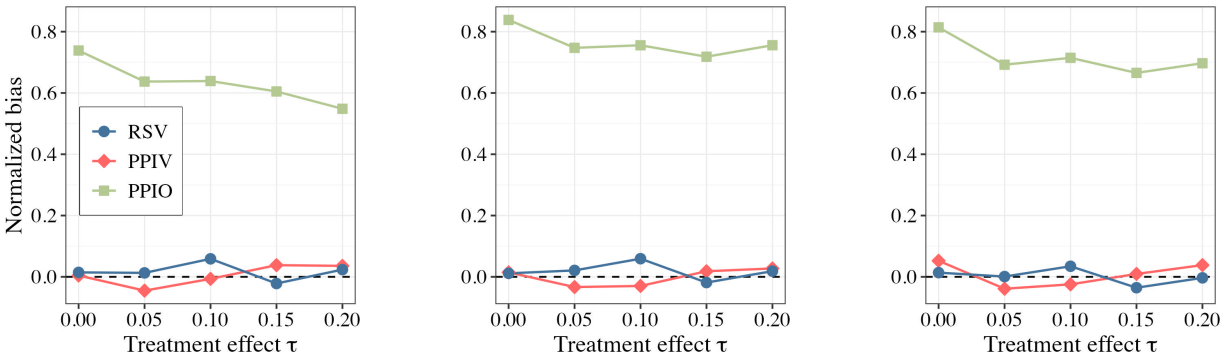
In summary, in the presence of a randomly selected validation sample $\widetilde{S}=v$, researchers may turn to our method to improve efficiency. By exploiting the post-outcome structure of the remotely sensed variable, our method efficiently incorporates information from the observational sample $\widetilde{S}=o$, while some existing methods do not.

Panel A: 0–2 Kilometer Geographic Band



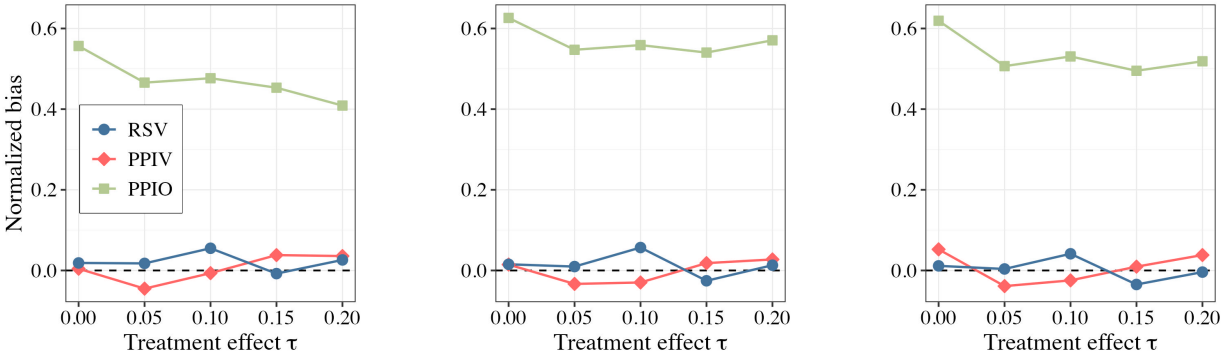
(a) $n_v = 100$. (b) $n_v = 250$. (c) $n_v = 500$.

Panel B: 0–5 Kilometer Geographic Band



(d) $n_v = 100$. (e) $n_v = 250$. (f) $n_v = 500$.

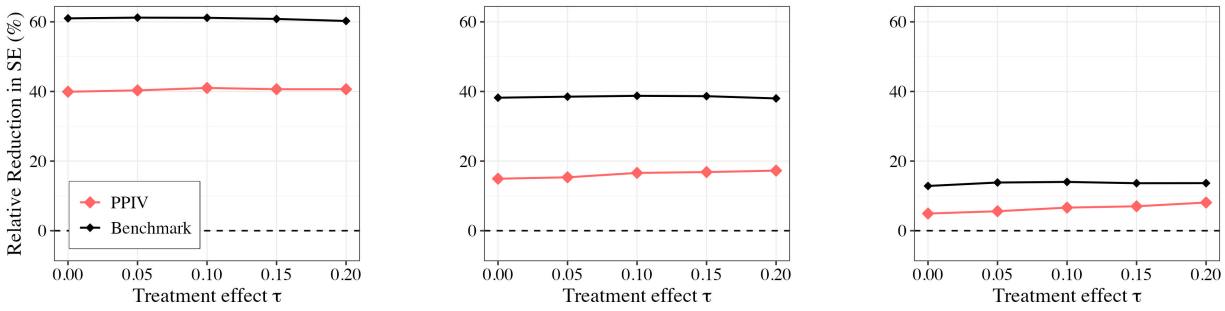
Panel C: 0–10 Kilometer Geographic Band



(g) $n_v = 100$. (h) $n_v = 250$. (i) $n_v = 500$.

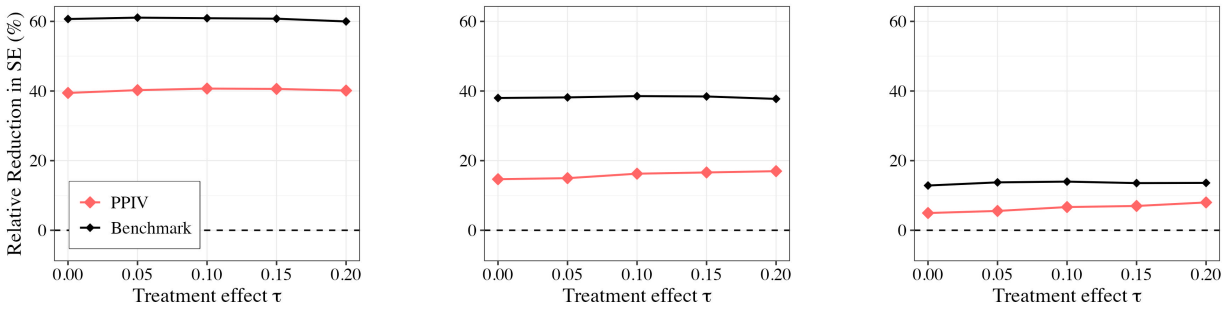
Figure G.5: In the forest cover design with three samples, we compare the normalized average bias of four methods: our proposal using $\widetilde{S} \in \{e,v,o\}$, PPIO using $\widetilde{S} \in \{e,v,o\}$, PPIV using $\widetilde{S} \in \{e,v\}$, and a benchmark regression using $\widetilde{S} = v$. PPIO is the only method that exhibits substantial bias. The different panels vary α_o. The different plots vary n_v. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

Panel A: 0–2 Kilometer Geographic Band



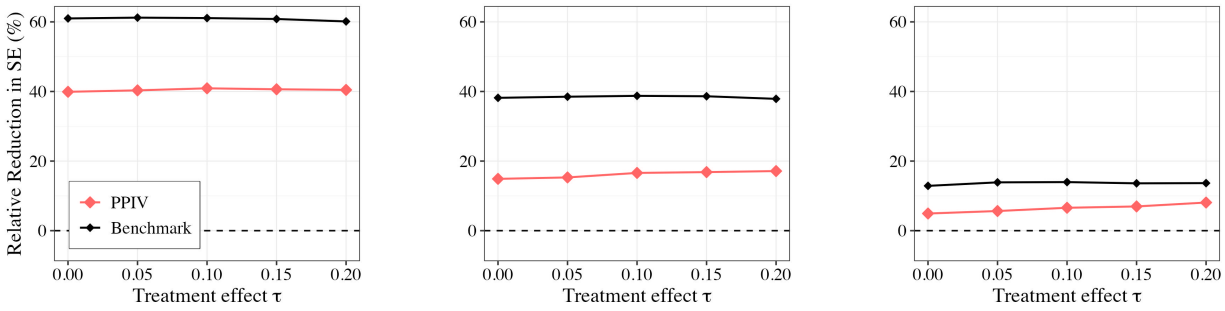
(a) $n_v = 100$. (b) $n_v = 250$. (c) $n_v = 500$.

Panel B: 0–5 Kilometer Geographic Band



(d) $n_v = 100$. (e) $n_v = 250$. (f) $n_v = 500$.

Panel C: 0–10 Kilometer Geographic Band



(g) $n_v = 100$. (h) $n_v = 250$. (i) $n_v = 500$.

Figure G.6: In the forest cover design with three samples, we compare the standard errors of the three unbiased methods: our proposal using $\widetilde{S} \in \{e,v,o\}$, PPIV using $\widetilde{S} \in \{e,v\}$, and a benchmark regression using $\widetilde{S} = v$. Specifically, we report the percent reduction in standard errors of our method relative to each unbiased alternative. Our method is the most efficient, since it uses all three samples. The different panels vary α_o. The different plots vary n_v. For each value of the synthetic treatment effect τ, we conduct 500 simulations.

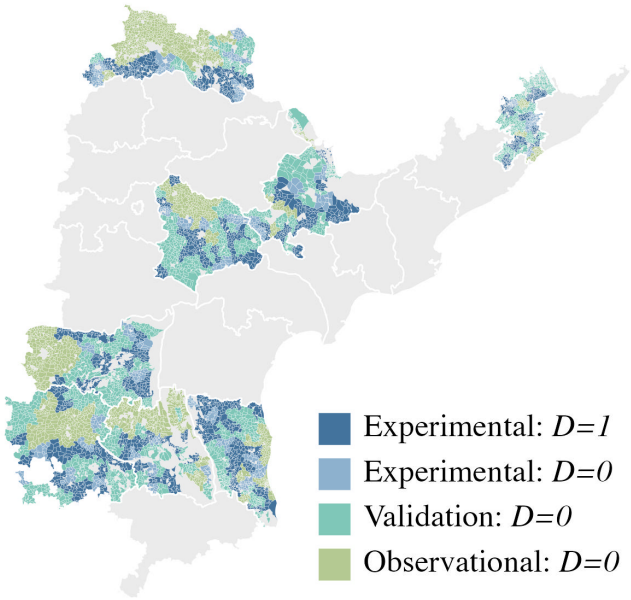


Figure H.1: We illustrate the experimental, validation, and observational samples in the Smartcards application (Muralidharan et al., 2023).

H Additional Empirical Results

H.1 Smartcards and Village-Level Poverty

In this section, we provide additional details and results for our re-analysis of the Smartcards experiment of Muralidharan et al. (2016, 2023) in Section 7.1. Appendix Figure H.1 illustrates our village classification.

We implement our method using probability random forests to estimate the nuisance functions that enter the representation. Specifically, we use the `R` package `ranger` with 1,000 trees `min.node.size = 20`, and `max.depth = 10`. Random forests satisfy stability conditions, which allow us to eliminate cross-fitting; the argument is a straightforward extension of Proposition 4, using stability in place of independence to handle the stochastic equicontinuity terms. See e.g. Chernozhukov et al. (2025, Theorem 2). Alternatively, we could use the limited complexity of random forests, along the lines of Chernozhukov et al. (2025, Theorem 3).

Appendix Table H.1 reports the point estimates and standard errors associated with Figure 5 in the main text. We also report the point estimates and standard errors for the differences between the unbiased benchmark and our method. The differences are statistically insignificant.

	Consumption	Low income	Middle income
RSV	-0.0525	-0.0261	-0.0450
	(0.0436)	(0.0220)	(0.0360)
Benchmark	-0.0746	-0.0235	-0.0397
	(0.0445)	(0.0148)	(0.0274)
Difference	-0.0222	0.0026	0.0053
	(0.0271)	(0.0142)	(0.0246)

Table H.1: Our method's point estimates approximately recovers the unbiased benchmark estimates. For each poverty outcome, we report point estimates and standard errors for the benchmark and for our method. We also report the difference between the benchmark and our method as well as its standard error. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

Plausibility of identifying assumptions. Appendix Figures H.2 and H.3 demonstrate that the stability and no-direct-effect assumptions are plausible for each poverty outcome.

First, stability requires that four conditional distributions of the remotely sensed variable should be the same across samples: $f_R(r \mid \widetilde{S} \in \{e,v\}, D=d, Y=y) = f_R(r \mid \widetilde{S} \in \{o,v\}, D=d, Y=y)$, for each (d,y) stratum.[28] Appendix Figure H.2 evaluates the conditions for $d=0$ and $y \in \{0,1\}$ by subsetting and visualizing densities. The conditional distributions almost coincide.

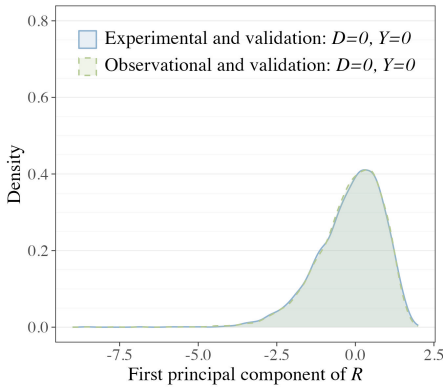
To complement the visual evidence, we conduct Kolmogorov-Smirnov tests for equality of the conditional distributions across samples. For each poverty outcome, we cannot reject the null of equal distributions at the 5% level.

Second, with incomplete observational cases, we require no direct effects (Assumption 3(ii)): the treatment should affect the remotely sensed variable only through the outcome. Using "oracle" data access, we can visually assess this condition; the densities $f_R(r \mid \widetilde{S} \in \{e,v\}, D=0, Y=y)$ and $f_R(r \mid \widetilde{S} \in \{e,v\}, D=1, Y=y)$ should coincide for $y \in \{0,1\}$. In Figure H.3, the conditional densities closely align for each poverty outcomes

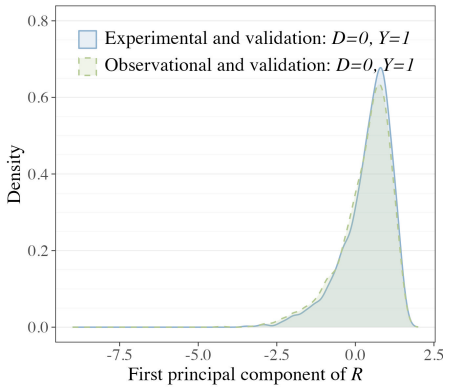
Again, we conduct Kolmogorov-Smirnov tests. For the "consumption" outcome, we can marginally reject the null of equal distributions at the 5% level for the $Y=0$ case ($p=0.048$)

[28] This condition follows from the proof of Lemma 1.

Panel A: Consumption

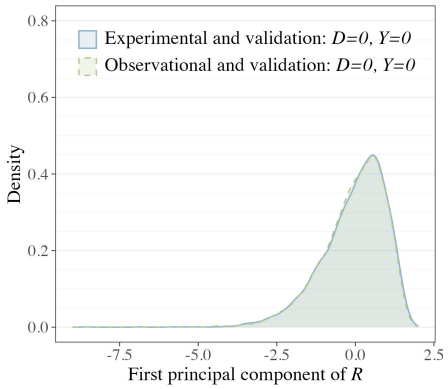


(a) Densities of $R \mid \widetilde{S}, D=0, Y=0$.

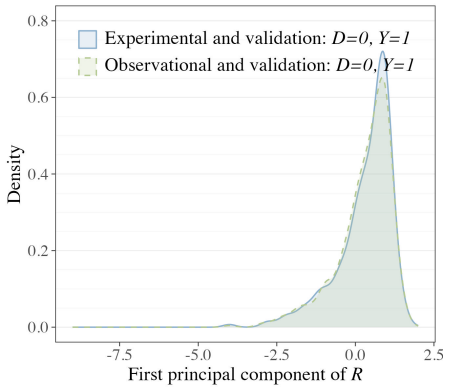


(b) Densities of $R \mid \widetilde{S}, D=0, Y=1$.

Panel B: Low Income

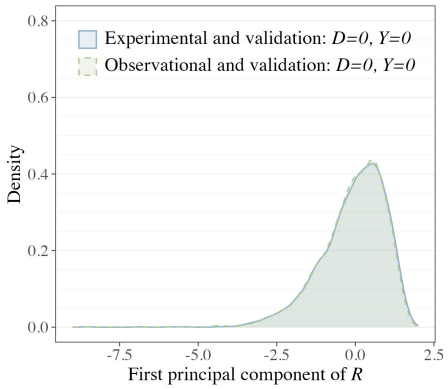


(c) Densities of $R \mid \widetilde{S}, D=0, Y=0$.

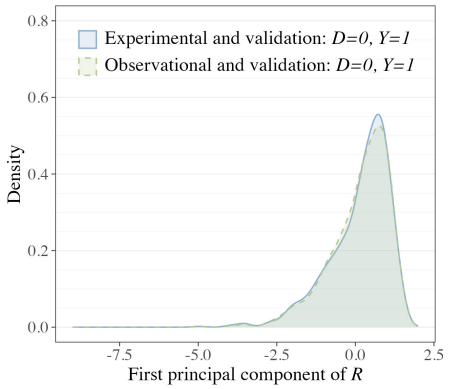


(d) Densities of $R \mid \widetilde{S}, D=0, Y=1$.

Panel C: Middle Income



(e) Densities of $R \mid \widetilde{S}, D=0, Y=0$.



(f) Densities of $R \mid \widetilde{S}, D=0, Y=1$.

Figure H.2: Our stability assumption (Assumption 2(i)) is plausible for each poverty outcome in the Smartcards experiment. We compare $f_R(R \mid \widetilde{S} \in \{e,v\}, D=d, Y=y)$ with $f_R(R \mid \widetilde{S} \in \{o,v\}, D=d, Y=y)$ for $d=0$ and $y \in \{0,1\}$. Because R is high-dimensional, we visualize the density of its standardized first principal component.

	Consumption	Low income	Middle income
Relevance, $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$	0.2349	0.5068	0.2145
	(0.0810)	(0.1715)	(0.0708)

Table H.2: Satellite images are relevant to the poverty outcomes in the presence of spillovers. For each poverty outcome, we report $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ and its standard error using our learned representation. The estimates correspond to Corollary C.4. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

and the $Y=1$ case ($p=0.033$). For the "low income" and "middle income" outcomes, we cannot reject the null of equality at the 5% level.

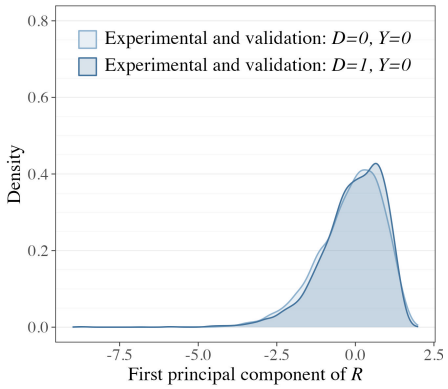
Comparing representations. Appendix Figure H.4 visualizes how our optimal representation $H^*(R)$ compares to the simple representation $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$. Even though the remotely sensed variable is high-dimensional, these representations are scalars. Therefore, we can visualize each village as a point in a plot with $H^*(R)$ on the vertical axis and $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$ on the horizontal axis. For the sake of visualization, we also plot the binscatter and the best fitting curve.

There are some notable differences. The optimal representation $H^*(R)$ often takes negative values, while the simple representation $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$ is within the unit interval $(0,1)$. It is optimal to extrapolate beyond observed villages. Using $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$ in place of $H^*(R)$ in Algorithm 1 would only interpolate among observed villages. The representations $H^*(R)$ and $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$ are generally correlated, but their relationship is nonlinear. As discussed in the main text, these differences matter for the precision of downstream inference.

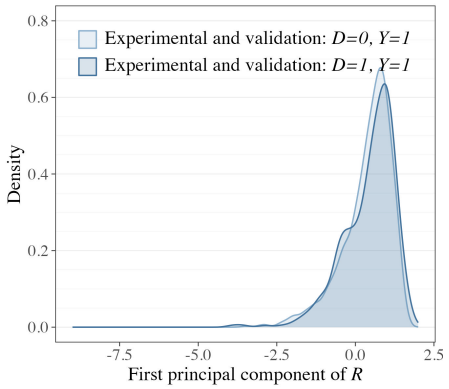
Spillovers. By assuming no spillovers across mandals, our method estimates the average global treatment effect from the full sample. Therefore, under the conditions of Corollary C.3, Figure 5 and Table 2 report average global treatment effect estimates and relevance in the presence of spillovers.

As a robustness check, we now repeat our empirical analysis after subsetting the villages in the manner described in Corollary C.4. The subsetting step eliminates about 27% of the total number villages. In particular, about 27% of villages in the full sample have at least one

Panel A: Consumption

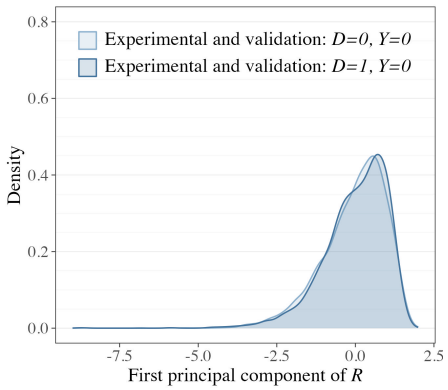


(a) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=0$.

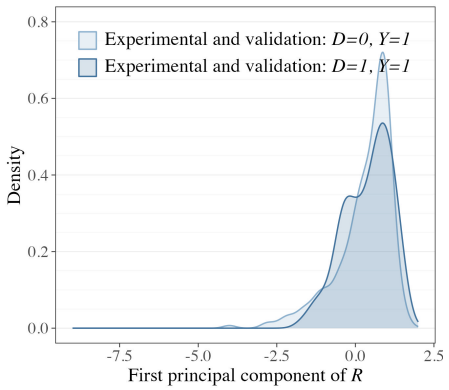


(b) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=1$.

Panel B: Low Income

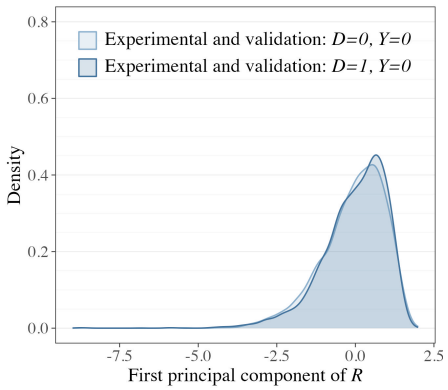


(c) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=0$.

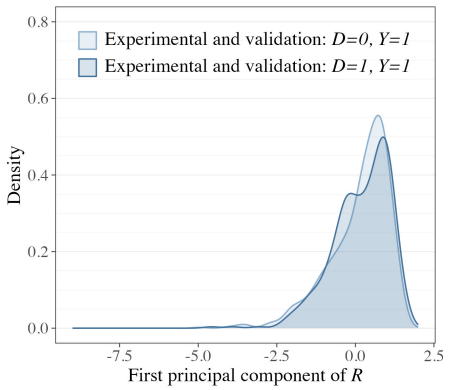


(d) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=1$.

Panel C: Middle Income

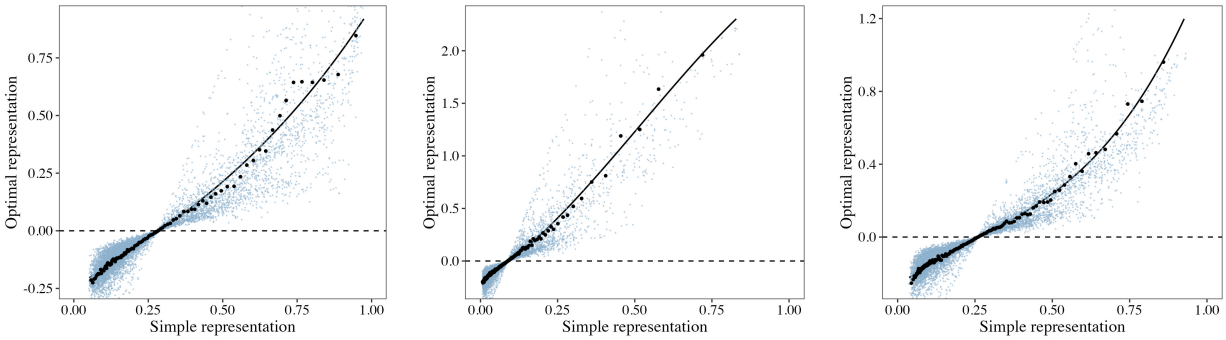


(e) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=0$.



(f) Densities of $R \mid \widetilde{S} \in \{e,v\}$, D, $Y=1$.

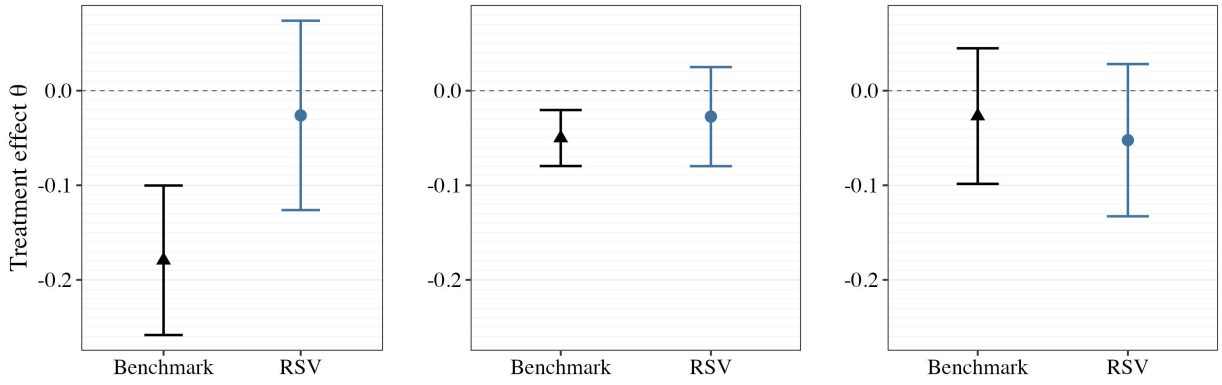
Figure H.3: No direct effects (Assumption 3(ii)) is plausible in the Smartcards experiment, for each poverty outcome. We compare $f_R(R \mid \widetilde{S} \in \{e,v\}, D=0, Y=y)$ with $f_R(R \mid \widetilde{S} \in \{e,v\}, D=1, Y=y)$ for $y \in \{0,1\}$. Because R is high-dimensional, we visualize the density of its standardized first principal component.



(a) Consumption. (b) Low Income. (c) Middle Income.

Figure H.4: We contrast optimal versus simple representations of the remotely sensed variable. The optimal representation is $H^*(R)$ from Section 4.1, combining three predictions. The simple representation is one prediction: $\Pr(Y=1 \mid R, \widetilde{S} \in \{o,v\})$.

village within a 20 km radius exposed to opposite treatment status. The distance between two villages is measured using their centroids, following Muralidharan et al. (2023).



(a) Consumption. (b) Low income. (c) Middle income.

Figure H.5: Our method's estimates approximately recover the unbiased benchmark estimates in the presence of spillovers. For each poverty outcome, we report 90% confidence intervals for the benchmark and for our method. The estimates correspond to Corollary C.4, i.e. using the subsetted sample. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

Under the conditions of Corollary C.4, Figure H.5 and Table H.2 report global average treatment effect estimates and relevance in the presence of spillovers. We find qualitatively similar results as before. Here, the benchmark uses the same subset of villages as our method.

Time variation. We discuss time variation in detail in Appendix C.4. Here, we summarize the consequences for the interpretation of results presented in Section 7.1.

Under the conditions of Theorem C.4, Figure 5 reports the effect of 2010 Smartcards adoption on 2013 poverty, for treated, untreated, and buffer mandals. Our results apply when D is an early adoption treatment deployed in 2010, when Y is a poverty outcome collected in 2013, and when R is a remotely sensed variable collected separately in, say, 2019.

With incomplete cases, there is an implicit Markov restriction that researchers must assess. Concretely, 2010 Smartcards may affect 2013 poverty; 2013 poverty may affect 2019 poverty; and 2013 and 2019 poverty may affect 2019 satellite images. However, in this example, 2010 Smartcards cannot affect 2019 poverty via any channel besides 2013 poverty. Figure H.3 provides empirical evidence supporting this assumption.

H.2 Payments for Ecosystem Services and Crop Burning

We present additional results for the empirical application in Section 7.2, based on Jack et al. (2025). Specifically, Table H.3 presents additional variations of the two-step method. In Table 4, we used the authors' "maximum accuracy" classifier for whether a field has not been burned as the predicted outcome. Now, we also use the authors' "balanced accuracy" classifier as the predicted outcome. Finally, we use the continuous random forest prediction as the predicted outcome. Both variations of the two-step method appears to have economically meaningful attenuation bias (Proposition 1).

	Maximum accuracy	Balanced accuracy	Random forest prediction
Common practice	0.075	0.067	0.030
	(0.038)	(0.046)	(0.017)

Table H.3: Variations of the common practice exhibit attenuation bias. The different variations use different predicted outcomes within the two-step method: the authors' "maximum accuracy" classifier, their "balanced accuracy" classifier, or continuous random forest predictions. We separately regress each predicted outcome on the treatment. Bootstrap standard errors in parentheses, based on 5000 replications, are clustered at the village level.