# Program Evaluation with Remotely Sensed Outcomes

Ashesh Rambachan
MIT

Rahul Singh
Harvard

Davide Viviano*
Harvard

First draft: November 2024
This draft: March 2026

**Abstract**

We study causal inference in experiments and quasi-experiments, where the economic outcome is imperfectly measured by a remotely sensed variable (RSV). The RSV is low-cost and scalable, such as satellite imagery or mobile phone activity, and predictive of the economic outcome in observational data. In our model, the RSV is post-outcome: variation in the economic outcome causes variation in the RSV. For example, changes in environmental quality cause changes in satellite imagery, not vice versa. Under this assumption, we propose a formula to nonparametrically identify the causal parameter by combining experimental and observational data. We develop a method for $n^{-1/2}$ inference, robust to misspecification, without restricting the algorithms used to process remotely sensed variables.

**Keywords:** Causal inference, data fusion, experiments, machine learning.

*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, Seema Jayachandran, Namrata Kala, Sylvia Klosin, Ben Olken, Jonathan Roth, Jesse Shapiro, Matthieu Stigler, and Alex Tetenov, as well as audiences at Bocconi, CEMFI, CREST/Sciences Po, Harvard/MIT, the NBER Summer Institute, the Online Causal Inference Seminar, Opportunity Insights, 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. As a consequence, researchers increasingly rely on remotely sensed variables (Appendix Figure A.1). Examples include night luminosities 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 cheaply available, and predictive of the outcomes of interest in observational data, but still imperfect.[2] Our research question is 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 that is affected by the outcome of interest. The experimental sample contains a randomly treatment assignment and the remotely sensed variable, but the economic outcome is missing (entirely or mostly) due to logistical or cost constraints. The observational sample links the outcome to the remotely sensed variable, but it is from a different non-experimental context. This implies that the treatment variable is missing or confounded, and this population may exhibit different treatment effects from the target experimental population. What the two samples have in common 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 a government program to reduce crop burning, we may have an experimental sample containing randomized contracts and satellite images of fields, as well as an observational sample linking 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 that the conditional distribution of the remotely sensed variable, given the outcome and treatment, is *stable* across samples. This assumption is plausible when the same measurement technology is used across samples, e.g. the same satellite generates

[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 McCord, 2022; Fowlie et al., 2019; Corral et al., 2025; Josephson et al., 2026).

the images in both datasets. Therefore, we study a setting where the treatment mechanism and outcome mechanism vary across samples, but the *sensing* mechanism is the same.

Our primary contribution is a novel formula to identify the treatment effect, even when no experimental outcomes are collected, by leveraging stability of the sensing mechanism. 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 providing a set of conditional moment restrictions, we also formalize formal diagnostic tests for the key 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, similar to testing of overidentifying restrictions.

As a secondary contribution, we develop a method for $n^{-1/2}$ inference on the treatment effect for arbitrarly complex machine learning methods. Building on a long-standing literature on optimal instruments (Chamberlain, 1987; Newey, 1993), we show how to summarize (high-dimensional) remote sensed variables into low-dimensional representation that should depends on predicted outcomes, treatments, and sample indicators for efficient inference. 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 converge slowly), allowing for arbitrary machine-learning estimator. This property comes from the structure of the moment conditions we obtain, which are linear in the choice of the remote sensed variable. From a practical perspective, this result allows for complex neural network estimators in compressing high-dimensional satellite images.

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.

For the setting we study (Theorems 1-2), we prove that some existing methods are biased.

In 50% of papers in general interest economics journals from 2015–2024 that use remotely sensed outcomes, many economists use 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] This method suffers from attenuation bias (Proposition 1), even if samples are randomly selected (Proposition C.2); it systematically underreports the effectiveness of government interventions. Another important method is called prediction powered inference (Angelopoulos et al., 2023), which adds bias correction terms. This method can be arbitrarily 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). Such a distribution shift is plausible when the two samples come from different populations, time periods, or regions.

We compare our method to some existing methods 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). We show that our method is approximately unbiased with nominal coverage. This comes from the plausibility of our assumptions in these settings, shown through a series of diagnostic checks. By contrast, the two step method and prediction powered inference are generally biased.[4]

## 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 some formal comparisons in Section 3.2.

In the surrogacy model (Prentice, 1989; Athey et al., 2025; Kallus and Mao, 2024), the auxiliary variable is called a surrogate, and it is pre-outcome: it is a "short term" measurement that mediates between the treatment and "long term" outcome of interest. 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 to attenuation bias (Proposition 1). The negative control literature extends

[3]We survey AEA journals, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*. The other 50% use a similar logic, without an explicit formula.

[4]Prediction powered inference is unbiased if (i) a randomly selected validation sample is available, and (ii) the observational sample is discarded. However, discarding the observational sample is inefficient (Remark 4).

the surrogacy model to address unobserved confounding (Ghassami et al., 2022; Imbens et al., 2024). In these extensions, misusing a remotely sensed variable incurs bias.

Several recent papers propose methods for inference with remotely sensed outcomes that require correct specification of a model relating the remotely sensed variable to the outcome. Proctor et al. (2023) propose a multiple imputation method that models the outcome as a function of the remotely sensed variable across samples, akin to the surrogacy method. Alix-Garcia and Millimet (2023); Torchiana et al. (2025) model misclassification processes for binary and panel outcomes. By contrast, our method does not require correct specification of any model relating the remotely sensed variable to the outcome. This distinction is important when the remotely sensed variable is unstructured, e.g. a satellite image.

One major difference between our method and prediction powered inference (Angelopoulos et al., 2023; Kluger et al., 2025) is the data requirement. Prediction powered inference requires joint observations of the treatment, outcome, and remotely sensed variable, in both treatment arms; it is 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) 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). See Remark 4 for comparisons in other data environments.

Another major difference between our method and prediction powered inference is the identifying assumption. Ji et al. (2025) show that prediction powered inference uses machine learning predictions as surrogates. We show that it implicitly requires stability of the outcome mechanism (Proposition 2). By contrast, our method requires stability of the sensing mechanism. Our method allows the experimental and observational samples to have different treatment effects, which is important when the samples are from different time periods or regions.

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 distinctions leads to a distinct identifying formula.

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 ($S = o$).

Within the experimental sample ($S = e$), we observe pre-treatment covariates $X \in \mathcal{X}$ and a binary treatment $D \in \{0,1\}$. However, the outcome $Y \in \mathcal{Y}$ is missing entirely. In its place, the researcher has access to a remotely sensed variable $R \in \mathcal{R}$. (In Section 5.1, we allow for the outcome to be only partially missing in the experimental sample.) We are agnostic about 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 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.[5]

**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. We refer to the auxiliary dataset as the observational sample ($S = o$). For these units, we observe baseline covariates $X$, the outcome $Y$, and the 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

[5]This definition of potential outcomes rules out spillovers across units. We extend our framework to allow for spillovers in Section 5.3.

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. As we discuss below, 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. For units in the experimental sample $(S=e)$, we observe $(X,D,R)$. For units in the observational sample $(S=o)$, we observe $(X,D,Y,R)$ if there are complete cases or we observe $(X,Y,R)$ 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 in the experiment is costly, requiring surveyors to monitor tree cover (Jayachandran et al., 2017) or to record crop management practices across rural areas between the harvest and the next tilling (when crop-burning typically occurs) (Jack et al., 2025). By contrast, satellite imagery $R$ is cheaply available at granular resolution and can accurately predict 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, which links the remotely sensed variable $R$ to the directly measured 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). Consequently, these auxiliary samples are not well-suited to serve as the primary basis for treatment effect estimation in the experiments. In both cases, however, they constitute possible observational samples in our framework. ▲

**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 $Y$. Measuring poverty in the experiment requires conducting detailed household surveys to collect consumption data, 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 now formalize this setting by presenting three assumptions under which the researcher can combine the experimental and observational samples 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. 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 empirical applications involving remotely sensed variables, such as Examples 1 and 2, this assumption is satisfied by design since 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: $\Pr(S=e|X)$ 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 measurement process (i.e., how the outcome generates the remotely sensed variable) is the same in both samples. In Sections 6-7, we provide evidence illustrating that Assumption 2(i) is plausible across three empirical applications in environmental and development economics.

Importantly, this condition does not require stability of the underlying treatment effects, which may differ across samples. In the language of causal inference, the treatment mechanism $\Pr(D=1|S,X)$ and the outcome mechanism $f_{Y(d)}(y|S,X,D)$ may depend on the sample $S$. Only the sensing mechanism must be stable: $f_R(r|S=e,X,D,Y)=f_R(r|S=o,X,D,Y)$. When the sensing mechanism is a shared technology—e.g. satellite images from the same satellite—stability becomes more plausible.

The remaining aspects 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 RSV distribution does not degenerate for any stratum. Assumption 2(iv) requires that we observe some data from both the experimental and observational samples.

**Example 1 (continued):** Returning to Example 1, satellite images are informative about 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 digital traces 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 same effects, outcome mechanisms, nor 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—a further 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: $D \perp\!\!\!\perp R \mid S=o,X,Y$.

Assumption 3 requires only one of two conditions to hold.

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

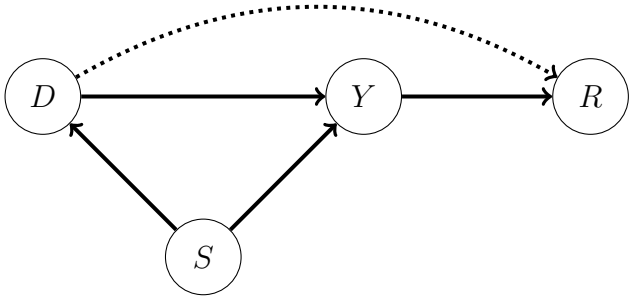



Figure 1: Causal graph for remotely sensed variables under Assumptions 3(i) versus 3(ii). Complete cases allow the dotted line. Assumption 2 rules out the line from $S$ to $R$.

**Incomplete observational cases:** If Assumption 3(i) fails, we have incomplete cases: no observations of $(X,D,Y,R)$ where $D$ varies. Without joint observations of the outcome and treatment, a further 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, Assumption 3(ii) 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. It becomes more plausible when $Y$ is a vector of outcomes, which may capture all mechanisms through which the treatment affects the remotely sensed variable.

While Assumption 3(ii) is a condition placed for the observational sample, combining it with Assumption 2(i) implies no direct effects for the experimental sample. Therefore, we will subsequently refer to no direct effects for both samples: $D \perp\!\!\!\perp R | X,Y$.

Together, Assumption 2(i) and Assumption 3(ii) imply that $(S,D) \perp\!\!\!\perp R \,|\, X,Y$. Importantly, these assumptions are jointly testable, as we discuss in Section 4.

How sensitive are our results to violations of Assumptions 3(i) and 3(ii)? In other words, what if there are incomplete observational cases yet direct effects? In Section 5.4, we demonstrate that the causal parameter remains partially identified.

To summarize, Figure 1 illustrates our identifying assumptions as a causal graph. The treatment affects the outcome, which in turn affects the remotely sensed variable. In other words, 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. Appendix Table A.1 summarizes the implications of our main assumptions.

# 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 then compare our identification strategy to alternative approaches, clarifying how our assumptions and data requirements differ. In our setting, several common methods are infeasible, biased, or inefficient.

## 3.1 Main Result

We derive a novel formula for data combination, without restricting the distribution nor dimension of the remotely sensed variable.

In this section, we impose that the outcome is discrete with finite support $\mathcal{Y}=\{y_1,...,y_K\}$ 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 binary (e.g., crop burning versus no crop burning). Section 5.5 extends our results to continuous outcomes.

Throughout the paper, 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 the settings with incomplete or complete observational cases, where either Assumption 3(ii) or 3(i) holds, respectively.

### 3.1.1 Incomplete Observational Cases: No Direct Effects

To begin, 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 outcomes we wish to identify.

**Lemma 1** (Identification as conditional densities)**.** Suppose Assumptions 1, 2, and 3(ii) hold. Denote by $\delta_d^e(X,R) := f_R(R \mid S=e,X,D=d)$ and $\delta_k^o(X,R) := f_R(R \mid S=o,X,Y=y_k)$ the conditional densities of the remotely sensed variable in the experimental and observational samples, respectively. Then, for any $d \in \{0,1\}$,

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

*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 would 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 $\theta(X) := (\mu_1(1,X) - \mu_1(0,X), \ldots, \mu_{K-1}(1,X) - \mu_{K-1}(0,X))^\top$ be the conditional average treatment effect on the probability of each discrete outcome value. Identification of $\theta(X)$ implies identification of the conditional average effect $\tau(X) := \mathbb{E}\{Y(1) - Y(0) \mid S = e, X\}$.

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

$$\mathbb{E}\{\Delta^e(X) - \Delta^o(X)^\top \theta(X) \mid X, R\} = 0 \text{ almost surely, where,}$$
$$\Delta^e(X) := \frac{1\{D=1, S=e\}}{\Pr(D=1, S=e \mid X)} - \frac{1\{D=0, S=e\}}{\Pr(D=0, S=e \mid X)},$$
$$\Delta^o_k(X) := \frac{1\{Y=y_k, S=o\}}{\Pr(Y=y_k, S=o \mid X)} - \frac{1\{Y=y_K, S=o\}}{\Pr(Y=y_K, S=o \mid X)}.$$

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.[6]

*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 $\mathcal{X} = \varnothing$. The treatment and outcome variation simplifies to scalars $\Delta^e$ and $\Delta^o$. Theorem 1 simplifies to a simple ratio:

$$\tau = \frac{\mathbb{E}[\Delta^e \mid R]}{\mathbb{E}[\Delta^o \mid R]},$$

where

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

This ratio appears to be a new formula for data combination, where the numerator and denominator are from different samples. Intuitively, 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 $Y$. ▲

[6] Following the notation of Ai and Chen (2003), we denote the exogeneous regressor by $X$.

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. In the standard fashion, 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 $\tau$ are identified.

**Example 3 (continued):** Return to the simple setting of binary $Y$ and no $X$ for intuition. Consider any measurable function $H:\mathcal{R}\to\mathbb{R}$. Corollary 2 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 demonstrates 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. Intuitively, 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 are sufficient 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; see Remark 1.

**Remark 1** (Overidentifying restrictions and a specification test)**.** Our framework generates testable overidentifying restrictions. In the setting with incomplete observational cases, Corollary 1 shows that any predictive 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 test of overidentifying restrictions using a $J$-statistic (e.g., Sargan, 1958; Hansen, 1982). An analogous specification test can be constructed for the setting with complete observational cases 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 $\mathbb{E}[H(R)\Delta^e]/\mathbb{E}[H(R)\Delta^o]$ becomes 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 different from zero, using existing frameworks which test for weak instruments.

### 3.1.2 Complete Observational Cases: Allowing for Direct Effects

Until this point, we have focused on the setting of Assumption 3(ii), which allows for incomplete observational cases but rules out direct effects of the treatment on the remotely sensed variable. Our identification result also holds for the setting of Assumption 3(i), which requires complete observational cases yet allows for direct effects. As mentioned earlier, 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 $\mu(d,X) := (\mu_1(d,X),\ldots\mu_{K-1}(d,X))^\top$.

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

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

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

*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,

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

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

## 3.2 Comparison with Alternative Approaches

Finally, 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 directly extends to the more general setting.

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

For expositional convenience, suppose Assumption 3(ii) holds: 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, and (ii) apply the predictor $m(R)$ to the experimental sample and calculate treatment effects on the predicted outcomes. This method appears in about 50% of our surveyed papers in general interest economics journals from 2015-2024 that use remotely sensed outcomes. While intuitive, this method generally incurs bias for the causal parameter $\tau$.

This two-step method implicitly targets the estimand

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

Within this expression, the first step estimates the function $m(R)$. The second step evaluates and averages $m(R)$ over the treated and untreated subgroups in the experimental sample.

As a starting point, if the remotely sensed variable fails to predict the outcome, i.e., 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. This 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.

More generally, even if the remotely sensed variable is predictive of the outcome, the implicit target $\widetilde{\tau}$ is biased away from $\tau$, i.e. the ATE in the experimental sample.

**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. as long as $\mathrm{Var}(Y \mid R, S=o) > 0$ almost surely.

*Proof.* See Appendix B.1.4 for the proof, as well as the generalization for discrete $Y$. □

Proposition 1(i) characterizes the implicit estimand of a 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, existing 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. In other words, empirical papers that use this method systematically under-report the impacts of environmental and anti-poverty interventions. The attenuation factor $\kappa$ reflects the strength of the correlation between the remotely sensed variable and the outcome.[7]

We generalize Proposition 1 for several empirically relevant settings. Appendix C.1.1 provides an analogous result 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 in causal inference. The implicit target $\widetilde{\tau}$ of the two-step method is the well-known surrogate formula (Prentice, 1989; Athey et al., 2025), so it recovers the causal parameter $\tau$ under the standard surrogacy assumptions $(D,S) \perp\!\!\!\perp Y \mid R,X$. These surrogacy assumptions state that $R$ fully mediates the effect of the treatment on the outcome, i.e. that the surrogate variable $R$ 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 which lead to the estimand $\tilde{\tau}$, have been extended to include negative controls (e.g., Ghassami et al., 2022; Imbens et al., 2024). The surrogacy method and its negative control generalizations would all incur the bias formalized in Proposition 1 when applied to post-outcome variables.

[7] A related result in Pelletier et al. (2026) shows 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. Proposition 1 fully characterizes the estimand and bias of common practice when the remotely sensed variable is post-outcome.

Similar to our method, the single negative control method (Park et al., 2024) uses one auxiliary variable. However, 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.

### 3.2.2 Prediction-Powered Inference

Next, we compare our approach with prediction-powered inference (PPI) (Angelopoulos et al., 2023; Pelletier et al., 2026). The PPI method begins with a pretrained machine learning predictor of a partially missing outcome variable $Y$. In our notation, denote the arbitrary predictor for the treated units by $m_1(R)$ and the arbitrary predictor for the untreated units by $m_0(R)$.

Similar to the intuitive method described above, the PPI method computes a difference of means using predicted outcomes for incomplete observations of $(D,R)$. The PPI method also adds bias correction terms, called "rectifiers," using complete observations of $(D,Y,R)$. Within Example 3, the estimand of PPI is therefore

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

First, suppose that we have incomplete observational cases. By definition, we do not have access to 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. Therefore, at least one PPI rectifier (i.e., $\mathbb{E}[Y-m_d(R) \mid D=d,S=o]$) cannot be computed, and so PPI is infeasible.

Second, suppose that we have complete observational cases. We now demonstrate that 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.[8] There may be bias even in the absence of direct effects (i.e., even if both Assumptions 3(i) and 3(ii) hold.)

**Proposition 2** (Bias of PPI)**.** 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 $\tilde{\tau}^o := \mathbb{E}[Y \mid D=1,S=o] - \mathbb{E}[Y \mid D=0,S=o]$ denote the difference in means in the observational sample. Then

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

1. $\tilde{\tau}^{\text{PPI}} = \tau + (1-\kappa_1)(\tilde{\tau}^o - \tau) + (\kappa_1 - \kappa_0)\delta$ 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 $\tilde{\tau}^o$ and $\tau$ (with possibly $\tilde{\tau}^o \neq \tau$), there exists a data-generating process with a relevant $R$ (i.e., $R \not\perp\!\!\!\perp Y | D, S$) such that $\tilde{\tau}^{\text{PPI}} = \tilde{\tau}^o$.

*Proof.* See Appendix B.1.5. □

Proposition 2(i) characterizes the implicit estimand of PPI: a mixture of the desired estimand $\tau$, the difference in means in the observational sample $\tilde{\tau}^o$, and the difference in baselines across samples $\delta$, with weights that depend on the predictors $m_1(R)$ and $m_0(R)$. Therefore, 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, PPI is generally biased.

We conclude that PPI implicitly relies on stability of the treatment effect and outcome mechanism across samples. By contrast, our method only relies on stability of the sensing mechanism. This distinction is salient in applications such as Example 1, where the observational sample is collected on different units and at different times than the experimental endline. 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 and/or the treated–untreated contrasts differ across samples ($\delta \neq 0$ and/or $\tilde{\tau}^o \neq \tau$), which are precisely the conditions under which Proposition 2(i) shows that PPI may be biased.

Proposition 2(ii) further shows that PPI can have arbitrary bias: it may be targeting $\tilde{\tau}^o > \tau$, $\tilde{\tau}^o < \tau$, or something else entirely. In other words, empirical papers that use this method may be over- or under-reporting the effectiveness of environmental and anti-poverty interventions.

### 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, 2025) establish identification by assuming stability of the distribution of the latent variable given the noisy measurement. In our notation, they require outcome stability: $f_Y(y \mid R, D, X, S=e) = f_Y(y \mid R, D, X, S=o)$.

This paper studies a different regime, where the noisy measurement is plausibly post-outcome: the remotely sensed variable is caused by the outcome of interest. We instead assume that the sensing mechanism is stable: $f_R(r \mid Y,D,X,S=e) = f_R(r \mid Y,D,X,S=o)$. 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.

For intuition, the structural equation $Y = R+U$ is a special case of latent variable stability and of Berkson measurement error. Meanwhile, the structural equation $R = Y+U$ is a special case of our model and of classical measurement error. This is the essential difference.

# 4 Estimation and Inference

As our secondary contribution, we show how our identification result guides the choice of representation for the remotely sensed variable, achieving precise inference on the treatment effect.

We also prove that our confidence intervals are highly robust. We derive valid $n^{-1/2}$ inference without rate conditions and without 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 is similar, using the identification results in Section 3.1.2; see Appendix C.6.2 for details. We generalize our estimation algorithm to include covariates in Section 5.6.

## 4.1 Choice of the Representation

For intuition, consider the scenario with a binary outcome (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; which representation is a reasonable choice in practice?

To answer this question, we may view $\tau$ 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 representation that achieves efficiency

within a class of models satisfying (1) is $H^*(R) = \frac{\mathbb{E}(\Delta^o|R)}{\sigma^2(\theta,R)}$, where $\sigma^2(\theta,R) := \mathbb{E}\{(\Delta^e - \Delta^o\theta)^2 \,|\, R\}$. 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 representation that achieves efficiency within a class of models satisfying Theorem 1 is

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

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

This formula reveals that improving the efficiency of downstream causal estimates requires three types of predictions: (i) predictions of each outcome category given from remotely sensed variable, appearing in the numerator $\mathbb{E}(\Delta^o \,|\, R)$; (ii) a 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, to improve efficiency of downstream causal estimates using remotely sensed variables, we should not only predict the outcome $Y$ using observational data, but also the treatment $D$ using experimental data, 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, using a family of representations.

## 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. For clarity, Algorithm 1 provides an overview of our estimation algorithm using sample splitting with TRAIN and TEST folds, 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, and Algorithm E.2 describes the algorithm with the complete observational cases.

Sample splitting may be eliminated under complexity restrictions that tolerate simple machine learning procedures. See e.g., Chernozhukov et al. (2020) 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 $\widehat{\Pr}(Y = y_k | S = o, R)$, sacrificing some precision to reduce computation.

**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 $\mathbb{E}_{\text{TRAIN}}(\cdot) = \frac{\sum_{i\in\text{TRAIN}}(\cdot)}{|\{i\in\text{TRAIN}\}|}$ and $\mathbb{E}_{\text{TEST}}(\cdot) = \frac{\sum_{i\in\text{TEST}}(\cdot)}{|\{i\in\text{TEST}\}|}$.

**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 $\widehat{H}(R) \in \mathbb{R}^{K-1}$ intended to approximate the efficient representation $H^*(R)$ (see Algorithm E.1 for an 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 $\widehat{\tau} = \sum_{j=1}^{K-1} y_j \widehat{\theta}_j + y_K\left(1 - \sum_{j=1}^{K-1}\widehat{\theta}_j\right)$.

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

8: **return** $\widehat{\tau}$ and its confidence interval.

## 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. To accommodate all forms of remotely sensed variables simultaneously, we analyze our estimator under weak regularity conditions.

For example, the prediction may be constructed from a complex machine learning algorithm, e.g. a deep convolutional neural network applied to satellite images or even a pre-trained model. For this realistic setting, 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. We require only that the predictions, and hence the representation estimator, have some probability limit; they may be misspecified.

**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 possible $\widetilde{H}(R) \neq H^*(R)$. This limit is correlated with outcome variation: $\mathbb{E}\{\widetilde{H}(R)(\Delta_o)^\top\}$ is non-singular, 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 $\tau$ 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 so Assumption 4 requires no complexity restriction (Chernozhukov et al., 2018, 2023) and no rate requirement (Chamberlain, 1987; Newey, 1993).

Concretely, within Algorithm 1, all of the predictors $\widehat{\Pr}(Y=y_k \mid S=o,R)$, $\widehat{\Pr}(D=1 \mid S=e,R)$, and $\widehat{\Pr}(S=e \mid R)$ 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 $y_k \in \{1,...,y_K\}$ as the marginal probabilities.

**Proposition 3** (Inference with known marginal probabilities)**.** Suppose Theorem 1's conditions and Assumption 4 hold. Suppose the marginal probabilities are known and bounded away from zero. Then for $\lambda=(y_1-y_K,...,y_{K-1}-y_K)^\top$ 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 $\widehat{\tau}$ is semiparametrically efficient for $\tau:=\lambda^\top\theta$ satisfying the conditional moment restriction $\mathbb{E}\{\Delta^e-(\Delta^{o\top})\theta \mid R\}=0$, with known marginal probabilities.

*Proof.* See Appendix B.2.1. □

**Proposition 4** (Inference with estimated marginal probabilities)**.** Suppose Theorem 1's conditions and Assumption 4 hold. If the marginal probabilities and their estimators are bounded away from zero, then $\sqrt{n}(\widehat{\tau}-\tau) \rightsquigarrow \mathcal{N}(0,\lambda^\top AVA^\top\lambda)$ for $V$ defined in Appendix Lemma B.6.

*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 that we must take into account.

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 D.6.2, and involves less computation.

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

# 5 Extensions and Generalizations

Our framework for identification and estimation naturally 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) partial identification when our main assumptions are relaxed; (v) continuous outcomes; (vi) low dimensional covariates. In this section, we provide a brief overview of each extension, while deferring formal details to the Appendix.

## 5.1 A Validation Sample

In empirical applications, researchers may also observe outcomes for a (possibly small) subset of experimental units, which we call the validation sample. This third sample is easily incorporated 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, all our identification and estimation results go through 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: $\widetilde{S} \perp\!\!\!\perp R | X,D,Y$. As before, only the sensing mechanism must be invariant.

Remark 4 below explains how a particular variation of PPI becomes unbiased by appropriately leveraging the validation sample and placing an extra assumption.[9] Extensive simulations demonstrate that, in the presence of a validation sample, our method is more efficient.

## 5.2 A Quasi-Experimental Sample

In many empirical applications with remotely sensed variables, researchers do not randomize treatment themselves but instead exploit quasi-experimental variation (e.g., Burgess et al.,

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

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 parameters for quasi-experiments: the local average treatment effect (LATE) using instruments, or the average treatment effect on the treated (ATT) using difference-in-differences. Overall, the technique in our main identification result is easily integrated into classic identification arguments for quasi-experiments.

## 5.3 Some Spillovers

When evaluating conservation and anti-poverty interventions, empirical researchers often discuss spillovers: how treatment of one unit may affect the potential outcome of another unit. In Appendix C.3, we extend our framework to allow for some simple types of spillovers.

First, we expand the potential outcome notation to reflect interdependence. In this richer notation, 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, by limiting spillovers to be 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 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 small departures. Appendix C.4 develops a partial identification framework for this purpose, summarized 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 the researcher is willing to bound the magnitude of the direct effect, and is able to calculate the correlation between the outcome variation and the representation $H(X,R)$. We prove that the identified set is an interval whose size is increasing in the direct effect bound and decreasing in the calculated correlation.

Next, we relax both 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 that recovers the identified set above when the stability violation vanishes.

Researchers may report these identified sets for a range of plausible violation magnitudes, in order to assess how robust their 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 a study's conclusions.

## 5.5 Continuous Outcomes

Appendix C.5 extends our identification and estimation results to settings with continuous outcomes. We 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 remotely sensed variable's conditional density $f_R(r \mid S=o,X,D,Y)$, is uniformly continuous in $Y$—we control the discretization error in Lemma 1. This yields an approximate version of Theorem 1, with a discretization error that we bound explicitly and that vanishes as the bin width shrinks to zero.

Researchers can therefore apply our estimation algorithm to the discretized outcome and adjust inference for the discretization error, for example by using bias-aware confidence intervals. Finer bins reduce discretization error but increase the dimension of the discrete system.

## 5.6 Low Dimensional Covariates

Section 3 derived identification with unrestricted 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.6 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 $\Pr(D=d,S=e \mid X)$ and $\Pr(Y=y,S=o \mid X)$ as smooth functions of $X$ using kernels or series, then estimate the conditional average treatment effects via analogous methods, before averaging over $X$. Appendix

C.6.3 discusses how to apply standard nonparametric methods to our moment conditions.

# 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.[10]

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 researchers 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 pre-trained 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 achieve an 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 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 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

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

in $\mathbb{R}^{4000}$) (Rolf et al., 2021). Both of these pretrained embeddings have been extensively 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.

In Appendix F.3 we provide a series of test that validate the plausability of our assumptions in both scenarios.

## 6.1 Simulation Design

We focus on 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 explicitly allows $\alpha_e \neq \alpha_o$, so that baseline outcomes may shift across samples. In the forest cover setting, $\alpha_e$ is calibrated from grid cells in the experimental sample and $\alpha_o$ is calibrated from grid cells in the observational sample.[11] Similarly, in the Smartcards setting, $\alpha_e$ and $\alpha_o$ 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 that enforces both stability and no direct effects by construction. For each experimental unit, we draw the baseline $Y_i(0) \sim \text{Bernoulli}(\alpha_e)$, randomly assign treatment as $D_i \sim \text{Bernoulli}(0.5)$, then set $Y_i = Y_i(0) + D_i\tau$. We then draw the remotely sensed variable $R_i$ from the empirical distribution of $R \mid Y$ 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 $n_o = 500$. In the Smartcards setting, we set $n_e = 3000$ and $n_o = 1000$. In the Smartcards setting, we also truncate the remotely sensed variable to $\mathbb{R}^{1050}$ for computational tractability. We report results for alternative choices of sample sizes in Appendix G.1. Appendix G.2 provides additional details on the simulation design.

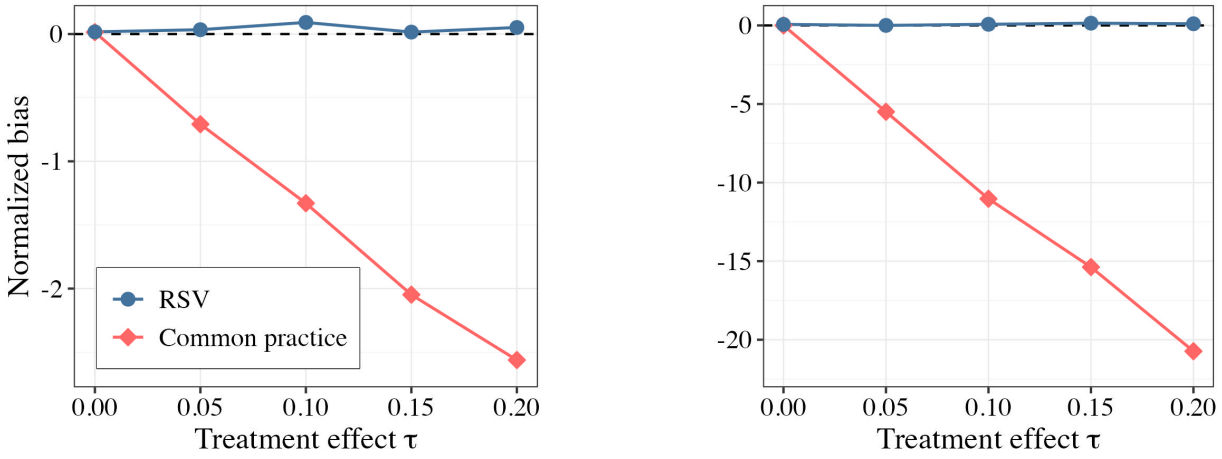
We compare two estimators: the commonly used two-step method (Section 3.2.1), which trains a predictor of $Y$ from $R$ in the observational sample and regresses predicted outcomes on

[11] In the main text, we report results using the two kilometer band as the observational sample. Appendix G.1 provides similar results in which $\alpha_o$ is calibrated using the five or ten kilometer band instead.

$D$ in the experimental sample; and our method (Algorithm 1). Prediction-powered inference 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 across empirical settings. Figure 2 visualizes the results. The commonly used method exhibits large bias that is proportional to the treatment effect. This pattern confirms Proposition 1: this method targets $\kappa\tau$ for some $\kappa \in [0,1)$. Importantly, 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.

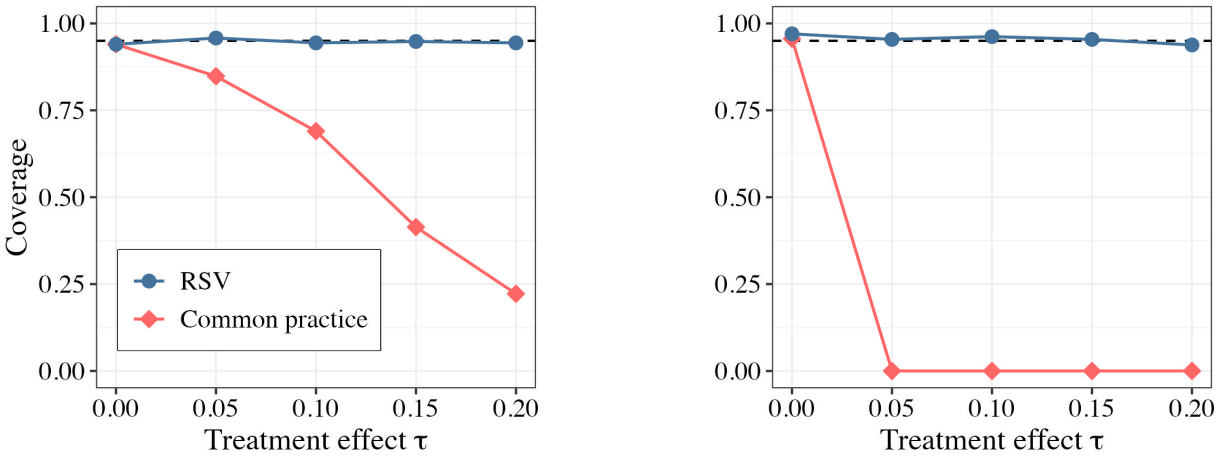



(a) Forest Cover in Uganda.

(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 $\tau$, we conduct 500 simulations.

Next, we compare the quality of the confidence intervals by comparing their coverage, across treatment effect values and across empirical settings. Figure 3 visualizes the results. The 95% confidence intervals based on commonly used method can significantly under-cover—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.




(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 $\tau$, we conduct 500 simulations.

**Remark 4** (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-sampled 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.4). Second, a PPI method that only uses the 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.5).[12] Therefore, even in the presence of a validation sample, 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.

# 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

[12] Only this variation of PPI is unbiased, since $\mathbb{E}\{Y - m_d(R) | D = d, S = e\} = \mathbb{E}\{Y - m_d(R) | D = d, S = v\} \neq \mathbb{E}\{Y - m_d(R) | D = d, S = o\}$. It is inefficient since it throws away the sample $S = o$.

(using satellite images in the experimental sample) against an "oracle" difference-in-means (using direct outcome measurements in the experimental sample). Our method matches the benchmark, despite limited access to outcomes from the experimental sample.

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 underestimates 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 binary measure of village-level poverty constructed from administrative data in 2012–2013. To assess robustness across different poverty definitions, we consider three possible 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 formal 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.[13]

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 $\widetilde{S}=o$ consists of non-study villages, for which we observe $(Y,R)$. Appendix Figure F.2 provides a map. Appendix H.1 provides additional details.

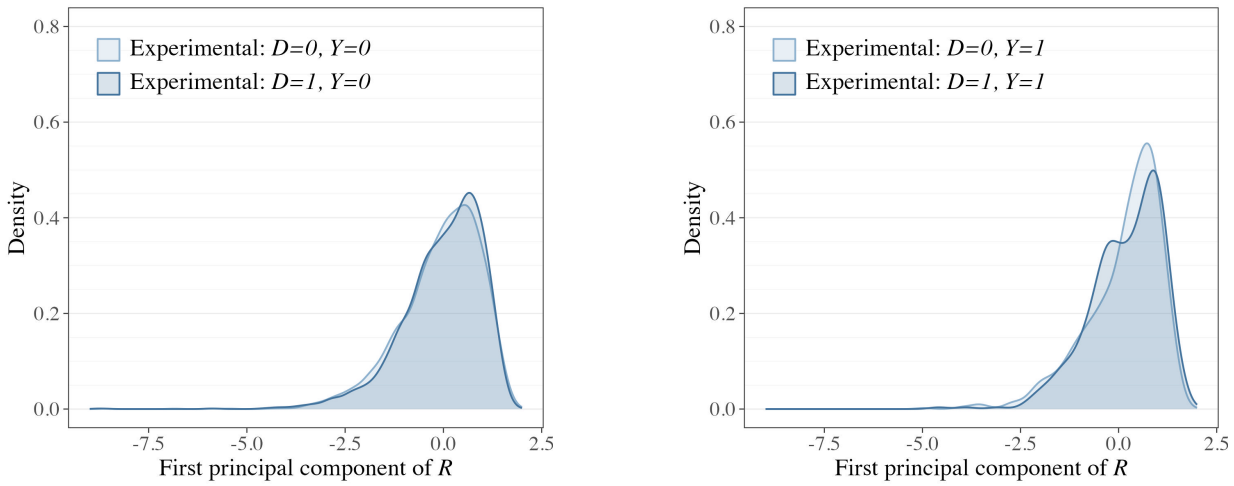
Two aspects of this exercise are noteworthy. First, the experimental sample has 136 mandals while the validation sample has 155 mandals. Whereas the benchmark has "oracle" data access,

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

using outcomes for all 291 mandals, our method has "imperfect" data access, using outcomes for only the 155 mandals in the validation sample. Second, the validation sample is non-randomly selected: rather containing treated and untreated villages, it only contains untreated villages. As a consequence, the study contains no joint observations of $(D=1,Y,R)$. Therefore, the PPI rectifier for the treated group cannot be implemented, and so PPI is infeasible in this setting.[14]

**Plausibility of identifying assumptions.** The treatment was randomized in this experiment (Assumption 1). Figure F.3 provides evidence that stability is plausible (Assumption 2).

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 $y \in \{0,1\}$. The conditional densities closely align, and Kolmogorov–Smirnov tests cannot reject the null of equality at the 5% level.




(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 $f_R \mid \widetilde{S} \in \{e,v\}, D=0, Y=y)$ 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.

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.

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

| | Consumption | Low income | Middle income |
|---|---|---|---|
| Relevance | 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.

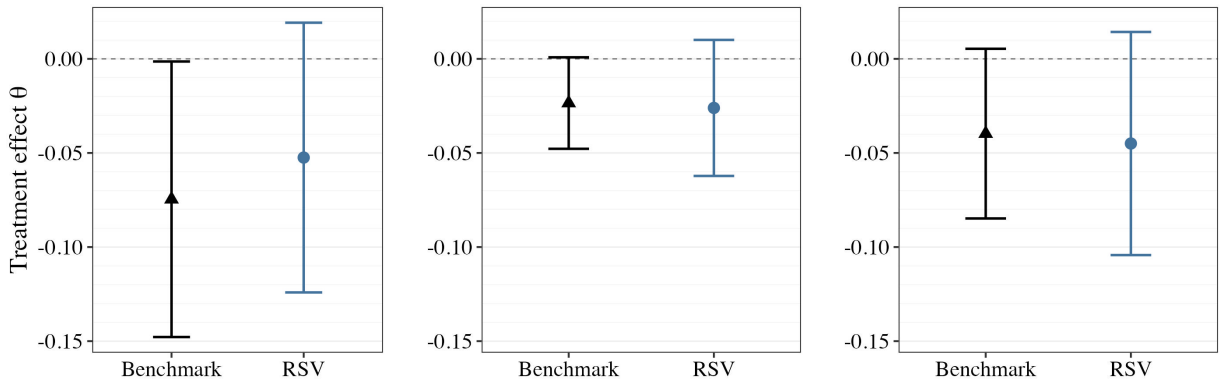
We assess stability and no direct effects for the other two outcomes Appendix H.1.

**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 155 mandal outcomes but leveraging an observational sample.

Our method recovers the treatment effects estimated 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 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 confidence intervals are nearly the same lengths as the benchmark. For the consumption outcome, our interval is essentially of the same length. For the middle income outcome, our interval is 31% longer. On the one hand, the benchmark uses more outcomes from the Smartcards experiment. On the other hand, our method leverages additional information from the observational sample, which effectively compensates for the missing experimental outcomes.

These findings have a practical implication: our method may allow economists to significantly reduce survey costs. An economist could forgo poverty surveys for a large fraction of villages in the Smartcards experiment, and instead rely on free satellite images as well as an observational sample linking satellite images with poverty measurements. If surveying




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

Figure 5: Our method recovers the unbiased benchmark. 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.

each individual in a village costs \$0.50, this choice could save \$3 million.[15]

**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 $H'(R)=\Pr(Y=1\,|\,R,S=o)$, 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 has 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 the two representations in Appendix H.1.

[15]Appendix Table F.1 summarizes the number of villages and their populations, which we use to calculate savings. This calculation is highly conservative; Viviano and Rudder (2024) 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 $\Pr(Y=1 \mid R, S=o)$. The $J$ statistic compares treatment effect estimate based on these two representations. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

**Remark 5** (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 Smarcards experiment in Appendix H.1.

**Remark 6** (Timing)**.** Outcomes were measured at the end of the experiment in 2013, so we study the effect on poverty in 2013. Meanwhile, 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 these 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.[16]

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.[17]

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 a few 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 data, there are no joint observations of $(D,Y = 0,R)$, 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 spot check sample do

[16] 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.

[17] 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 in using their "balanced accuracy" classifier in Appendix H.2.

not provide unbiased estimates of potential outcomes, making PPI biased (Proposition 2).[18] On the other hand, stability of the sensing mechanism is plausible, since it requires that the relationship between what is observed from 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 in 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 specific infrared spectral bands. Other behavioral 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 relevant. We confirm that the learned representation is correlated with outcome variation: $\mathbb{E}_n\{\hat{H}(R)\hat{\Delta}^o\} = 0.262$, with standard error 0.048.[19]

**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, controlling for stratum fixed effects; (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.[20] We interpret this discrepancy as an economically meaningful attenuation bias (Proposition 1).

**Specification test and efficiency gains.** As before, our model generates testable overidentifying restrictions (Remark 1). Once again, we implement this test by comparing estimates

[18] For example, $\tau \neq \tilde{\tau}^o$ 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.

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

[20] 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.

| | Validation | Observational |
|---|---|---|
| Optimal representation | | 0.1417 |
| | | (0.0782) |
| Simple representation | | 0.1929 |
| | | (0.0989) |
| Common practice | | 0.0752 |
| | | (0.0380) |
| $J$-statistic | | 1.7439 |
| $p$-value | | 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 $\Pr(Y=1 \mid R, S=o)$. The $J$ statistic compares treatment effect estimate based on these two representations. Bootstrap standard errors, based on 5000 replications, are clustered at the village level.

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

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 Discussion

We study remotely sensed variables as *post-outcome*: variation in the economic outcome causes changes 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 with jointly observed treatments, outcomes, and remotely sensed variables (for both treatment values). When such an auxiliary sample is unavailable, PPI is infeasible. Even when such an auxiliary sample is available, PPI is biased whenever the experimental and auxiliary samples differ in baseline outcomes or treatment effects (Proposition 2).[21]

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 digital traces reflect economic conditions—operates the same way regardless of where and when the data were collected. 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 $R$ should reflect the outcome $Y$ 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).

At the same time, the observational sample should be constructed to make stability (Assumption 2) credible by selecting an observational sample reflecting similar sensing mechanism (e.g., a region with similar footprints, or similar soil characteristics to the target sample). Discrepancies in the outcome mechanism (perhaps due to timing of data collection) are permitted.

**Inference with Remotely Sensed Outcomes.** Valid inference only requires the remotely sensed variable to be predictive of the outcome. However, more efficient inference leverages three predictions from the remotely sensed variable: predictions of the outcome

[21] In terms of Section 5.1, PPI is biased when using an observational sample rather than a randomly sampled validation sample.

$Y$, the treatment $D$, and the sample indicator $S$. We aggregate these three predictions into an 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, while 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 frameworks for weak instrument detection.

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 and can be formalized as a $J$-test.

# 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. *The Wharton School Research Paper*.

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

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). Inference for regression with variables generated by AI or machine learning. *arXiv:2402.15585*.

Bluhm, R. and G. C. McCord (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.

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

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:2011.06158*.

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), 441–445.

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.

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.

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 (2020). Adversarial estimation of Riesz representers. *arXiv:2101.00009*.

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.

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.

Donaldson, D. and A. Storeygard (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives 30*(4), 171–198.

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

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

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., 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.

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. *National Bureau of Economic Research*.

Imbens, G., N. Kallus, X. Mao, and Y. Wang (2024, 10). 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. Technical report, Haas School of Business, UC Berkeley.

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. *arXiv:2501.09731*.

Johannemann, J., V. Hadad, S. Athey, and S. Wager (2019). Sufficient representations for categorical variables. *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:2501.18577*.

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.

Michaels, G., D. Nigmatulina, F. Rauch, T. Regan, N. Baruah, and A. Dahlstrand (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 R. F. Engle and D. L. McFadden (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.

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. *National Bureau of Economic Research*.

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. *Handbook of Econometrics 6*, 5469–5547.

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*(1), 4392.

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

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 *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 (2024). Policy design in experiments with unknown interference. *arXiv: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: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** **3**

**B Proofs for Theoretical Results in the Main Text** **4**

B.1 Proofs for Section 3 . . . . . 4

B.1.1 Proof of Lemma 1 . . . . . 4

B.1.2 Proof of Theorem 1 . . . . . 4

B.1.3 Proof of Theorem 2 . . . . . 6

B.1.4 Proof of Proposition 1 . . . . . 6

B.1.5 Proof of Proposition 2 . . . . . 8

B.2 Proofs for Section 4 . . . . . 10

B.2.1 Proof of Proposition 3 . . . . . 10

B.2.2 Proof of Proposition 4 . . . . . 12

**C Additional Theoretical Results and Discussion** **14**

C.1 Bias of Alternative Approaches: Extensions . . . . . 14

C.1.1 Bias of Common Practice with Complete Observational Cases . . . . 14

C.1.2 Bias of Common Practice with Random Sample Selection . . . . . 15

C.2 Quasi-Experiments with Remotely Sensed Outcomes . . . . . 15

C.2.1 Instrumental Variables . . . . . 16

C.2.2 Two-Period Difference-in-Differences . . . . . 17

C.3 Some Spillovers . . . . . 19

C.3.1 Generalized Potential Outcomes . . . . . 19

C.3.2 Average Global Treatment Effect . . . . . 21

C.4 Partial Identification under Approximate Identifying Assumptions . . . . . 23

C.4.1 Approximately No Direct Effects . . . . . 23

C.4.2 Approximate Stability and Approximately No Direct Effects . . . . . 25

C.5 Continuous Outcomes . . . . . 27

C.6 Estimation with Covariates . . . . . 29

C.6.1 Discrete Covariates and Incomplete Observational Cases . . . . . 29
C.6.2 Discrete Covariates and Complete Observational Cases . . . . . 30
C.6.3 Continuous Covariates . . . . . 32

**D Proofs for Additional Theoretical Results** **34**
D.1 Bias of Alternative Approaches: Extensions . . . . . 34
D.1.1 Proof of Proposition C.1 . . . . . 34
D.1.2 Proof of Proposition C.2 . . . . . 35
D.2 Quasi-Experiments with Remotely Sensed Outcomes . . . . . 36
D.2.1 Proof of Theorem C.1 . . . . . 36
D.2.2 Proof of Theorem C.2 . . . . . 37
D.3 Some Spillovers . . . . . 38
D.3.1 Proof of Lemma C.1 . . . . . 38
D.3.2 Proof of Theorem C.3 . . . . . 39
D.3.3 Proof of Corollary C.3 . . . . . 40
D.3.4 Proof of Corollary C.4 . . . . . 40
D.4 Partial Identification under Approximate Identifying Assumptions . . . . . 41
D.4.1 Proof of Lemma C.2 . . . . . 41
D.4.2 Proof of Theorem C.4 . . . . . 42
D.4.3 Proof of Lemma C.3 . . . . . 43
D.4.4 Proof of Theorem C.4.2 . . . . . 43
D.5 Continuous Outcomes . . . . . 44
D.5.1 Proof of Proposition C.3 . . . . . 44
D.5.2 Proof of Corollary C.5 . . . . . 45
D.5.3 Proof of Proposition C.4 . . . . . 46
D.6 Estimation with Covariates . . . . . 47
D.6.1 Proof of Proposition C.5 . . . . . 47
D.6.2 Proof of Proposition C.6 . . . . . 47
D.7 Additional Lemmas . . . . . 50

**E Additional Estimation Algorithms** **53**

**F Additional Details on Empirical Settings** **56**
F.1 Forest Cover Details . . . . . 56

F.2 Smartcards Details . . . . . 57
F.3 Plausibility of Identifying Assumptions . . . . . 59

**G Additional Monte Carlo Simulation Results** **61**
G.1 Incomplete Observational Cases . . . . . 61
G.2 Incomplete Observational Cases with Validation Sample . . . . . 65

**H Additional Empirical Results** **69**
H.1 Smartcards Experiment and Village-Level Poverty . . . . . 69
H.2 Payments for Ecosystem Services Experiment and Crop Burning . . . . . 72

# A Additional Model Details

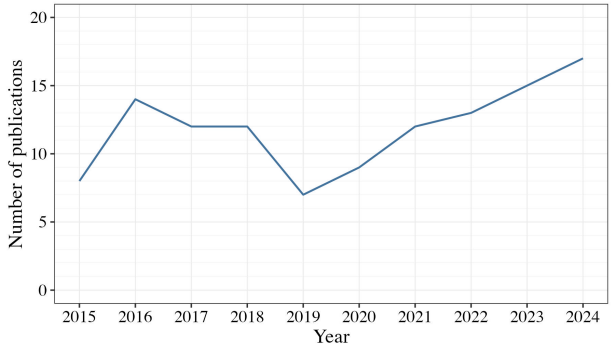



Figure A.1: Remotely sensed variables are increasingly popular in published papers.

Figure A.1 illustrates the increasing popularity of remotely sensed variables in empirical research. First, we collected papers published in the 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 subsetted to the papers with remotely sensed variables in their main empirical analysis.

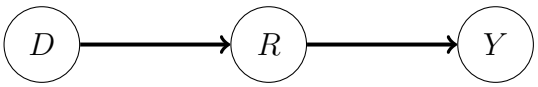



Figure A.2: Causal graph of surrogacy model.

Figure A.2 illustrate 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

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

where the first equality applies Assumption 2 and the second equality applies Assumption 3(ii).

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 the second equality applies Assumption 3(i).

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,

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

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 $f_R(R|X)$ 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 $\gamma_{y_k}$ 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 $(S,D) \perp\!\!\!\perp R \mid Y$,

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

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), $\mathbb{E}\{Y(d)\,|\,S=e\}=\mathbb{E}(Y\,|\,D=d,S=e)$, 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 $Y = \Delta^e$, $W = \Delta^o$, and $U = Y - (\Delta^o)^\top \theta$, $Z = \widetilde{H}(R)$, and $\widehat{Z} = \widehat{H}(R)$. The estimator $\widehat{\theta}$ can then be written as $\widehat{\theta} := \mathbb{E}_n[\widehat{Z}W^\top]^{-1}\mathbb{E}_n[\widehat{Z}Y]$.

While proving this result, we use standard notation for cross-fitting. Let there be $L$ folds, each denoted by $I_\ell$ with $\ell \in [L]$. Each fold contains $n_\ell = n/L$ observations. The complement of $I_\ell$ 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 $I_{-\ell}$.

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

*Proof.* We express the difference as

$$\mathbb{E}_n[\widehat{Z}W^\top] - \mathbb{E}[ZW^\top] =$$

$$\mathbb{E}_n[(\widehat{Z} - Z)W^\top] + (\mathbb{E}_n[ZW^\top] - \mathbb{E}[ZW^\top]).$$

Since the marginal probabilities are bounded away from zero and $\mathbb{E}[\|\widetilde{H}(r)\|^2] < \infty$ by Assumption 4, the $\mathbb{E}_n[ZW^\top] - \mathbb{E}[ZW^\top] = o_p(1)$ by the weak law of large numbers. By the Cauchy-Schwarz inequality,

$$\|\mathbb{E}_n[(\widehat{Z} - Z)W^\top]\| \leq \left(\mathbb{E}_n[\|\widehat{Z} - Z\|^2]\right)^{1/2} \left(\mathbb{E}_n[\|\Delta^o\|^2]\right)^{1/2} = o_p(1),$$

where the last equality follows since the marginal probabilities are bounded away from zero and by Assumption 4. □

**Lemma B.2.** Under Proposition 3's conditions, $\sqrt{n}\mathbb{E}_n[\widehat{Z}U] = \sqrt{n}\mathbb{E}_n[ZU] + o_p(1)$.

*Proof.* We express the difference as

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

Focusing on the foldwise quantity, it suffices to control

$$\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i \in I_\ell}(\widehat{Z}_i - Z_i)U_i\right\}^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i \in I_\ell}(\widehat{Z}_i - Z_i)U_i\right\}^2 \mid I_{-\ell}\right]\right].$$

Due to cross-fitting, the inner expectation is

$$\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i \in I_\ell}(\widehat{Z}_i - Z_i)U_i\right\}^2 \mid I_{-\ell}\right] = \frac{1}{n_\ell}\mathbb{E}\left[\sum_{i,j \in I_\ell}(\widehat{Z}_i - Z_i)U_i U_j^\top (\widehat{Z}_j - Z_j)^\top \mid I_{-\ell}\right]$$

$$= \frac{1}{n_\ell}\mathbb{E}\left[\sum_{i\in I_\ell}(\widehat{Z}_i - Z_i)U_iU_i^\top(\widehat{Z}_i - Z_i)^\top \mid I_{-\ell}\right] = \mathbb{E}[(\widehat{Z}_i - Z_i)U_iU_i^\top(\widehat{Z}_i - Z_i)^\top \mid I_{-\ell}]$$
$$= \mathbb{E}[(\widehat{Z}_i - Z_i)\mathbb{E}[U_iU_i^\top \mid Z_i, I_{-\ell}](\widehat{Z}_i - Z_i)^\top \mid I_{-\ell}].$$

We next use that $\mathbb{E}[U_iU_i^\top \mid Z_i, I_{-\ell}] = \mathbb{E}[U_iU_i^\top \mid Z_i] = \Sigma_{UU}$ for some matrix with $\|\Sigma_{UU}\| < \infty$ since the marginal probabilities are bounded away from zero. Consequently,

$$\mathbb{E}\left[\left\{n_\ell^{1/2}\frac{1}{n_\ell}\sum_{i\in I_\ell}(\widehat{Z}_i - Z_i)U_i\right\}^2 \mid I_{-\ell}\right] = \mathbb{E}[(\widehat{Z}_i - Z_i)\Sigma_{UU}(\widehat{Z}_i - Z_i)^\top \mid I_{-\ell}] = o_p(1)$$

where the final equality follows by Assumption 4 and $\|\mathbb{E}[(\widehat{Z}_i - Z_i)\Sigma_{UU}(\widehat{Z}_i - Z_i)^\top \mid I_{-\ell}]\| \leq \mathbb{E}[\|(\widehat{Z}_i - Z_i)\Sigma_{UU}(\widehat{Z}_i - Z_i)^\top\| \mid I_{-\ell}] \leq \|\Sigma_{UU}\|^2\mathbb{E}[\|\widehat{Z}_i - Z_i)\|^2 \mid I_{-\ell}]$. □

**Lemma B.3.** Under Proposition 3's conditions, $\sqrt{n}\mathbb{E}_n[ZU] \rightsquigarrow \mathcal{N}(0, B)$ for $B = \mathbb{E}[U^2ZZ^\top]$.

*Proof.* By Theorem 1, $\mathbb{E}[ZU] = 0$, and so consequently $\mathbb{E}[ZU] = 0$ by iterated expectations. Moreover, $\mathbb{E}[\|ZU\|^2] < \infty$ under our assumptions. Consequently, the result follows by the multivariate Lindeberg-Levy central limit theorem. □

We are now ready to prove Proposition 3. By definition, $\mathbb{E}_n[\widehat{Z}Y] = \mathbb{E}_n[\widehat{Z}W^\top]\widehat{\theta}$ and $\mathbb{E}[ZY] = \mathbb{E}[ZW^\top]\theta$. Consequently,

$$\mathbb{E}_n[\widehat{Z}W^\top](\widehat{\theta} - \theta) = \mathbb{E}_n[\widehat{Z}Y] - \mathbb{E}_n[\widehat{Z}W^\top]\theta$$
$$= \mathbb{E}_n[\widehat{Z}\big(Y - W^\top\theta\big)] = \mathbb{E}_n[\widehat{Z}U].$$

Consequently,

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

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

Finally, since $\tau = \sum_{k=1}^{K-1}\lambda^\top\theta$ for $\lambda = (y_1 - y_K, \ldots, y_{K-1} - y_K)$ by definition, the result is an immediate consequence of Lemma 3 and the Delta method.

Efficiency follows directly from Chamberlain (1987) and Newey (1993), noting that $\hat{\theta}$ is semiparametric efficient for $\theta$ within the class of estimators satisfying $\mathbb{E}[\Delta^e - \theta^\top\Delta^o | R] = 0$ almost surely and therefore it is $\tau$ since it is defined as a linear transformation of $\theta$. □

### B.2.2 Proof of Proposition 4

With unknown marginal probabilities, we extend the notation introduced in the proof of Lemma 3. In particular, define $\widehat{\Delta}^e = \frac{1\{D=1,S=e\}}{\mathbb{E}_n[1\{D=1,S=e\}]} - \frac{1\{D=0,S=e\}}{\mathbb{E}_n[1\{D=0,S=e\}]}$. Analogously define $\widehat{\Delta}^o_j = \frac{1\{Y=y_j,S=o\}}{\mathbb{E}_n[1\{Y=y_j,S=o\}]} - \frac{1\{Y=y_j,S=o\}}{\mathbb{E}_n[1\{Y=y_j,S=o\}]}$ and $\widehat{\Delta}^o = \left(\widehat{\Delta}^o_1,...,\widehat{\Delta}^o_{K-1}\right)^\top$. We then use the lightened notation: $\widehat{Y} = \widehat{\Delta}^e$, $\widehat{W} = \widehat{\Delta}^o$, and $\widehat{U} = \widehat{Y} - \widehat{W}^\top\theta$. The estimator $\widehat{\theta}$ can then be written as $\widehat{\theta} := \mathbb{E}_n[\widehat{Z}\widehat{W}^\top]^{-1}\mathbb{E}_n[\widehat{Z}\widehat{Y}]$, and the estimator of the ATE in the experimental sample is $\widehat{\tau} = \lambda^\top\widehat{\theta}$ for $\lambda = (y_1 - y_K,...,y_{K-1} - y_K)^\top$.

**Lemma B.4.** Under Proposition 4's conditions, $\mathbb{E}_n[\widehat{Z}\widehat{W}^\top] = \mathbb{E}[ZW^\top] + o_p(1)$.

*Proof.* To prove this result, we will proceed in three steps. First, by an argument similar to the proof of Lemma B.1, $\mathbb{E}_n[\widehat{Z}\widehat{W}^\top] = \mathbb{E}_n[Z\widehat{W}^\top] + o_p(1)$ since $\|\widehat{W}^\top\|$ is finite almost surely under our assumptions. Second, we will next show that $\mathbb{E}_n[Z\widehat{W}^\top] = \mathbb{E}_n[ZW^\top] + o_p(1)$. To show this, we express the difference

$$\mathbb{E}_n(Z\widehat{W}^\top) - \mathbb{E}_n(ZW) = \mathbb{E}_n[Z(\widehat{W} - \widehat{W})^\top] \leq \mathbb{E}_n[\|Z\|^2]^{1/2}\mathbb{E}_n[\|\widehat{W} - W\|^2]^{1/2}.$$

Since $\mathbb{E}_n[\|Z\|^2] = \mathbb{E}[\|Z\|^2] + o_p(1)$ when $\mathbb{E}[\|Z\|^2] < \infty$ by the weak law of large numbers, it suffices to study

$$\mathbb{E}_n[\|\widehat{W} - W\|^2] = \frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{n_\ell}\sum_{i\in I_\ell}\|\widehat{W}_i - W_i\|^2 = \frac{1}{L}\mathbb{E}_\ell[\|\widehat{W} - W\|^2],$$

where $\mathbb{E}_\ell[\cdot] = \frac{1}{n_\ell}\sum_{i\in I_\ell}(\cdot)$. Unpacking the notation of $\mathbb{E}_\ell\{\|\hat{W}_i - W_i\|^2\}$,

$$\begin{aligned}
\hat{W}_j - W_j &= \widehat{\Delta}^o_j - \Delta^o_j = \frac{1_{Y=j,S=o}}{\mathbb{E}_\ell(1_{Y=j,S=o})} - \frac{1_{Y=K,S=o}}{\mathbb{E}_\ell(1_{Y=K,S=o})} - \left\{\frac{1_{Y=j,S=o}}{\mathbb{E}(1_{Y=j,S=o})} - \frac{1_{Y=K,S=o}}{\mathbb{E}(1_{Y=K,S=o})}\right\} \\
&= 1_{Y=j,S=o}\left\{\frac{1}{\mathbb{E}_\ell(1_{Y=j,S=o})} - \frac{1}{\mathbb{E}(1_{Y=j,S=o})}\right\} - 1_{Y=K,S=o}\left\{\frac{1}{\mathbb{E}_\ell(1_{Y=K,S=o})} - \frac{1}{\mathbb{E}(1_{Y=K,S=o})}\right\} \\
&= 1_{Y=j,S=o}\left\{\frac{\mathbb{E}(1_{Y=j,S=o}) - \mathbb{E}_\ell(1_{Y=j,S=o})}{\mathbb{E}_\ell(1_{Y=j,S=o})\mathbb{E}(1_{Y=j,S=o})}\right\} - 1_{Y=K,S=o}\left\{\frac{\mathbb{E}(1_{Y=K,S=o}) - \mathbb{E}_\ell(1_{Y=K,S=o})}{\mathbb{E}_\ell(1_{Y=K,S=o})\mathbb{E}(1_{Y=K,S=o})}\right\}.
\end{aligned}$$

With population and empirical counts bounded away from zero,

$$|\hat{W}_j - W_j| \lesssim |\mathbb{E}(1_{Y=j,S=o}) - \mathbb{E}_\ell(1_{Y=j,S=o})| + |\mathbb{E}(1_{Y=K,S=o}) - \mathbb{E}_\ell(1_{Y=K,S=o})|.$$

By Hoeffding's inequality and the union bound, with probability $1-2\delta$,

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

Therefore with probability $1-\delta$ for all $i \in [n]$ and $j \in \{1,\dots,K-1\}$ simultaneously,

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

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

$$\mathbb{E}_\ell\{\|\hat{W} - W\|^2\} \lesssim \mathbb{E}_\ell\left\{\frac{\ln(2(K-1)/\delta)}{n_\ell}\right\} = \frac{\ln(2(K-1)/\delta)}{n_\ell}.$$

For all $\delta > 0$, this quantities vanishes for large $n_\ell = n/L$, so $\mathbb{E}_\ell\{\|\hat{W} - W\|^2\} = o_p(1)$. The continuous mapping theorem implies the desired result. Finally, we notice that $\mathbb{E}_n[ZW^\top] = \mathbb{E}[ZW^\top] + o_p(1)$ by the same argument as the proof of Lemma B.1. □

**Lemma B.5.** Under Proposition 4's conditions, $\sqrt{n}\mathbb{E}_n[\widehat{Z}\widehat{U}] = \sqrt{n}\mathbb{E}_n[Z\widehat{U}] + o_p(1)$.

*Proof.* The argument is similar to the first step in the proof of Lemma B.4 using the fact that $\|\widehat{U}\|$ is finite almost surely under our assumptions. □

**Lemma B.6.** Under Proposition 4's conditions, $\sqrt{n}\mathbb{E}_n[Z\widehat{U}] \rightsquigarrow \mathcal{N}(0, V)$, where $V = v^\top \Sigma v$, $\Sigma = \mathrm{Var}(B)$,

$$B = \begin{pmatrix} Z1\{D=1, S=e\} \\ 1\{D=1, S=e\} \\ Z1\{D=0, S=e\} \\ 1\{D=0, S=e\} \\ Z1\{Y=\cdot, S=o\} \\ 1\{Y=\cdot, S=o\} \\ Z1\{Y=y_K, S=o\} \\ 1\{Y=y_k, S=o\} \end{pmatrix}$$

with $1\{Y = \cdot, S = o\} = (1\{Y = y_1, S = o\}, \dots, 1\{Y = y_K - 1\})^\top$, $v = \nabla h(\mathbb{E}[B])$, and $h(B) = \left(\frac{Z1\{D=1,S=e\}}{1\{D=1,S=e\}} - \frac{Z1\{D=0,S=e\}}{1\{D=0,S=e\}}\right) - \left(\frac{Z1\{Y=\cdot,S=o\}}{1\{Y=\cdot,S=o\}} - \frac{Z1\{Y=y_K,S=o\}}{1\{Y=y_K,S=o\}}\right)^\top \theta$ for $\frac{Z1\{Y=\cdot,S=o\}}{1\{Y=\cdot,S=o\}}$ denoting element-wise division.

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

$$\widehat{U} = \widehat{Y} - \widehat{W}^\top\theta = \frac{1\{D=1,S=e\}}{\mathbb{E}_n[1\{D=1,S=e\}]} - \frac{1\{D=0,S=e\}}{\mathbb{E}_n[1\{D=0,S=e\}]} - \left(\frac{1\{Y=\cdot,S=o\}}{\mathbb{E}_n[1\{Y=\cdot,S=o\}]} - \frac{1\{Y=y_K,S=o\}}{\mathbb{E}_n[1\{Y=y_K,S=o\}]}\right)\theta.$$

Therefore, for $h(B)$ defined in the statement, $\sqrt{n}\mathbb{E}_n[Z\widehat{U}] = \sqrt{n}h(\mathbb{E}_n[B])$. We observe that under Assumption 4, $\mathbb{E}[\|Z\|^2] < \infty$, and so by the Lindeberg-Levy central limit theorem, $\sqrt{n}(\mathbb{E}_n(B) - \mathbb{E}(B)) \rightsquigarrow \mathcal{N}(0, \Sigma)$. The result then follows by an application of the Delta method. □

**Lemma B.7.** Under Proposition 4's conditions, $\sqrt{n}(\widehat{\theta}-\theta) \rightsquigarrow \mathcal{N}(0, AVA^\top)$ for $A = \mathbb{E}[\widetilde{H}(R)(\Delta^o)^\top]^{-1}$ and $V$ defined in Lemma B.6.

*Proof.* We observe that $\mathbb{E}_n[\widehat{Z}\widehat{Y}] = \mathbb{E}_n[\widehat{Z}\widehat{W}^\top]\widehat{\theta}$ and $\mathbb{E}[ZY] = \mathbb{E}[ZW^\top]\theta$. As a result,

$$\mathbb{E}_n[\widehat{Z}\widehat{W}^\top](\widehat{\theta}-\theta) = \mathbb{E}_n[\widehat{Z}\widehat{Y}] - \mathbb{E}_n[\widehat{Z}\widehat{W}^\top]\theta = \mathbb{E}_n[\widehat{Z}\widehat{U}]$$

and $\sqrt{n}(\widehat{\theta}-\theta) = \sqrt{n}\mathbb{E}_n[\widehat{Z}\widehat{W}^\top]^{-1}\mathbb{E}_n[\widehat{Z}\widehat{U}]$. Lemma B.5 and Lemma B.6 imply that $\sqrt{n}\mathbb{E}_n[\widehat{Z}\widehat{U}] \rightsquigarrow \mathcal{N}(0, V)$. Lemma B.4 and the continuous mapping theorem imply that $\mathbb{E}_n[\widehat{Z}\widehat{W}^\top]^{-1} \xrightarrow{p} \mathbb{E}[ZW^\top]^{-1}$. The result then follows by Slutsky's theorem. □

Since $\widehat{\tau} = \lambda^\top\widehat{\theta}$ and $\tau = \lambda^\top\theta$ for $\lambda = (y_1 - y_K, \ldots, y_{K-1} - y_K)^\top$, the result is an immediate consequence of Lemma B.7 and the Delta method. □

# 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

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

**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 $v_d(r,y) := \frac{\Pr(Y=y|D=d,S=o)}{\Pr(Y=y|D=d,S=e)} \frac{f_R(r|D=d,S=e)}{f_R(r|D=d,S=o)}$;

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 $w_1(r,y) := \frac{\Pr\{Y(0)=y|S=e\}}{\Pr\{Y(1)=y|S=e\}} \frac{f_R(r|D=1,S=e)}{f_R(r|D=0,S=e)}$.

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 $\text{Var}(Y|R,S=o) > 0$ 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. Instrument overlap: $\Pr(Z=1 \mid S=e)$ is 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) $\theta^{LATE} := \mathbb{E}\{Y(1)-Y(0) \mid D(1) > D(0), S=e\}$ 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 $Z,D \perp\!\!\!\perp R | S=o,Y$. It now requires that the instrument $Z$ and the treatment $D$ only effect 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 $\theta^{LATE}$ 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 $d \in \{0,1\}$,

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

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 $z \in \{0,1\}$, $d \in \{0,1\}$ 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, $\theta^{LATE}$ is also identified. Estimation and inference can then follow by suitably stacking moments and extending the discussion provided in Appendix C.6.2.

### 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: $\Pr(D=1)$ is bounded away from zero and one.

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

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

Under Assumption C.4, if we were to observe the outcomes in the quasi-experimental sample, then the average treatment effect on the treated $\theta^{ATT}:=\mathbb{E}\{Y_2(1)-Y_2(0)\,|\,S=e,D=1\}$ 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 $D\perp\!\!\!\perp R_t|S=o,Y_t$. 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 $\theta^{ATT}$ 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 $d \in \{0,1\}$,

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

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, $\theta^{ATT}$ is also identified. Estimation and inference can again follow by suitably stacking moments and extending the discussion provided in Appendix C.6.2.

Our two-period difference-in-differences identification result extends directly 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 $\mathbb{E}(Y_t \,|\, S=e, D_t=d)$. 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.[22] 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 $\theta_n$ 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 $Y_i = Y_i(d_i, \mathbf{d}_{-i})$.

ii. Randomization: $D_i \perp\!\!\!\perp \{Y_i(d_i, \mathbf{d}_{-i})\} \,|\, S_i = e$.[23]

iii. Overlap: $\Pr(D_i = 1 \,|\, S_i = e)$ 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 $\theta_n$ 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.

[22] Formally, we may denote unobserved heterogeneity by $\eta_i$ 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 $Y_i^{\text{dir}}(d_i, \eta_i) = \int Y_i(d_i, \mathbf{d}_{-i}, \eta_i)\} \text{dPr}(\mathbf{d}_{-i} | D_i = d_i, S_i = e, \eta_i)$.

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

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 $\theta_n$ 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 $S_i \perp R_i \,|\, (Y_i, D_i, X_i)$: conditional on a unit's *own* outcome, treatment, and covariates, the distribution of the unit's RSV $R_i$ would be the same had it belonged to the experimental or observational samples. In particular, the unit's RSV 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 treatment statuses of other mandals. We think of this as a plausible approximation; intuitively the satellite image of a village should only depends on variables observed in that village.

In summary, our method extends as long as spillovers are in treatment effects, not in RSV distributions. 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 $\theta_n$ after dropping 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 Partial Identification under Approximate Identifying 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)). In practice, researchers may worry that these conditions hold only approximately—for instance, that the treatment has a small direct effect on the remotely sensed variable, or that the measurement process exhibits minor instability across samples. In this section, we develop partial identification results that allow researchers to assess how sensitive their conclusions are to departures from these assumptions.

We proceed in two steps. First, in Section C.4.1, we maintain stability but relax the no-direct-effects condition, bounding the magnitude of any direct effect. Second, in Section C.4.2, we further relax stability itself. Throughout, we omit covariates ($X=\varnothing$) and restrict attention to binary outcomes $\mathcal{Y}=\{0,1\}$. Our arguments extend pointwise in $x$ for discrete covariates.

### C.4.1 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 $R$ conditional on the outcome $Y$. We continue to work with incomplete observational cases, so $\Pr(D=0\,|\,S=o)=1$ and no treatment variation is observed in the observational sample.

For any measurable function $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$, define

$$\tilde{\tau}_H := \frac{\mathbb{E}[\Delta^e H(R)]}{\mathbb{E}[\Delta^o H(R)]}, \tag{C.1}$$

where $\Delta^e := \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)} - \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)}, \Delta^o = \frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}$. When Assumption 3(ii) (no direct effects) holds, Section 3 showed that $\tilde{\tau}_H=\tau$. When direct effects are present, $\tilde{\tau}_H$ remains identified but may differ from $\tau$. Our goal is to bound this discrepancy.

To state our result, for the chosen representation $H$, define the *direct-effect bias* as, for $y\in\{0,1\}$,

$$\delta_y(H) := \mathbb{E}[H(R)\,|\,S=e,Y=y,D=1] - \mathbb{E}[H(R)\,|\,S=e,Y=y,D=0]. \tag{C.2}$$

This quantity measures how much the treatment shifts the distribution of $H(R)$ holding the outcome fixed.

**Lemma C.2** (Bias under direct effects with incomplete observational cases)**.** 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,

$$\tilde{\tau}_H-\tau=\frac{(1-\mu(1))\delta_0(H)+\mu(1)\delta_1(H)}{\mathbb{E}[\Delta^o H(R)]},$$

where $\mu(1)=\mathbb{E}[Y(1)\,|\,S=e]$.

*Proof.* See Appendix D.4.1. □

Lemma C.2 shows that the discrepancy between $\tilde{\tau}_H$ and $\tau$ is governed by a weighted average of the direct-effect biases $\delta_0(H)$ and $\delta_1(H)$, with weights determined by the treated potential outcome distribution, scaled by the "first-stage" strength $\mathbb{E}[\Delta^o H(R)]$. If the researcher is willing to bound the magnitude of the direct effect, then the following result delivers an identified set for $\tau$.

**Theorem C.4** (Partial identification under bounded direct effects and incomplete observational cases)**.** Maintain the conditions of Lemma C.2. Fix $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$ and suppose $|\delta_y(H)|\leq\bar{\delta}_H$ for $y\in\{0,1\}$. Then,

$$\tau\in\left[\tilde{\tau}_H-\frac{\bar{\delta}_H}{|\mathbb{E}[\Delta^o H(R)]|},\tilde{\tau}_H+\frac{\bar{\delta}_H}{|\mathbb{E}[\Delta^o H(R)]|}\right]\cap[-\mu(0),1-\mu(0)],\tag{C.3}$$

where $\mu(0)=\mathbb{E}[Y(0)\,|\,S=e]$.

Moreover, for any $\tau^\star$ in (C.3), there exists a data-generating process satisfying the maintained assumptions with $|\delta_y(H)|\leq\bar{\delta}_H$ that reproduces the observed values of $\mathbb{E}[\Delta^e H(R)]$ and $\mathbb{E}[\Delta^o H(R)]$ and yields $\tau=\tau^\star$.

*Proof.* See Appendix D.4.2. □

There are several features of Theorem C.4 worth emphasizing. First, the width of the interval scales with the ratio $\bar{\delta}_H/|\mathbb{E}[\Delta^o H(R)]|$—the magnitude of the direct effect relative to the outcome-driven "first stage." The bound is most informative when the remotely sensed variable is highly predictive of the outcome (large $|\mathbb{E}[\Delta^o H(R)]|$) and the scope for direct effects is limited (small $\bar{\delta}_H$). In the environmental application of Examples 1, the treatment (e.g., a PES contract) is unlikely to directly alter satellite reflectance except through its effect on the outcome (e.g., crop burning or deforestation), so $\bar{\delta}_H$ 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. In particular, for any $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$, $\mathbb{E}[H(R)\,|\,S=e,D=0]=(1-\mu(0))\mathbb{E}[H(R)\,|$

$S=o,Y=0]+\mu(0)\mathbb{E}[H(R)\,|\,S=o,Y=1]$, so $\mu(0)$ is identified from observed moments in the experimental control group and the observational sample.

Finally, because the bound $\bar{\delta}_H$ and the first-stage strength $|\mathbb{E}[\Delta^o H(R)]|$ both depend on the choice of representation $H$, Theorem C.4 highlights that the researcher has a meaningful degree of freedom: she can select $H$ to load on features of $R$ for which direct effects are most plausibly small. For instance, in satellite imagery applications, the researcher might construct $H$ from spectral bands that respond primarily to the outcome of interest—such as near-infrared reflectance for vegetation cover, or thermal signatures for crop burning—while avoiding bands that could be affected by the treatment through other channels.

**Remark C.1** (Breakpoint analyses.)**.** Researchers can use Theorem C.4 to conduct a breakpoint analysis: for each $H$, report the smallest value $\bar{\delta}_H$ at which the identified set includes zero (or, more generally, at which a conclusion of interest is overturned) (Masten and Poirier, 2020). This value $\bar{\delta}_H^{\text{bp}} = |\tilde{\tau}_H| \cdot |\mathbb{E}[\Delta^o H(R)]|$ is the minimum direct-effect magnitude (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** (Using multiple moments.)**.** If we have a class $\mathcal{H}$ and, for each $H \in \mathcal{H}$, a bound $\bar{\delta}_H$ such that $|\delta_y(H)| \leq \bar{\delta}_H$ for $y \in \{0,1\}$, then intersecting the single-$H$ bounds tightens identification:

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

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

### C.4.2 Approximate Stability and Approximately No Direct Effects

We now relax both identifying assumptions simultaneously, allowing for both direct effects of $D$ on $R$ conditional on $Y$ (departures from Assumption **??**(ii)) and instability across samples (departures from Assumption **??**). We continue to work with incomplete observational cases ($\Pr(D=0\,|\,S=o)=1$), no covariates, and binary outcomes.

In addition to the direct-effect bias $\delta_y(H)$ defined in (C.2), we now introduce the *stability violation*: for $y \in \{0,1\}$,

$$\gamma_y(H) := \mathbb{E}[H(R)\,|\,S=e,Y=y,D=0] - \mathbb{E}[H(R)\,|\,S=o,Y=y,D=0]. \tag{C.4}$$

This quantity measures how much the distribution of $H(R)$, conditional on the outcome and no treatment, differs across the experimental and observational samples. Under stability (Assumption 2), $\gamma_y(H)=0$ for both $y\in\{0,1\}$.

**Lemma C.3** (Bias under direct effects and stability violations)**.** Suppose Assumption 1 holds with $\Pr(D=0\,|\,S=o)=1$. Fix $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$. Then

$$\mathbb{E}[\Delta^e H(R)]=\tau(\mathbb{E}[\Delta^o H(R)]+s(H))+b(H), \tag{C.5}$$

where $s(H):=\gamma_1(H)-\gamma_0(H)$ captures differential stability violations across outcome values, and $b(H):=(1-\mu(1))\delta_0(H)+\mu(1)\delta_1(H)$ captures the direct-effect bias as in Lemma C.2.

*Proof.* See Appendix D.4.3. □

Equation (**??**) makes transparent how the two violations interact. The direct-effect bias $b(H)$ enters additively, just as in the stability-only case of Lemma C.2. The stability violation $s(H)$ enters multiplicatively, distorting the effective first stage from $\mathbb{E}[\Delta^o H(R)]$ to $\mathbb{E}[\Delta^o H(R)]+s(H)$. Solving for $\tau$ yields

$$\tau=\frac{\tilde{\tau}_H\cdot\mathbb{E}[\Delta^o H(R)]-b(H)}{\mathbb{E}[\Delta^o H(R)]+s(H)},$$

which reduces to $\tau=\tilde{\tau}_H$ when $b(H)=s(H)=0$, and to the expression in Lemma C.2 when $s(H)=0$ alone.

If the researcher is willing to bound both the direct effect and the stability violation, then Lemma **??** delivers an identified set for $\tau$.

**Theorem C.5** (Partial identification with bounded direct effects and stability violations)**.** Maintain the conditions of Lemma **??**. Fix $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$. Suppose $|\delta_y(H)|\leq\bar{\delta}_H$ and $|\gamma_y(H)|\leq\bar{\gamma}_H$ for $y\in\{0,1\}$. Let $A:=\mathbb{E}[\Delta^e H(R)]$ and $B:=\mathbb{E}[\Delta^o H(R)]$. Then,

$$\tau\in\left\{\frac{A-b}{B+s}:b\in[-\bar{\delta}_H,\bar{\delta}_H],s\in[-2\bar{\gamma}_H,2\bar{\gamma}_H],B+s\neq 0\right\}. \tag{C.6}$$

If $|B|>2\bar{\gamma}_H$, the identified set in (C.6) is contained in the interval

$$\tau\in\left[\tilde{\tau}_H-\frac{\bar{\delta}_H+2\bar{\gamma}_H|\tilde{\tau}_H|}{|B|-2\bar{\gamma}_H},\tilde{\tau}_H+\frac{\bar{\delta}_H+2\bar{\gamma}_H|\tilde{\tau}_H|}{|B|-2\bar{\gamma}_H}\right]. \tag{C.7}$$

*Proof.* See Appendix D.4.4. □

We again emphasize that the choice of representation $H$ interacts with width of these bounds. Stability is most plausible for representations that depend on the sensing technology

rather than on sample-specific features—for instance, spectral indices derived from a common satellite platform. Choosing $H$ to load on technologically stable features can simultaneously reduce the plausible value of $\bar{\gamma}_H$ and maintain a strong first stage $|B|$, tightening the bounds in Theorem .

## C.5 Continuous Outcomes

In this section, we extend our results to outcomes that are continuous and bounded. Recall that $f_W(\cdot|...)$ is our general symbol for the Radon-Nikodym derivative. Let $f_{Y(d)}(y\,|\,S,X,R)$ denote the conditional density of $Y(d)$ given $S$, $X$, and $R$.

**Assumption C.10** (Bounded outcome)**.** Suppose that $Y(d)\in[-U,U]^p$ for some $U<\infty$. Suppose that for $d\in\{0,1\}$ and $s\in\{e,o\}$, $f_{Y(d)}(y\,|\,S=s,X,D,R)>0$ for $y\in[-U,U]^p$ almost surely.

We discretize the continuous outcome $Y$ into a discrete approximation $\tilde{Y}_\varepsilon$ by its support. Specifically, define the grid $\mathcal{Y}_\varepsilon=\{y_1,\cdots,y_K\}$, where each grid value is the center of a bin of radius $\varepsilon$. As notation, let $B_\varepsilon(y)$ define an $l_\infty$-ball of radius $\varepsilon>0$ around the value $y$, i.e. $B_y(\varepsilon)=\{y'\in\mathcal{Y}:\|y-y'\|_\infty\leq\varepsilon\}$. Define the discretized parameter $\theta_\varepsilon=(\Pr(Y(1)\in B_\varepsilon(y_1)|S=e)-\Pr(Y(0)\in B_\varepsilon(y_1)|S=e),\cdots,\Pr(Y(1)\in B_\varepsilon(y_{K-1})|S=e)-\Pr(Y(0)\in B_\varepsilon(y_{K-1})|S=e))^\top$.

We assume that the conditional density of the remotely sensed variable $f_R(r\,|\,S=o,D=d,Y=y)$ is continuous in the outcome $y$.

**Assumption C.11.** Suppose that for all $y,y'\in\mathcal{Y},d\in\{0,1\}$

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

for a continuous positive function $\omega(t)$ with $\omega(0)=0$. Let $\inf_r|f_R(r)|\geq\underline{\ell}>0$.

**Proposition C.3.** Let Assumptions 1, 2, and 3(ii) hold with $X=\varnothing$, and $Y$ has (possibly continuous) support $\mathcal{Y}\subset\mathbb{R}$. In addition, let Assumptions C.10-C.11 hold. Then, almost surely

$$|\mathbb{E}\{\Delta^e-\Delta_\varepsilon^{o\top}\theta_\varepsilon|R\}|\leq\frac{2\omega(\varepsilon)}{\underline{\ell}}$$

where $\Delta^e=\frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)}-\frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)}$, and

$$\Delta_\varepsilon^o=\left(\frac{1\{Y\in B_\varepsilon(y_1),S=o\}}{\Pr(Y\in B_\varepsilon(y_1),S=o)}-\frac{1\{Y\in B_\varepsilon(y_K),S=o\}}{\Pr(Y\in B_\varepsilon(y_1),S=o)},\cdots,\frac{1\{Y\in B_\varepsilon(y_{K-1}),S=o\}}{\Pr(Y\in B_\varepsilon(y_{K-1}),S=o)}-\frac{1\{Y\in B_\varepsilon(y_K),S=o\}}{\Pr(Y\in B_\varepsilon(y_{K-1}),S=o)}\right).$$

*Proof.* See Appendix D.5.1. □

**Corollary C.5.** Suppose the conditions of Proposition C.3 hold. Consider any measurable function $H:\mathcal{R}\to\mathbb{R}^J$ with $J\geq K-1$, $\mathbb{E}[\|H(R)\|]<\infty$, and $\mathbb{E}[H(R)\Delta_\varepsilon^{o\top}]$ invertible. Define $\widetilde{\tau}_e = \lambda_\varepsilon^\top \mathbb{E}[H(R)\Delta_\varepsilon^{o,\top}]^{-1}\mathbb{E}[H(R)\Delta^e]$ where $\lambda_\varepsilon = (y_1-y_K,\ldots,y_{K-1}-y_K)$ for the grid $\mathcal{Y}_\varepsilon=\{y_1,\cdots,y_K\}$. Then,

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

*Proof.* See Appendix D.5.2. □

Proposition C.3 shows that our identification result extends to a continuous outcome after discretization. In this case, the discretized parameter $\theta_\varepsilon$ approximately satisfies the conditional moment restriction holds up to a small error that depends on the modulus of continuity of the conditional density of $R\,|\,S=o,Y,D$. Consequently, we could extend our estimation argument to estimate the discretized parameter $\theta_\varepsilon$ (and the associated discretized ATE), adjusting inference for the residual bias, for example, using bias-aware inference. Corollary C.5 further shows that we can further bound the bias for the reduced-form estimand implied by applying our estimation algorithm (Algorithm E.1) to the discretized outcomes.

We establish an analogous result for the setting with complete observational cases. Now define $\mu_\varepsilon(d)=\Big(\Pr(Y(d)\in B_\varepsilon(y_1)\,|\,S=e),\ldots,\Pr(Y(d)\in B_\varepsilon(y_{K-1})\,|\,S=e)\Big)^\top$ for $d\in\{0,1\}$.

**Proposition C.4.** Let Assumptions 1, 2, and 3(i) hold with $X=\varnothing$, and $Y$ has (possibly continuous) support $\mathcal{Y}\subset\mathbb{R}$. In addition, let Assumptions C.10-C.11 hold. Then, for each $d\in\{0,1\}$ almost surely,

$$\left|\mathbb{E}\left\{\widetilde{\Delta}^e(d)-\widetilde{\Delta}_\varepsilon^{o\top}(d)\mu_\varepsilon(d)\;\middle|\;R\right\}\right| \leq \frac{\omega(\varepsilon)}{\underline{\ell}},$$

where

$$\widetilde{\Delta}^e_(d)=\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e)}-\frac{1\{Y\in B_\varepsilon(y_K),D=d,S=o\}}{\Pr(Y\in B_\varepsilon(y_K),D=d,S=o)},$$

$$\widetilde{\Delta}^o_\varepsilon(d)=\left(\frac{1\{Y\in B_\varepsilon(y_1),D=d,S=o\}}{\Pr(Y\in B_\varepsilon(y_1),D=d,S=o)}-\frac{1\{Y\in B_\varepsilon(y_K),D=d,S=o\}}{\Pr(Y\in B_\varepsilon(y_K),D=d,S=o)},\ldots,\frac{1\{Y\in B_\varepsilon(y_{K-1}),D=d,S=o\}}{\Pr(Y\in B_\varepsilon(y_{K-1}),D=d,S=o)}-\frac{1\{Y\in B_\varepsilon(y_K),D=d,S=o\}}{\Pr(Y\in B_\varepsilon(y_K),D=d,S=o)}\right)^\top.$$

*Proof.* See Appendix D.5.3. □

By a similar argument to Corollary C.5, we can further bound the bias for the reduced-form estimand implied by applying our estimation algorithm (Algorithm E.2) to the discretized outcomes.

## C.6 Estimation with Covariates

In this section, we present estimation algorithms allow for covariates. We first consider the case with discrete covariates and then generalize to allow for continuous covariates.

### C.6.1 Discrete Covariates and Incomplete Observational Cases

In this section, we describe our estimation algorithm with incomplete observational cases (Assumption 3(ii)) when 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 the representation that achieves efficiency within a class of models satisfying Theorem 1

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

where $H^*(X,R)\in\mathbb{R}^{K-1}$, $\mathbb{E}\{\Delta^o(X)\,|\,X,R\}\in\mathbb{R}^{K-1}$, and $\sigma^2(\theta,X,R)\in\mathbb{R}$. Its $j$-th component is the scalar $H^*_j(X,R)=\frac{\mathbb{E}\{\Delta^o_j(X)|X,R\}}{\sigma^2(\theta,X,R)}$, where $\Delta^o_j(X)=\frac{1\{Y=y_j,S=o\}}{\Pr(Y=y_j,S=o|X)}$.

Like in the main text, the 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 ATE on test. Algorithm E.1 details the estimation algorithm using sample splitting, though our analysis below will allow 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.12** (Limit)**.** For each $x\in\mathcal{X}$, the learned representation has some mean square limit: $\mathbb{E}\left[\|\widehat{H}(R,x)-\widetilde{H}(R,x)\|^2|\,X=x\right]=o_p(1)$ for some limit $\widetilde{H}(R,x)\in\mathbb{R}^{K-1}$ with $\mathbb{E}[\|\widetilde{H}(R,x)\|^2|\ X=x]<\infty$, Moreover, the limit is correlated with outcome variation: $A(x):=\mathbb{E}[\widetilde{H}(x,R)\Delta_o(x)^\top\,|\,X=x]$ is non-singular for each $x\in\mathcal{X}$, and its smallest singular value is bounded away from zero.

In the following statements, we generalize our asymptotic results in Section 4 to this setting with discrete covariates. We refer to $\Pr(D=d,S=e\,|\,X=x)$ for $d\in\{0,1\}$ and $\Pr(Y=y_j,S=o\,|\,X=x)$ for $j=1,\ldots,K$ as the conditional cell probabilities and $\Pr(X=x\,|\,S=e)$ as the marginal cell probabilities. Let $n_x=\sum_{i=1}^n 1\{X_i=x\}$ denote the number of observations with $X_i=x$

**Proposition C.5.** Suppose Theorem 1's conditions and Assumption C.12 hold. If the conditional cell probabilities are known and bounded away from zero, then for each $x\in\mathcal{X}$

$\sqrt{n_x}\left(\widehat{\theta}(x)-\theta(x)\right) \rightsquigarrow \mathcal{N}(0, A(x)B(x)A(x)^\top)$, 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 $\widehat{\theta}(x)$ is semiparametrically efficient for $\theta(x)$ satisfying (**??**) with conditional cell probabilities known to the researcher.

*Proof.* See Appendix D.6.1. □

**Proposition C.6** (Inference with unknown counts)**.** Suppose Theorem 1 and Assumption C.12 hold. If the marginal and conditional cell probabilities as well as their counting estimators are bounded away from zero, then $\sqrt{n}(\widehat{\tau}-\tau) \rightsquigarrow \mathcal{N}(0, V_\tau)$ for some $V_\tau > 0$ defined in the proof.

*Proof.* See Appendix D.6.2. □

### C.6.2 Discrete Covariates and Complete Observational Cases

We generalize our estimation algorithm to the setting with complete observational cases under Assumption 3(i) with discrete covariates. In this setting, the observational sample contains $(X,D,Y,R)$ so that treatment $D$ is observed in the observational sample, and we allow direct effects of treatment on the remotely sensed variable $R$.

For each treatment $d \in \{0,1\}$ and covariate cell $x \in \mathcal{X}$, define the experimental-sample outcome probabilities $\mu_j(d,x) := \Pr(Y(d) = y_j \mid S = e, X = x)$ for $j = 1,\ldots,K$ and collect $\mu(d,x) := (\mu_1(d,x),\ldots,\mu_{K-1}(d,x))^\top \in \mathbb{R}^{K-1}$, where $\mu_K(d,x) := 1 - \sum_{j=1}^{K-1}\mu_j(d,x)$. The conditional average treatment effect is then $\tau(x) := \sum_{j=1}^{K} y_j(\mu_j(1,x)-\mu_j(0,x))$, and the ATE in the experimental sample is $\tau := \mathbb{E}[\tau(X) \mid S=e]$.

**Choice of Representation:** Corollary 2 implies that for any measurable representation $H_d(X,R)$ with $\mathbb{E}[\widetilde{\Delta}^o(X)^\top H_d(X,R)]$ is non-singular, $\mathbb{E}[H_d(X,R)\left(\widetilde{\Delta}^e(d,x)-\widetilde{\Delta}^o(d,x)^\top\mu(d,x)\right) \mid X = x] = 0$. To choose $H_d(X,R)$, we apply the standard optimal-instrument formula for conditional moment models (Chamberlain, 1987; Newey, 1993). In general, for a conditional moment restriction, $\mathbb{E}[m(\theta) \mid W] = 0$, the efficient instrument is proportional to $H^*(W) \propto \mathbb{E}[\frac{\partial m(\theta)}{\partial \theta^\intercal} \mid W]\,\mathrm{Var}(m(\theta) \mid W)^{-1}$. In this case, fixing $d \in \{0,1\}$, $m(\theta) = \widetilde{\epsilon}_d$ for $\widetilde{\epsilon}_d = \widetilde{\Delta}^e(d,x) - \widetilde{\Delta}^o(d,x)^\intercal\mu(d,x)$. It follows immediately that

$$H_d^*(X,R) = \frac{\mathbb{E}\{\Delta^o(d,X) \mid X,R\}}{\sigma_d^2(\mu,X,R)} \in \mathbb{R}^{K-1}, \quad \sigma_d^2(\mu,X,R) := \mathbb{E}[\{\widetilde{\Delta}^e(d,X)-\widetilde{\Delta}^o(d,X)^\top\mu(d,X)\}^2 \mid X,R].$$

Although $\tau$ depends on $\mu(1,x)-\mu(0,x)$, the structure of the conditional moment restrictions implies that the efficient representation of the stacked system of moments across

treatments simply concatenates the arm-specific efficient representations $H_d^*(X,R)$. More specifically, define the stacked residual

$$\widetilde{\epsilon}=(\widetilde{\epsilon}_0,\widetilde{\epsilon}_1)=\begin{pmatrix}\widetilde{\Delta}^e(0,X)-\widetilde{\Delta}^o(0,X)^\top\mu(0,X)\\ \widetilde{\Delta}^e(1,X)-\widetilde{\Delta}^o(1,X)^\top\mu(1,X)\end{pmatrix},$$

and the stacked parameter $\mu(X)=\left(\mu(0,X)^\top,\mu(1,X)^\top\right)^\top$. The stacked conditional moment is then $\mathbb{E}[\widetilde{\epsilon}\,|\,X,R]=0$.

Since every term in $\widetilde{\Delta}^e(d,X),\widetilde{\Delta}^o(d,X)$ is multiplied by $1\{D=d\}$, it follows that $\widetilde{\epsilon}_0\widetilde{\epsilon}_1=0$ almost surely, and therefore $\mathrm{Cov}(\widetilde{\epsilon}_0,\widetilde{\epsilon}_1\,|\,X,R)=0$. Thus the conditional covariate matrix is block diagonal with

$$\Sigma(X,R):=\mathbb{E}[\widetilde{\epsilon}\widetilde{\epsilon}^\top\,|\,X,R]=\begin{pmatrix}\sigma_0^2(X,R) & 0\\ 0 & \sigma_1^2(X,R)\end{pmatrix}.$$

Moreover, since $\widetilde{\epsilon}_0$ only depends on $\mu(0,X)$ and $\widetilde{\epsilon}_1$ only depends on $\mu(1,X)$, it also follows that

$$\frac{\partial\widetilde{\epsilon}}{\partial\mu(X)^\top}=\begin{pmatrix}-\widetilde{\Delta}^o(0,X)^\top & 0\\ 0 & -\widetilde{\Delta}^o(1,X)^\top\end{pmatrix}.$$

Applying the same formula for the efficient instrument to the stacked moment condition then yields that the efficient representation $H^*(X,R)$ is also block diagonal with blocks equal to the single-arm efficient representation:

$$H^*(X,R):=\begin{pmatrix}H_0^*(X,R) & 0\\ 0 & H_1^*(X,R)\end{pmatrix}.$$

Consequently, for purposes of estimating $\tau$, our estimation algorithm will learn two treatment-specific representations $H_0^*(X,R)$ and $H_1^*(X,R)$ and then combine the resulting estimates of $\mu(0,X)$ and $\mu(1,X)$ into an estimate of the ATE in the experimental sample.

Finally, note that for $d\in\{0,1\}$ and $x\in\mathcal{X}$, we have the following algebraic simplification:

$$\begin{aligned}&(\widetilde{\Delta}^e(d,x)-\widetilde{\Delta}^o(d,x)^\top\mu(d,X))^2=\\ &\frac{1\{D=d,S=e\}}{\Pr(D=d,S=e\,|\,X=x)^2}+\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_k,D=d,S=o\,|\,X=x)^2}+\\ &\sum_{j=1}^{K-1}\mu_j^2(d,x)\left(\frac{1\{Y=y_j,D=d,S=o\}}{\Pr(Y=y_j,D=d,S=o\,|\,X=x)^2}+\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o\,|\,X=x)^2}\right)+\\ &\sum_{j\neq k}\mu_j(d,x)\mu_k(d,x)\frac{1\{Y=y_K,D=d,S=o\}}{\Pr(Y=y_K,D=d,S=o\,|\,X=x)^2}.\end{aligned}$$

Algorithm E.2 contains the detailed estimation algorithm using sample splitting. Establishing the asymptotic properties of the estimation algorithm follows the same arguments as in our analysis of incomplete observational cases with discrete outcomes and discrete covariates.

### C.6.3 Continuous Covariates

As a final extension, Algorithm E.3 shows how to modify our inferential procedure to accommodate continuous covariates. We continue to impose Assumption 3(ii). We focus on low-dimensional covariates with $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$.

Our identification results already apply to continuous covariates as written. The conditional moment equation from Theorem 1 applies for each $x$:

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

where recall $\theta_j(x) = \Pr\{Y(1) = y_j \mid S = e, X = x\} - \Pr\{Y(0) = y_j \mid S = e, X = x\}$. Corollary 1 then implies that, given a representation $H : \mathcal{X} \times \mathcal{R} \to \mathbb{R}^J$ with $J \geq K - 1$,

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

The average treatment effect in the experimental sample is then identified as the linear functional $\tau = \mathbb{E}[\lambda^\top \theta(x)]$ for $\lambda = (y_1 - y_K, \ldots, y_{K-1} - y_K)$.

With continuous covariates, the key modification to our estimation algorithm is that the conditional expectations $\mathbb{E}[H(X,R)\Delta^o(X)^\top \mid X = x]$ and $\mathbb{E}[H(X,R)\Delta^e(X) \mid X = x]$ must be estimated as smooth functions of $x$ using nonparametric methods (e.g., kernel smooth, local polynomials, or series). Valid inference would now additionally require standard smoothness conditions on the conditional expectations as functions of $x$, and undersmoothing or bias correction for their nonparametric estimates to ensure that they contribute only second-order terms. With these modifications, the estimation algorithm proceeds in the same manner: divide the sample into TRAIN and TEST folds; learn the representation on TRAIN, and apply the learned representation to estimate the ATE on TEST.

Analyzing the estimation algorithm 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 RSV moment condition permits misspecified representation learning—continues to hold.

**A Series-Based Estimation Algorithm (Algorithm E.3 )** As notation, given a representation $H(X_i, R_i) \in \mathbb{R}^{K-1}$, define the stacked primitive vector $B_i \in \mathbb{R}^m$ for $m = K(K+2)$

by stacking

$$H(X_i,R_i)1\{D_i=1,S_i=e\},H(X_i,R_i)1\{D_i=0,S_i=e\},1\{D_i=1,S_i=e\},1\{D_i=0,S_i=e\} \text{ and}$$

$$H(X_i,R_i)1\{Y_i=y_1,S_i=o\},...,H(X_i,R_i)1\{Y_i=y_K,S_i=o\},1\{Y_i=y_1,S_i=o\},...,1\{Y_i=y_K,S_i=o\}.$$

Define $g(x)=\mathbb{E}[B \mid X=x]$. Given $g(x)$, we can calculate $\mathbb{E}[H(X,R)\Delta^e(X) \mid X=x]$ and $\mathbb{E}[H(X,R)\Delta^o(X)^\top \mid X=x]$. Consequently, there exists the a smooth mapping $\psi(\cdot)$ such that $\theta(x)=\psi(g(x))$, and therefore $\tau(x)=\lambda^\top\psi(g(x))$ and $\tau=\mathbb{E}[\lambda^\top\psi(g(x)) \mid S_i=e]$.

We note that our characterization of the efficient representation $H^*(R,X)$ in Appendix C.6.1 continues to apply. The estimation algorithm approximates the conditional mean $g(x)$ by series least squares using a basis $p_J:\mathcal{X}\to\mathbb{R}^J$. Denote $P_i=p_J(X_i)$, and let $\widehat{Q}_{\text{TRAIN}}:=\frac{1}{n_{\text{TRAIN}}}\sum_{i\in\text{TRAIN}}P_iP_i^\top$ with $\widehat{Q}_{\text{TEST}}$ defined analogously. Define $\widehat{\beta}_{\text{TRAIN}}\in\mathbb{R}^{J\times m}$ as the OLS coefficient matrix from regressing $B_i$ on $P_i$: $\widehat{\beta}_{tr}=\arg\min_{\beta\in\mathbb{R}^{J\times m}}\frac{1}{n_{\text{TRAIN}}}\sum_{i\in\text{TRAIN}}\|B_i-P_i^\top\beta\|^2$. $\widehat{\beta}_{\text{TEST}}$ is defined analogously. The sieve estimators are then $\widehat{g}_{\text{TRAIN}}(x):=p_J(x)^\top\widehat{\beta}_{\text{TRAIN}}$ and $\widehat{g}_{\text{TRAIN}}(x):=p_J(x)^\top\widehat{\beta}_{\text{TEST}}$.

**Properties: Discussion** As mentioned, with continuous covariates, the key modification is that the conditional moments $g(x)=\mathbb{E}[B \mid X=x]$ entering $\theta(x)=\psi(g(x))$ must be estimated as smooth functions of $x$ using nonparametric methods. Establishing analogues of our discrete covariate asymptotic results for Algorithm E.3 then requires standard sieve regularity conditions. These include: (i) smoothness conditions: each component of the conditional mean $g(\cdot)$ belongs to a Holder/Sobolev class of order $s$ on compact support $\mathcal{X}$ so that $g(\cdot)$ admits a sieve approximation $p_J(x)^\top\beta_J$ with approximation error $\|g(\cdot)-p_J(\cdot)^\top\beta_J\|$ shrinking at a known rate $J\to\infty$; (ii) sieve growth and variance control: the basis $p_J(\cdot)$ has a well-conditioned Gram matrix and $J=J_n$ grows slowly enough relative to sample size to delver consistency and a linearization of the series least squares estimator; and (iii) positivity: the conditional cell probabilities are bounded away from zero on $\mathcal{X}$ and $A(x)=\mathbb{E}[\widetilde{H}(X,R)\Delta^o(X)^\top \mid X=x]$ is nonsingular uniformly on $\mathcal{X}$.

Under (i)-(iii), then standard arguments yield pointwise asymptotic normality of the estimators $\widehat{\theta}(x),\widehat{\tau}(x)$ after either undersmoothing (i.e., choosing $J_n$ so the sieve bias is negligible) or explicit bias correction (e.g., Newey, 1997; Chen, 2007). Moreover, since $\tau=\mathbb{E}[\tau(X) \mid S=e]$ is a smooth functional of $g$, the plug-in estimator $\widehat{\tau}$ obtained by averaging $\widehat{\tau}(X)$ over the experimental covariate distribution is $\sqrt{n}$-asymptotically normal provided the sieve approximation bias for this function is $o(n^{-1/2})$ and again the sieve dimension satisfies standard growth restrictions.

Finally, as in the main text, the representation-learning step enters only as a nuisance:

with sample splitting and the infinite-order Neyman orthogonality of the moment restriction, it suffices that the learned representation $\widehat{H}(X,R)$ converges in mean square to some limit $\widetilde{H}(X,R)$ that is correlated with outcome variation, and no explicit convergence rate is required; like in Assumption C.12.

# 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,

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

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

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

Because $S \perp\!\!\!\perp \{Y(0),Y(1)\}$ and we are in the setting with incomplete observational cases,

$$f_Y(y\,|\,S=o)=f_{Y(0)}(y\,|\,S=o)=f_{Y(0)}(y\,|\,S=e)=f_Y(y\,|\,D=0,S=e).$$

Next, we prove that $f_R(r\,|\,S=o)=f_R(r\,|\,D=0,S=e)$. First, we write

$$f_R(R|S=o)=\int_y f_R(R|Y=y,S=o)f_Y(y|S=o)\mathrm{d}y$$
$$f_R(R|D=0,S=e)=\int_y f_R(R|Y=y,D=0,S=e)f_Y(y|D=0,S=e)\mathrm{d}y.$$

By Assumptions 2 and 3(ii), $(S,D) \perp\!\!\!\perp R\,|\,Y$, hence $f_R(R|Y,S=o)=f_R(R|Y,D=0,S=e)$. Above, we proved that $f_Y(y\,|\,S=o)=f_Y(y\,|\,D=0,S=e)$.

In summary, we have shown

$$w_0(r,y)=\frac{f_Y(y\,|\,S=o)}{f_Y(y\,|\,D=0,S=e)}\frac{f_R(r\,|\,D=0,S=e)}{f_R(r\,|\,S=o)}=1$$
$$w_1(r,y)=\frac{f_Y(y\,|\,S=o)}{f_Y(y\,|\,D=1,S=e)}\frac{f_R(r\,|\,D=1,S=e)}{f_R(r\,|\,S=o)}=\frac{f_Y(y\,|\,D=0,S=e)}{f_Y(y\,|\,D=1,S=e)}\frac{f_R(r\,|\,D=1,S=e)}{f_R(r\,|\,D=0,S=e)}.$$

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,

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

Next, notice that

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

where the first equality applies Assumption C.2 and the second equality applies Assumption C.3.

Combining the previous displays, we arrive at the general result

$$\delta_{d,z}^e(R) = \int \delta_y^o(R) f_Y(y \mid S=e,D=d,Z=z) \mathrm{d}y.$$

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

$$\delta_{d,z}^e(R) = \delta_1^o(R)\alpha(d,z) + \delta_0^o(R)\{1-\alpha(d,z)\} = \delta_0^o(R) + \{\delta_1^o(R) - \delta_0^o(R)\}\alpha(d,z).$$

2. We apply Bayes' rule to rewrite

$$\begin{aligned} \delta_y^o(R) &= f_R(R \mid S=o,Y=y) = \frac{\Pr(Y=y,S=o \mid R) f_R(R)}{\Pr(Y=y,S=o)}, \\ \delta_{d,z}^e(R) &= f_R(R \mid S=e,D=d,Z=z) = \frac{\Pr(D=d,Z=z,S=e \mid R) f_R(R)}{\Pr(D=d,Z=z,S=e)}. \end{aligned}$$

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,

$$\begin{aligned} \delta_{d,t}^e(R_t) &:= f_{R_t}(R_t \mid S=e,D=d) \\ &= \int f_{R_t,Y_t}(R_t,y \mid S=e,D=d) \mathrm{d}y \\ &= \int f_{R_t}(R_t \mid S=e,D=d,Y_t=y) f_{Y_t}(y \mid S=e,D=d) \mathrm{d}y. \end{aligned}$$

Next, notice that

$$\begin{aligned} f_{R_t}(R_t \mid S=e,D=d,Y_t=y) &= f_{R_t}(R_t \mid S=o,D=d,Y_t=y) \\ &= f_{R_t}(R_t \mid S=o,Y_t=y) \\ &=: \delta^o_{y,t}(R_t), \end{aligned}$$

where the first equality applies Assumption C.5 and the second equality applies Assumption C.6.

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 $\eta_i$, as defined in Footnote 22,

$$\begin{aligned} \mathbb{E}(Y_i \mid D_i=d,S_i=e) &= \mathbb{E}\{Y_i(D_i,\mathbf{D}_{-i},\eta_i) \mid D_i=d,S_i=e\} \\ &= \mathbb{E}[\mathbb{E}\{Y_i(D_i,\mathbf{D}_{-i},\eta_i) \mid D_i=d,S_i=e,\eta_i)\} \mid D_i=d,S_i=e] \\ &= \mathbb{E}\{Y_i^{\mathrm{dir}}(d,\eta_i) \mid D_i=d,S_i=e\} \\ &= \mathbb{E}\{Y_i^{\mathrm{dir}}(d,\eta_i) \mid S_i=e\} \end{aligned}$$

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,

$$\begin{aligned}\delta^e_{d,i}(R_i) &:= f_{R_i}(R_i \mid S_i = e, D_i = d) \\ &= \int f_{R_i,Y_i}(R_i, y \mid S_i = e, D_i = d)\mathrm{d}y \\ &= \int f_{R_i}(R_i \mid S_i = e, D_i = d, Y_i = y) f_{Y_i}(y \mid S_i = e, D_i = d)\mathrm{d}y.\end{aligned}$$

Next, notice that

$$\begin{aligned}f_{R_i}(R_i \mid S_i = e, D_i = d, Y_i = y) &= f_{R_i}(R_i \mid S_i = o, D_i = d, Y_i = y) \\ &= f_{R_i}(R_i \mid S_i = o, Y_i = y) \\ &=: \delta^o_{y,i}(R_i),\end{aligned}$$

where the first equality applies Assumption C.8 and the second equality applies Assumption C.9.

Combining the previous displays, we arrive at the general result

$$\delta^e_{d,i}(R_i) = \int \delta^o_{y,i}(R_i) f_{Y_i}(y \mid S_i = e, D_i = d)\mathrm{d}y.$$

When $Y$ is binary,

$$\begin{aligned}f_{Y_i}(1 \mid S_i = e, D_i = d) &= \mathbb{E}(Y_i \mid S_i = e, D_i = d) =: \mu_i(d) \\ f_{Y_i}(0 \mid S_i = e, D_i = d) &= 1 - \mathbb{E}(Y_i \mid S_i = e, D_i = d) = 1 - \mu_i(d).\end{aligned}$$

By the proof of Lemma C.1, $\mu_i(d) = \mathbb{E}\{Y_i^{\mathrm{dir}}(d) | S_i = e\}$. Therefore, the general result specializes to

$$\delta^e_{d,i}(R_i) = \delta^o_{1,i}(R_i)\mu_i(d) + \delta^o_{0,i}(R_i)\{1 - \mu_i(d)\} = \delta^o_{0,i}(R_i) + \{\delta^o_{1,i}(R_i) - \delta^o_{0,i}(R_i)\}\mathbb{E}\{Y_i^{\mathrm{dir}}(d) | S_i = e\}.$$

2. We apply Bayes' rule to rewrite

$$\begin{aligned}\delta^o_{y,i}(R_i) &= f_{R_i}(R_i \mid S_i = o, Y_i = y) = \frac{\Pr(Y_i = y, S_i = o \mid R_i) f_{R_i}(R_i)}{\Pr(Y_i = y, S_i = o)}, \\ \delta^e_{d,i}(R_i) &= f_{R_i}(R_i \mid S_i = e, D_i = d) = \frac{\Pr(D_i = d, S_i = e \mid R_i) f_{R_i}(R_i)}{\Pr(D_i = d, S_i = e)}.\end{aligned}$$

Substituting these expressions into the previous step and canceling $f_{R_i}(R_i)$ gives

$$\mathbb{E}(\Delta^e_i | R_i) = \mathbb{E}(\Delta^o_i | R_i)\theta_i, \quad \theta_i = \mathbb{E}\{Y_i^{\mathrm{dir}}(1) - Y_i^{\mathrm{dir}}(0) | S_i = e\}.$$

We conclude that $\theta_n = \frac{1}{n}\sum_{i=1}^{n}\theta_i = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}(\Delta_i^e|R_i)}{\mathbb{E}(\Delta_i^o|R_i)}$.

### 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 22,

$$\begin{aligned}
\mathbb{E}\{Y_1^{\text{dir}}(1,\eta_1)|S_1 = e\} &= \mathbb{E}\{Y_1(1,D_2,...,D_n,\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,D_2,...,D_m,\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{m-1},\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|D_1 = 1, S_1 = e\} \\
&= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|S_1 = e\}
\end{aligned}$$

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 22,

$$\begin{aligned}\mathbb{E}\{Y_1^{\text{dir}}(1,\eta_1)|S_1=e\} &= \mathbb{E}\{Y_1(1,D_2,...,D_n,\eta_1)|D_1=1,S_1=e\}\\ &= \mathbb{E}\{Y_1(1,D_2,...,D_m,\eta_1)|D_1=1,S_1=e\}\\ &= \mathbb{E}\{Y_1(1,\mathbf{1}_{m-1},\eta_1)|D_1=1,S_1=e\}\\ &= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|D_1=1,S_1=e\}\\ &= \mathbb{E}\{Y_1(1,\mathbf{1}_{n-1},\eta_1)|S_1=e\}\end{aligned}$$

by spillovers only within the radius, the subsetting rule, spillovers only within the radius, and randomization.

## D.4 Partial Identification under Approximate Identifying Assumptions

### D.4.1 Proof of Lemma C.2

Fix $H$ with $\mathbb{E}[\Delta^o H(R)] \neq 0$. We decompose the argument into three pieces.

- **(A) Reduce $\mathbb{E}[\Delta^e H(R)]$ and $\mathbb{E}[\Delta^o H(R)]$ to differences in conditional means.**
  - By the definition of $\Delta^e$ (and randomization in $S=e$ by Assumption 1),

$$\mathbb{E}[\Delta^e H(R)] = \mathbb{E}[H(R)\,|\,S=e,D=1] - \mathbb{E}[H(R)\,|\,S=e,D=0].$$

  - Since $\Pr(D=0\,|\,S=o)=1$, the definition of $\Delta^o$ gives

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

- **(B) Expand the experimental difference using the law of iterated expectations and $\delta_y(H)$.**
  - For each $d\in\{0,1\}$,

$$\mathbb{E}[H(R)\,|\,S=e,D=d] = (1-\mu(d))\mathbb{E}[H(R)\,|\,S=e,Y=0,D=d] + \mu(d)\mathbb{E}[H(R)\,|\,S=e,Y=1,D=d].$$

  - Use $\mathbb{E}[H(R)\,|\,S=e,Y=y,D=1] = \mathbb{E}[H(R)\,|\,S=e,Y=y,D=0] + \delta_y(H)$ to write

$$\begin{aligned}&\mathbb{E}[H(R)\,|\,S=e,D=1] - \mathbb{E}[H(R)\,|\,S=e,D=0]\\ &\quad = (\mu(1)-\mu(0))\Big(\mathbb{E}[H(R)\,|\,S=e,Y=1,D=0] - \mathbb{E}[H(R)\,|\,S=e,Y=0,D=0]\Big)\\ &\qquad + (1-\mu(1))\delta_0(H) + \mu(1)\delta_1(H).\end{aligned}$$

- **(C) Replace the $D=0$ difference by $\mathbb{E}[\Delta^o H(R)]$ using stability, then divide.**
  - By Assumption 2,

$$\mathbb{E}[H(R)\,|\,S=e,Y=y,D=0]=\mathbb{E}[H(R)\,|\,S=o,Y=y,D=0].$$

  - With $\theta=\mu(1)-\mu(0)$, this yields

$$\mathbb{E}[\Delta^e H(R)]=\theta\mathbb{E}[\Delta^o H(R)]+(1-\mu(1))\delta_0(H)+\mu(1)\delta_1(H).$$

    Dividing by $\mathbb{E}[\Delta^o H(R)]$ and using $\tilde{\theta}_H=\mathbb{E}[\Delta^e H(R)]/\mathbb{E}[\Delta^o H(R)]$ gives (**??**).

### D.4.2 Proof of Theorem C.4

Fix $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$. 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.

**Step 1: Necessity (outer bound).** By Lemma C.2,

$$\tilde{\theta}_H-\theta=\frac{(1-\mu(1))\delta_0(H)+\mu(1)\delta_1(H)}{\mathbb{E}[\Delta^o H(R)]}.$$

Since $Y\in\{0,1\}$ we have $\mu(1)\in[0,1]$, hence $(1-\mu(1))$ and $\mu(1)$ are nonnegative and sum to one. Therefore,

$$\Big|(1-\mu(1))\delta_0(H)+\mu(1)\delta_1(H)\Big|\leq(1-\mu(1))|\delta_0(H)|+\mu(1)|\delta_1(H)|\leq\bar{\delta}_H,$$

which implies

$$|\tilde{\theta}_H-\theta|\leq\frac{\bar{\delta}_H}{|\mathbb{E}[\Delta^o H(R)]|}.$$

This gives the first interval in (**??**). Finally, $\mu(1)=\mu(0)+\theta\in[0,1]$ implies $\theta\in[-\mu(0),1-\mu(0)]$, so intersecting yields (**??**).

**Step 2: Sharpness (attainability) with respect to observed moments over $H$.** Pick any $\theta^\star$ that belongs to the right-hand side of (**??**) and set $\mu(1)^\star\equiv\mu(0)+\theta^\star\in[0,1]$. Define

$$b^\star\equiv\mathbb{E}_{\text{obs}}[\Delta^e H(R)]-\theta^\star\mathbb{E}_{\text{obs}}[\Delta^o H(R)].$$

Because $\theta^\star$ lies in the interval, $|b^\star|\leq\bar{\delta}_H$.

We construct a DGP satisfying the maintained assumptions and reproducing the observed moments in (C.1) while yielding $\theta=\theta^\star$ as follows:

- Keep the observed distribution of $(Y,R)$ in $S=o$ unchanged, so $\mathbb{E}_\star[\Delta^o H(R)] = \mathbb{E}_{\text{obs}}[\Delta^o H(R)]$.
- Specify the conditional distribution of $R$ under $(S=e,D=1,Y=y)$ so that, for each $y\in\{0,1\}$,

$$\mathbb{E}_\star[H(R)\,|\,S=e,Y=y,D=1]=\mathbb{E}_\star[H(R)\,|\,S=e,Y=y,D=0]+b^\star,$$

  i.e., set $\delta_0(H)=\delta_1(H)=b^\star$; this respects $|\delta_y(H)|\le\bar{\delta}_H$.
- Choose potential outcomes in $S=e$ so that $\mathbb{E}_\star[Y(0)\,|\,S=e]=\mu(0)$ and $\mathbb{E}_\star[Y(1)\,|\,S=e]=\mu(1)^\star$, which implies $\theta=\theta^\star$. Randomize $D$ in $S=e$ as in Assumption 1.

It remains to verify that $\mathbb{E}_\star[\Delta^e H(R)]=\mathbb{E}_{\text{obs}}[\Delta^e H(R)]$. By Lemma C.2 applied under $\mathbb{E}_\star$,

$$\mathbb{E}_\star[\Delta^e H(R)]=\theta^\star\mathbb{E}_\star[\Delta^o H(R)]+(1-\mu(1)^\star)b^\star+\mu(1)^\star b^\star=\theta^\star\mathbb{E}_{\text{obs}}[\Delta^o H(R)]+b^\star=\mathbb{E}_{\text{obs}}[\Delta^e H(R)],$$

where the last equality uses the definition of $b^\star$. Thus the observed moments in (C.1) are reproduced and the constructed DGP yields $\theta=\theta^\star$. Since $\theta^\star$ was arbitrary, the proof completes.

### D.4.3 Proof of Lemma C.3

[XX TO FILL IN XX]

### D.4.4 Proof of Theorem C.4.2

From Lemma D.4 (Equation (D.4)), write

$$A=\theta(B+s)+b,\qquad s\equiv\gamma_1(H)-\gamma_0(H),\qquad b\equiv(1-\mu(1))\delta_0(H)+\mu(1)\delta_1(H).$$

Since $\mu(1)\in[0,1]$ and $|\delta_y(H)|\le\bar{\delta}_H$, we have $|b|\le\bar{\delta}_H$. Also, $|\gamma_y(H)|\le\bar{\gamma}_H$ implies $|s|\le 2\bar{\gamma}_H$. Solving $A=\theta(B+s)+b$ for $\theta$ yields the set inclusion (**??**).

If $|B|>2\bar{\gamma}_H$, then $|B+s|\ge|B|-|s|\ge|B|-2\bar{\gamma}_H>0$. Using $A=\tilde{\theta}_H B$, we can write

$$\theta-\tilde{\theta}_H=-\frac{b+\tilde{\theta}_H s}{B+s},$$

so

$$|\theta-\tilde{\theta}_H|\le\frac{|b|+|\tilde{\theta}_H||s|}{|B+s|}\le\frac{\bar{\delta}_H+2\bar{\gamma}_H|\tilde{\theta}_H|}{|B|-2\bar{\gamma}_H},$$

which gives (**??**).

## D.5 Continuous Outcomes

### D.5.1 Proof of Proposition C.3

Write $B_k := B_\varepsilon(y_k)$.

**Step 1.** Assumption 2(i) gives $S \perp\!\!\!\perp R \,|\, (D,Y)$ and Assumption 3(ii) gives $D \perp\!\!\!\perp R \,|\, Y$. Hence $(S,D) \perp\!\!\!\perp R \,|\, Y$, so for every $d \in \{0,1\}$ and $y \in \mathcal{Y}$,

$$f_R(r \,|\, S=e, D=d, Y=y) = f_R(r \,|\, S=o, Y=y) \quad \text{for all } r.$$

**Step 2.** By the law of total probability and Step 1,

$$f_R(r \,|\, S=e, D=d) = \int f_R(r \,|\, S=o, Y=y) d\Pr\{Y \le y \,|\, S=e, D=d\}.$$

By randomization (Assumption 1(i)) with $X = \varnothing$, $\Pr\{Y \le y \,|\, S=e, D=d\} = \Pr\{Y(d) \le y \,|\, S=e\}$, so

$$f_R(r \,|\, S=e, D=d) = \int f_R(r \,|\, S=o, Y=y) d\Pr\{Y(d) \le y \,|\, S=e\}. \tag{D.1}$$

Partitioning $\mathcal{Y}$ into $\{B_k\}_{k=1}^K$ gives

$$f_R(r \,|\, S=e, D=d) = \sum_{k=1}^K \int_{B_k} f_R(r \,|\, S=o, Y=y) d\Pr\{Y(d) \le y \,|\, S=e\}.$$

Fix $k$ and any $y \in B_k$. Assumption C.11 and $|y - y'| \le 2\varepsilon$ for $y, y' \in B_k$ imply

$$\sup_{r \in \mathcal{R}} |f_R(r \,|\, S=o, Y=y) - f_R(r \,|\, S=o, Y \in B_k)| \le \omega(\varepsilon).$$

Therefore, replacing $f_R(r \,|\, S=o, Y=y)$ by $f_R(r \,|\, S=o, Y \in B_k)$ inside each bin integral in (D.1) induces at most $\omega(\varepsilon)$ error (uniformly in $r$), yielding

$$f_R(r \,|\, S=e, D=d) = \sum_{k=1}^K \Pr\{Y(d) \in B_k \,|\, S=e\} f_R(r \,|\, S=o, Y \in B_k) + \mathrm{err}_d(r), \tag{D.2}$$

with $\sup_{r \in \mathcal{R}} |\mathrm{err}_d(r)| \le \omega(\varepsilon)$.

Subtracting (D.2) for $d=0$ from $d=1$ and using the baseline construction of $\theta_\varepsilon$ gives

$$f_R(r \,|\, S=e, D=1) - f_R(r \,|\, S=e, D=0) = \sum_{k=1}^{K-1} \theta_k \Big( f_R(r \,|\, S=o, Y \in B_k) - f_R(r \,|\, S=o, Y \in B_K) \Big) + \mathrm{err}(r), \tag{D.3}$$

where $\mathrm{err}(r) = \mathrm{err}_1(r) - \mathrm{err}_0(r)$ and hence $|\mathrm{err}(r)| \le 2\omega(\varepsilon)$.

**Step 3.** By Bayes' rule,

$$\mathbb{E}\left[ \frac{1\{D=d, S=e\}}{\Pr(D=d, S=e)} \,\middle|\, R=r \right] = \frac{f_R(r \,|\, S=e, D=d)}{f_R(r)}, \qquad \mathbb{E}\left[ \frac{1\{Y \in B_k, S=o\}}{\Pr(Y \in B_k, S=o)} \,\middle|\, R=r \right] = \frac{f_R(r \,|\, S=o, Y \in B_k)}{f_R(r)}.$$

Combining this with (D.3) yields

$$\mathbb{E}\big[\Delta^e - (\Delta_\varepsilon^o)^\top \theta \,|\, R=r\big] = \frac{\mathrm{err}(r)}{f_R(r)},$$

proving the desired result since Assumption C.11 implies $|f_R(r)| \geq \underline{l}$.

### D.5.2 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.

First, for any $d \in \{0,1\}$, observe

$$|\mathbb{E}[Y(d) \,|\, S=e] - \mathbb{E}[Y_\varepsilon(d) \,|\, S=e]| \leq \mathbb{E}[|Y(d) - Y_\varepsilon(d)| \,|\, S=e] \leq \varepsilon.$$

Defining $\tau_\varepsilon = \mathbb{E}[Y_\varepsilon(1) - Y_\varepsilon(0) \,|\, S=e]$, it follows that

$$|\tau - \tau_e| \leq |(\mathbb{E}[Y(1) \,|\, S=e] - \mathbb{E}[Y_\varepsilon(1) \,|\, S=e]) - (\mathbb{E}[Y(0) \,|\, S=e] - \mathbb{E}[Y_\varepsilon(0) \,|\, S=e])| \leq 2\varepsilon.$$

Defining $U_\varepsilon = \Delta^e - \Delta_\varepsilon^{o\top} \theta_\varepsilon$, Proposition C.3 establishes $\mathbb{E}[U_\varepsilon \mid R] \leq \frac{2\omega(\varepsilon)}{\underline{\ell}}$. By iterated expectations, $\mathbb{E}[H(R) U_\varepsilon] = \mathbb{E}[H(R) \mathbb{E}[U_\varepsilon \,|\, R]]$. Applying Holder's inequality, we then have that

$$\|\mathbb{E}[H(R) U_\varepsilon]\| \leq \frac{2\omega(\varepsilon)}{\underline{\ell}} \mathbb{E}[\|H(R)\|].$$

As notation, let $\widetilde{\theta}_\varepsilon = \mathbb{E}[H(R) \Delta_\varepsilon^{o,\top}]^{-1} \mathbb{E}[H(R) \Delta^e]$. By definition, it then follows that

$$0 = \mathbb{E}[H(R)(\Delta^e - \Delta_\varepsilon^{o\top} \widetilde{\theta}_\varepsilon)] = \mathbb{E}[H(R) \Delta^e] - \mathbb{E}[H(R) \Delta_\varepsilon^{o\top}] \widetilde{\theta}_\varepsilon.$$

At the same time,

$$\mathbb{E}[H(R) \Delta^e] = \mathbb{E}[H(R)(\Delta_\varepsilon^{o\top} \theta_\varepsilon + U_\varepsilon)] = \mathbb{E}[H(R) \Delta_\varepsilon^{o\top}] \theta_\varepsilon + \mathbb{E}[H(R) U_\varepsilon].$$

Combining the previous two display equations then yields

$$\mathbb{E}[H(R) \Delta_\varepsilon^{o\top}] \widetilde{\theta}_\varepsilon = \mathbb{E}[H(R) \Delta_\varepsilon^{o\top}] \theta_\varepsilon + \mathbb{E}[H(R) U_\varepsilon].$$

Since $\mathbb{E}[H(R) \Delta_\varepsilon^{o\top}]$ is invertible, it therefore follows that

$$\widetilde{\theta}_\varepsilon - \theta_\varepsilon = \mathbb{E}[H(R) \Delta_\varepsilon^{o\top}]^{-1} \mathbb{E}[H(R) U_\varepsilon].$$

It then follows that

$$\|\widetilde{\theta}_\varepsilon - \theta_\varepsilon\| \leq \|\mathbb{E}[H(R)\Delta_\varepsilon^{o\top}]^{-1}\| \frac{2\omega(\varepsilon)}{\underline{\ell}} \mathbb{E}[\|H(R)\|].$$

We next translate this bound back into the error for the ATE in the experimental sample. Observe that

$$\widetilde{\tau}_\varepsilon - \tau_\epsilon = \lambda_\varepsilon^\top (\widetilde{\theta}_\varepsilon - \theta_\varepsilon).$$

Consequently,

$$\|\widetilde{\tau}_\varepsilon - \tau_\epsilon\| \leq \|\lambda_\varepsilon\| \|\mathbb{E}[H(R)\Delta_\varepsilon^{o\top}]^{-1}\| \frac{2\omega(\varepsilon)}{\underline{\ell}} \mathbb{E}[\|H(R)\|].$$

Finally, to prove the result,

$$|\widetilde{\tau}_\varepsilon - \tau| \leq |\widetilde{\tau}_\varepsilon - \tau_\varepsilon| + |\tau_\varepsilon - \tau|.$$

by the triangle inequality. The result is then immediate. □

### D.5.3 Proof of Proposition C.4

Fix $d \in \{0,1\}$. By Assumption 2(i), $S \perp\!\!\!\perp R \mid (D,Y)$, hence $f_R(r \mid S=e, D=d, Y=y) = f_R(r \mid S=o, D=d, Y=y)$ for all $(r,y)$. By the law of total probability and randomization (Assumption 1(i)),

$$f_R(r \mid S=e, D=d) = \int f_R(r \mid S=o, D=d, Y=y) d\Pr\{Y(d) \leq y \mid S=e\}.$$

Partitioning $\mathcal{Y}$ into $\{B_\varepsilon(y_k)\}_{k=1}^K$ yields

$$f_R(r \mid S=e, D=d) = \sum_{k=1}^K \int_{B_\varepsilon(y_k)} f_R(r \mid S=o, D=d, Y=y) d\Pr\{Y(d) \leq y \mid S=e\}.$$

For any $y \in B_\varepsilon(y_k)$, the quantity $f_R(r \mid S=o, D=d, Y \in B_\varepsilon(y_k))$ is a within-bin average of $f_R(r \mid S=o, D=d, Y=y')$ over $y' \in B_\varepsilon(y_k)$; since $|y-y'| \leq \varepsilon$ within the bin, the continuity assumption implies

$$\sup_{r \in \mathcal{R}} \Big| f_R(r \mid S=o, D=d, Y=y) - f_R(r \mid S=o, D=d, Y \in B_\varepsilon(y_k)) \Big| \leq \omega(\varepsilon).$$

Replacing $f_R(r \mid S=o, D=d, Y=y)$ by $f_R(r \mid S=o, D=d, Y \in B_\varepsilon(y_k))$ inside each bin integral therefore changes $f_R(r \mid S=e, D=d)$ by at most $\omega(\varepsilon)$ (uniformly in $r$), giving

$$f_R(r \mid S=e, D=d) = \sum_{k=1}^K \Pr\{Y(d) \in B_\varepsilon(y_k) \mid S=e\} f_R(r \mid S=o, D=d, Y \in B_\varepsilon(y_k)) + \mathrm{err}_d(r),$$

with $\sup_r|\text{err}_d(r)|\leq\omega(\varepsilon)$. Writing the $k=K$ term as the baseline and using $\sum_{k=1}^{K}\Pr\{Y(d)\in B_k\,|\,S=e\}=1$ yields

$$f_R(r\,|\,S=e,D=d)-f_R(r\,|\,S=o,D=d,Y\in B_K)=$$
$$\sum_{k=1}^{K-1}\mu_{\varepsilon,k}(d)\Big(f_R(r\,|\,S=o,D=d,Y\in B_k)-f_R(r\,|\,S=o,D=d,Y\in B_K)\Big)+\text{err}_d(r),$$

where $B_k:=B_\varepsilon(y_k)$ and $\mu_{\varepsilon,k}(d)$ denotes the $k$th component of $\mu_\varepsilon(d)$. Dividing by $f_R(r)$ and applying Bayes' rule gives

$$\mathbb{E}\Big[\widetilde{\Delta}^e(d)-\widetilde{\Delta}_\varepsilon^{o\top}(d)\mu_\varepsilon(d)\Big|R=r\Big]=\frac{\text{err}_d(r)}{f_R(r)}.$$

Since $\inf_r f_R(r)\geq\underline{l}$ and $\sup_r|\text{err}_d(r)|\leq\omega(\varepsilon)$, we obtain

$$\Big|\mathbb{E}\Big\{\widetilde{\Delta}^e(d)-\widetilde{\Delta}_\varepsilon^{o\top}(d)\mu_\varepsilon(d)\;\Big|\;R=r\Big\}\Big|\leq\frac{\omega(\varepsilon)}{\underline{l}},$$

as claimed.

## D.6 Estimation with Covariates

### D.6.1 Proof of Proposition C.5

The proof follows the same argument as Proposition 3 applied within each covariate cell $x\in\mathcal{X}$.

### D.6.2 Proof of Proposition C.6

Define $Y(x)=\Delta^e(x)$, $W(x)=\Delta^o(x)$, $U(x)=Y(x)-\Delta^o(x)^\top\theta(x)$, and $Z(x)=H(x,R)$. Define $p_d(x)=\Pr(D=d,S=e\,|\,X=x)$ for $d\in\{0,1\}$, $q_j(x)=\Pr(Y=y_j,S=o\,|\,X=x)$ for $j\in\{1,\ldots,K\}$, and $\pi(x)=\Pr(X=x\,|\,S=e)$. Finally, let $n_x=\sum_{i=1}^n 1\{X_i=x\}$ and define the within-$x$ cell average $\mathbb{E}_{n,x}[\cdot]=n_x^{-1}\sum_{i:X_i=x}(\cdot)$.

Define $\widehat{\Delta}^e(x)=\frac{1\{D=1,S=e\}}{\mathbb{E}_n[1\{D=1,S=e\}|X=x]}-\frac{1\{D=0,S=e\}}{\mathbb{E}_n[1\{D=0,S=e|X=x\}]}$. Analogously define $\widehat{\Delta}_j^o(x)=\frac{1\{Y=y_j,S=o\}}{\mathbb{E}_n[1\{Y=y_j,S=o\}|X=x]}-\frac{1\{Y=y_j,S=o\}}{\mathbb{E}_n[1\{Y=y_j,S=o\}|X=x]}$ and $\widehat{\Delta}^o(x)=\Big(\widehat{\Delta}_1^o(x),\ldots,\widehat{\Delta}_{K-1}^o(x)\Big)^\top$. We then use the lightened notation: $\widehat{Y}(x)=\widehat{\Delta}^e(x)$, $\widehat{W}(x)=\widehat{\Delta}^o(x)$, $\widehat{U}(x)=\widehat{Y}(x)-\widehat{W}(x)^\top\theta(x)$, and $\widehat{Z}(x)=\widehat{H}(x,R)$.

In this lighter notation, the cell-specific estimator $\widehat{\theta}(x)$ can then be written as $\widehat{\theta}(x):=\mathbb{E}_n[\widehat{Z}(x)\widehat{W}(x)^\top]^{-1}\mathbb{E}_n[\widehat{Z}(x)\widehat{Y}(x)]$, and the estimator of the cell-specific ATE in the experimental sample is $\widehat{\tau}(x)=\lambda^\top\widehat{\theta}(x)$ for $\lambda=(y_1-y_K,\ldots,y_{K-1}-y_K)^\top$. The estimator of the ATE in the experimental sample is $\widehat{\tau}=\sum_{x\in\mathcal{X}}\widehat{\pi}(x)\widehat{\tau}(x)$.

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

The proof largely follows the same argument as the Proposition 4 applied within each covariate cell $x \in \mathcal{X}$. For completeness, we state the following two intermediate lemmas that parallel Lemma B.6 and Lemma B.7.

**Lemma D.1.** Under Proposition C.6's conditions, for any $x \in \mathcal{X}$, $\sqrt{n}\mathbb{E}_{n,x}[Z(x)\widehat{U}(x)] \rightsquigarrow N(0, V(x))$, where $V(x) = v^\top(x)\Sigma(x)v(x)$, $\Sigma = \mathrm{Var}(B(x) \mid X = x)$,

$$B(x) = \begin{pmatrix} Z(x)1\{D=1, S=e\} \\ 1\{D=1, S=e\} \\ Z(x)1\{D=0, S=e\} \\ 1\{D=0, S=e\} \\ Z(x)1\{Y=\cdot, S=o\} \\ 1\{Y=\cdot, S=o\} \\ Z(x)1\{Y=y_K, S=o\} \\ 1\{Y=y_k, S=o\} \end{pmatrix}$$

and $h(\cdot)$ defined as in Lemma B.6,and $v(x) = \nabla h(\mathbb{E}[B(x) \mid X = x])$.

*Proof.* The proof follows the same argument as Lemma B.6 applied within each covariate cell $x \in \mathcal{X}$. □

**Lemma D.2.** Under Proposition C.6's conditions, for any $x \in \mathcal{X}$, $\sqrt{n_x}(\widehat{\theta}(x) - \theta(x)) \rightsquigarrow N(0, A(x)V(x)A^\top(x))$ for $A(x) = \mathbb{E}[\widetilde{H}(x,R)\Delta^o(x)^\top \mid X = x]^{-1}$ and $V(x)$ defined in Lemma D.1.

*Proof.* The proof follows the same argument as Lemma B.7 applied within each covariate cell $x \in \mathcal{X}$. □

To prove Proposition C.6, we first observe that Lemma D.1 implies the linear expansion

$$\mathbb{E}_{n,x}[Z(x)\widehat{U}(x)] = h(\mathbb{E}[B(x) \mid X = x]) + v(x)^\top(\mathbb{E}_{n,x}[B(x)] - \mathbb{E}[B(x) \mid X = x]) + o_p(n_x^{-1/2}),$$

where $h(\mathbb{E}[B(x) \mid X = x]) = \mathbb{E}[Z(x)U(x) \mid X = x] = 0$ by Corollary 1. Consequently,

$$\mathbb{E}_{n,x}[Z(x)\widehat{U}(x)] = \frac{1}{n_x}\sum_{i:X_i=x} \phi_i(x) + o_p(n_x^{-1/2}),$$

where $\phi_i(x) = v(x)^\top (B(x) - \mathbb{E}[B(x) \mid X = x])$. Observe that $\mathbb{E}[\phi_o(x) \mid X = x] = 0$ and define $V(x) = \mathrm{Var}(\phi_i(x) \mid X = x)$. Lemma D.2 implies that $\widehat{\theta}(x) - \theta(x) = A(x)\mathbb{E}_{n,x}[Z(x)\widehat{U}(x)] + o_p(n_x^{-1/2})$. Consequently, we have the linear representation

$$\widehat{\theta}(x) - \theta(x) = \frac{1}{n_x} \sum_{i:X_i = x} A(x)\phi_i(x) + o_p(n_x^{-1/2}),$$

and since $\widehat{\tau}(x) = \lambda^\top \widehat{\theta}(x)$, $\tau(x) = \lambda^\top \theta(x)$, it further follows that

$$\widehat{\tau}(x) - \tau(x) = \frac{1}{n_x} \sum_{i:X_i = x} \lambda^\top A(x)\phi_i(x) + o_p(n_x^{-1/2}).$$

We also have that $o_p(n_x^{-1/2}) = o_p(n^{-1/2})$ uniformly over $x$ by assumption.

Next, we expand the marginal weight estimator $\widehat{\pi}(x) = \frac{\widehat{p}_{e,x}}{\widehat{p}_e}$ for $\widehat{p}_e := \frac{1}{n}\sum_{i=1}^n 1\{S_i = e\}$ and $\widehat{p}_{e,x} = \frac{1}{n}\sum_{i=1}^n 1\{S_i = e, X_i = x\}$. Since these probabilities and their estimators are bounded away from zero, the map $(a,b) \to a/b$ is smooth at $(p_{e,x}, p_e)$. Therefore, a standard Delta-method argument yields

$$\sqrt{n}(\widehat{\pi}(x) - \pi(x)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i(x) + o_p(1)$$

for $\xi_i(x) = \frac{1}{p_e}(1\{S_i = e, X_i = x\} - \pi(x) 1\{S_i = e\})$.

We next decompose the difference

$$\widehat{\tau} - \tau = \sum_{x \in \mathcal{X}} \widehat{\pi}(x)\widehat{\tau}(x) - \sum_{x \in \mathcal{X}} \pi(x)\tau(x) =$$

$$\underbrace{\sum_{x \in \mathcal{X}} \pi(x)(\widehat{\tau}(x) - \tau(x))}_{(I)} + \underbrace{\sum_{x \in \mathcal{X}} (\widehat{\pi}(x) - \pi(x))\tau(x)}_{(II)} + \underbrace{\sum_{x \in \mathcal{X}} (\widehat{\pi}(x) - \pi(x))(\widehat{\tau}(x) - \tau(x))}_{(III)}.$$

We then analyze each term.

Consider (III). Recall that $\max_{x \in \mathcal{X}} |\widehat{\pi}(x) - \pi(x)| = O_p(n^{-1/2})$ and $\max_{x \in \mathcal{X}} |\widehat{\tau}(x) - \tau(x)| = O_p(n^{-1/2})$. Therefore, (III) is bounded by

$$\left| \sum_{x \in \mathcal{X}} (\widehat{\pi}(x) - \pi(x))(\widehat{\tau}(x) - \tau(x)) \right| \leq |\mathcal{X}| \cdot \max_{x \in \mathcal{X}} |\widehat{\pi}(x) - \pi(x)| \cdot \max_{x \in \mathcal{X}} |\widehat{\tau}(x) - \tau(x)| = O_p(n^{-1}).$$

It therefore follows that $\sqrt{n}(III) = o_p(1)$.

Consider (I). Substituting our linearization of $\widehat{\tau}(x) - \tau(x)$, we have

$$\sum_{x \in \mathcal{X}} \pi(x)(\widehat{\tau}(x) - \tau(x)) = \sum_{x \in \mathcal{X}} \pi(x) \left( \frac{1}{n_x} \sum_{i:X_i = x} \lambda^\top A(x)\phi_i(x) \right) + o_p(n^{-1/2}).$$

Multiplying by $\sqrt{n}$, we then have that

$$\sqrt{n}(I) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{x \in \mathcal{X}} \pi(x) \frac{n}{n_x} 1\{X_i = x\} \lambda^\top A(x)\phi_i(x) + o_p(1).$$

Since $n_x/n \xrightarrow{p} \Pr(X_i = x)$ for all $x$, and so

$$\sqrt{n}(I) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{x \in \mathcal{X}} \frac{\pi(x)}{\Pr(X_i = x)} 1\{X_i = x\} \lambda^\top A(x) \phi_i(x) + o_p(1).$$

Finally, consider (II). Substituting our linearization for $\widehat{\pi}(x) - \pi(x)$, we have that

$$\sqrt{n} \sum_{x \in \mathcal{X}} (\widehat{\pi}(x) - \pi(x)) \tau(x) = \sum_{x \in \mathcal{X}} \tau(x) \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i(x) \right) + o_p(n^{-1/2})$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{x \in \mathcal{X}} \tau(x) \xi_i(x) + o_p(1).$$

We now combine these together to arrive at the linearization

$$\sqrt{n}(\widehat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i^{unknown} + o_p(1),$$

where

$$\psi_i^{unknown} = \sum_{x \in \mathcal{X}} \frac{\pi(x)}{\Pr(X = x)} 1\{X_i = x\} \lambda^\top A(x) \phi_i(x) + \sum_{x \in \mathcal{X}} \tau(x) \xi_i(x).$$

We define $V_\tau^{unknown} = \mathrm{Var}(\psi_i^{unknown})$. We observe that $\mathbb{E}[\psi_i^{unknown}] = 0$ since $\mathbb{E}[\phi_i(x) \mid X_i = x] = 0$ and $\mathbb{E}[\xi_i(x)] = 0$. We further observe that $\mathbb{E}[(\psi_i^{unknown})^2] < \infty$ under our assumptions. The result then follows by applying the Lindeberg-Levy central limit theorem. □

## D.7 Additional Lemmas

**Lemma D.3.** Suppose Theorem 1's conditions and Assumption C.12 hold. The following holds:

$$(\Delta^e(X) - \Delta^o(X)^\intercal \theta(X))^2 =$$

$$\frac{1\{D=1, S=e\}}{\Pr(D=1, S=e \mid X=x)^2} + \frac{1\{D=0, S=e\}}{\Pr(D=0, S=e \mid X=x)^2} +$$

$$\sum_{j=1}^{K-1} \theta_j^2(X) \left( \frac{1\{Y=y_j, S=o\}}{\Pr(Y=y_j, S=o \mid X=x)^2} + \frac{1\{Y=y_K, S=o\}}{\Pr(Y=y_K, S=o \mid X=x)^2} \right) +$$

$$\sum_{j \neq k} \theta_j(X) \theta_k(X) \frac{1\{Y=y_K, S=o\}}{\Pr(Y=y_K, S=o \mid X=x)^2}.$$

*Proof.* First observe

$$(\Delta^e(X) - \Delta^o(X)^\intercal \theta(X))^2 = (\Delta^e(X))^2 - 2(\Delta^o(X)^\intercal \theta(X)) \Delta^e(X) + (\Delta^o(X)^\intercal \theta(X))^2$$

$$= (\Delta^e(X))^2 + (\Delta^o(X)^\intercal \theta(X))^2,$$

where the second equality follows since the events in $\Delta^e(X)$ are exclusive of the events in $\Delta^o(X)$. Further observe that

$$(\Delta^e(X))^2 = \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e\,|\,X=x)^2} + \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e\,|\,X=x)^2}, \text{ and}$$

$$\begin{aligned}(\Delta^o(X)^\intercal\theta(X))^2 &= \left(\sum_{j=1}^{K-1}\theta_j(X)\Delta_j^o(X)\right)^2 \\ &= \sum_{j=1}^{K-1}\theta_j^2(X)\big(\Delta_j^o(X)\big)^2 + \sum_{j\neq k}\theta_j(X)\theta_k(X)\Delta_j^o(X)\Delta_k^o(X),\end{aligned}$$

where

$$\begin{aligned}\big(\Delta_j^o(X)\big)^2 &= \left(\frac{1\{Y=y_j,S=o\}}{\Pr(Y=y_j,S=o\,|\,X=x)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o\,|\,X=x)}\right)^2 \\ &= \frac{1\{Y=y_j,S=o\}}{\Pr(Y=y_j,S=o\,|\,X=x)^2} + \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o\,|\,X=x)^2}\end{aligned}$$

and

$$\Delta_j^o(X)\Delta_k^o(X) =$$
$$\left(\frac{1\{Y=y_j,S=o\}}{\Pr(Y=y_j,S=o)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o\,|\,X=x)}\right)\left(\frac{1\{Y=y_k,S=o\}}{\Pr(Y=y_k,S=o\,|\,X=x)} - \frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o\,|\,X=x)}\right) =$$
$$\frac{1\{Y=y_K,S=o\}}{\Pr(Y=y_K,S=o\,|\,X=x)^2}.$$

The result follows by combining these expressions. □

**Lemma D.4** (Bias under direct effects and stability violations)**.** Let $\mathcal{X}=\emptyset,\mathcal{Y}=\{0,1\}$. Let $\Delta^e := \frac{1\{D=1,S=e\}}{\Pr(D=1,S=e)} - \frac{1\{D=0,S=e\}}{\Pr(D=0,S=e)}, \Delta^o = \frac{1\{Y=1,S=o\}}{\Pr(Y=1,S=o)} - \frac{1\{Y=0,S=o\}}{\Pr(Y=0,S=o)}$. Suppose Assumption 1 holds and $\Pr(D=0\,|\,S=o)=1$. Fix $H$ with $\mathbb{E}[\Delta^o H(R)]\neq 0$. Define, for $y\in\{0,1\}$,

$$\delta_y(H) \equiv \mathbb{E}[H(R)\,|\,S=e,Y=y,D=1] - \mathbb{E}[H(R)\,|\,S=e,Y=y,D=0],$$

and define the (control-arm) stability violation on the $H$-scale

$$\gamma_y(H) \equiv \mathbb{E}[H(R)\,|\,S=e,Y=y,D=0] - \mathbb{E}[H(R)\,|\,S=o,Y=y,D=0].$$

Then

$$\mathbb{E}[\Delta^e H(R)] = \theta\Big(\mathbb{E}[\Delta^o H(R)] + \gamma_1(H) - \gamma_0(H)\Big) + (1-\mu(1))\delta_0(H) + \mu(1)\delta_1(H), \tag{D.4}$$

where $\mu(d)=\mathbb{E}[Y(d)\,|\,S=e]$ and $\theta=\mu(1)-\mu(0)$. Equivalently,

$$\theta = \frac{\mathbb{E}[\Delta^e H(R)] - \{(1-\mu(1))\delta_0(H) + \mu(1)\delta_1(H)\}}{\mathbb{E}[\Delta^o H(R)] + \gamma_1(H) - \gamma_0(H)}.$$

*Proof.* Because $\Pr(D=0 \mid S=o)=1$, $\mathbb{E}[\Delta^o H(R)] = \mathbb{E}[H(R) \mid S=o, Y=1, D=0] - \mathbb{E}[H(R) \mid S=o, Y=0, D=0]$. By randomization in $S=e$ (Assumption 1),

$$\mathbb{E}[\Delta^e H(R)] = \mathbb{E}[H(R) \mid S=e, D=1] - \mathbb{E}[H(R) \mid S=e, D=0].$$

Expand the experimental difference by conditioning on $Y$:

$$\mathbb{E}[H(R) \mid S=e, D=1] - \mathbb{E}[H(R) \mid S=e, D=0] = \theta\Big(\mathbb{E}[H(R) \mid S=e, Y=1, D=0] - \mathbb{E}[H(R) \mid S=e, Y=0, D=0]$$

where we used $\mathbb{E}[H(R) \mid S=e, Y=y, D=1] = \mathbb{E}[H(R) \mid S=e, Y=y, D=0] + \delta_y(H)$ and $\theta = \mu(1) - \mu(0)$. Finally, add and subtract the observational control-arm means:

$$\mathbb{E}[H(R) \mid S=e, Y=y, D=0] = \mathbb{E}[H(R) \mid S=o, Y=y, D=0] + \gamma_y(H),$$

so the bracketed difference equals $\mathbb{E}[\Delta^o H(R)] + \gamma_1(H) - \gamma_0(H)$. Substituting yields (D.4). □

# E Additional Estimation Algorithms

**Algorithm E.1** Estimation with discrete outcomes, discrete covariates, and incomplete observational cases: details

**Require:** Data $\{S_i, 1\{S_i=e\}D_i, 1\{S_i=o\}Y_i, X_i, R_i\}_{i=1}^n$ with $Y \in \mathcal{Y}=\{y_1,\dots,y_K\}$ and $X \in \mathcal{X}=\{x_1,\dots,x_{|\mathcal{X}|}\}$.

**Ensure:** $\widehat{\tau}$ and a confidence interval.

Randomly split indices into a training fold TRAIN and a test fold TEST.

**for all** $x \in \mathcal{X}$ **do**

Compute $\mathbb{E}_{\text{TRAIN}}(1\{D_i=d, S_i=e\} \mid X_i=x)$ for $d \in \{0,1\}$ and $\mathbb{E}_{\text{TRAIN}}(1\{Y_i=y_j, S_i=o\} \mid X_i=x)$ for $j=1,\dots,K$.

Train $\text{PRED}_Y(\cdot,x): \mathcal{R} \to [0,1]^K$ on $\{(R_i,Y_i): i \in \text{TRAIN}, X_i=x, S_i=o\}$ to estimate $\Pr(Y=\cdot \mid S=o, X=x, R)$.

Train $\text{PRED}_D(\cdot,x): \mathcal{R} \to [0,1]$ on $\{(R_i,D_i): i \in \text{TRAIN}, X_i=x, S_i=e\}$ to estimate $\Pr(D=1 \mid S=e, X=x, R)$.

Train $\text{PRED}_S(\cdot,x): \mathcal{R} \to [0,1]$ on $\{(R_i,S_i): i \in \text{TRAIN}, X_i=x\}$ to estimate $\Pr(S=e \mid X=x, R)$.

Define, for generic $(R,x)$,

$$\widehat{\mathbb{E}}\{\Delta^e(x) \mid R,x\} = \left\{ \frac{\text{PRED}_D(R,x)}{\mathbb{E}_{\text{TRAIN}}(1\{D_i=1, S_i=e\} \mid X_i=x)} - \frac{1-\text{PRED}_D(R,x)}{\mathbb{E}_{\text{TRAIN}}(1\{D_i=0, S_i=e\} \mid X_i=x)} \right\} \text{PRED}_S(R,x),$$

and for each $j=1,\dots,K-1$,

$$\widehat{\mathbb{E}}\{\Delta_j^o(x) \mid R,x\} = \left\{ \frac{\text{PRED}_{Y,j}(R,x)}{\mathbb{E}_{\text{TRAIN}}(1\{Y_i=y_j, S_i=o\} \mid X_i=x)} - \frac{\text{PRED}_{Y,K}(R,x)}{\mathbb{E}_{\text{TRAIN}}(1\{Y_i=y_K, S_i=o\} \mid X_i=x)} \right\} \{1-\text{PRED}_S(R,x)\}.$$

Stack $\widehat{\mathbb{E}}\{\Delta^o(x) \mid R,x\} = (\widehat{\mathbb{E}}\{\Delta_1^o(x) \mid R,x\}, \dots, \widehat{\mathbb{E}}\{\Delta_{K-1}^o(x) \mid R,x\})^\top$.

Compute

$$\widehat{\theta}_{\text{INIT}}(x) = \arg\min_{\theta \in \mathbb{R}^{K-1}} \sum_{i \in \text{TRAIN}: X_i=x} \left[ \widehat{\mathbb{E}}\{\Delta^e(x) \mid R_i,x\} - \widehat{\mathbb{E}}\{\Delta^o(x) \mid R_i,x\}^\top \theta \right]^2.$$

Compute $\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}}, R, x)$ using Lemma D.3 (plugging in the fitted probabilities/predictors above), and set

$$\widehat{H}(R,x) = \frac{\widehat{\mathbb{E}}\{\Delta^o(x) \mid R,x\}}{\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}}, R, x)} \in \mathbb{R}^{K-1}.$$

Compute $\mathbb{E}_{\text{TEST}}(1\{D_i=d, S_i=e\} \mid X_i=x)$ for $d \in \{0,1\}$ and $\mathbb{E}_{\text{TEST}}(1\{Y_i=y_j, S_i=o\} \mid X_i=x)$ for $j=1,\dots,K$.

**for all** $i \in \text{TEST}$ with $X_i=x$ **do**

Set

$$\widehat{\Delta}_i^e(x) = \frac{1\{D_i=1, S_i=e\}}{\mathbb{E}_{\text{TEST}}(1\{D_i=1, S_i=e\} \mid X_i=x)} - \frac{1\{D_i=0, S_i=e\}}{\mathbb{E}_{\text{TEST}}(1\{D_i=0, S_i=e\} \mid X_i=x)}.$$

**for** $j=1,\dots,K-1$ **do**

Set

$$\widehat{\Delta}_{i,j}^o(x) = \frac{1\{Y_i=y_j, S_i=o\}}{\mathbb{E}_{\text{TEST}}(1\{Y_i=y_j, S_i=o\} \mid X_i=x)} - \frac{1\{Y_i=y_K, S_i=o\}}{\mathbb{E}_{\text{TEST}}(1\{Y_i=y_K, S_i=o\} \mid X_i=x)}.$$

**end for**

Stack $\widehat{\Delta}_i^o(x) = (\widehat{\Delta}_{i,1}^o(x), \dots, \widehat{\Delta}_{i,K-1}^o(x))^\top$.

**end for**

Compute

$$\widehat{\theta}(x) = \left\{ \sum_{i \in \text{TEST}: X_i=x} \widehat{H}(R_i,x) \widehat{\Delta}_i^o(x)^\top \right\}^{-1} \left\{ \sum_{i \in \text{TEST}: X_i=x} \widehat{H}(R_i,x) \widehat{\Delta}_i^e(x) \right\}.$$

**end for**

Compute $\widehat{\theta} = \mathbb{E}_{\text{TEST}}\{\widehat{\theta}(X) \mid S=e\}$ (empirical mean over $i \in \text{TEST}$ with $S_i=e$).

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

Bootstrap the standard error $\widehat{v}$ of $\widehat{\tau}$ *holding* $\widehat{H}$ *fixed.*

Return the confidence interval $\widehat{\tau} \pm z_{1-\alpha/2} \widehat{v}/\sqrt{n}$.

**Algorithm E.2** Estimation with discrete outcomes, discrete covariates, and complete observational cases

**Require:** Data $\{S_i,D_i,1\{S_i=o\}Y_i,X_i,R_i\}_{i=1}^n$ 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.

Split indices into training fold TRAIN and test fold TEST.

**for all** $x\in\mathcal{X}$ **do**

  **for all** $d\in\{0,1\}$ **do**

    Estimate on TRAIN the cell probabilities $\widehat{p}_d^e(x)=\mathbb{E}_{\text{TRAIN}}(1\{D_i=d,S_i=e\}\,|\,X_i=x)$ and $\widehat{p}_{d,j}^o(x)=\mathbb{E}_{\text{TRAIN}}(1\{Y_i=y_j,D_i=d,S_i=o\}\,|\,X_i=x)$ for $j=1,...,K$.

    Fit nuisance models on TRAIN: $\text{PRED}_{Y,d}(\cdot,x)\approx\Pr(Y=\cdot,D=d\,|\,S=o,X=x,R)$, $\text{PRED}_D(\cdot,x)\approx\Pr(D=1\,|\,S=e,X=x,R)$, $\text{PRED}_S(\cdot,x)\approx\Pr(S=e\,|\,X=x,R)$. Define $\text{PRED}_{D=1}(R,x)=\text{PRED}_D(R,x)$ and $\text{PRED}_{D=0}(R,x)=1-\text{PRED}_D(R,x)$.

    Form the plug-in conditional moments (for generic $(R,x)$):

$$\widehat{\mathbb{E}}\{\widetilde{\Delta}^e(d,x)\,|\,R,x\}=\frac{\text{PRED}_{D=d}(R,x)\text{PRED}_S(R,x)}{\widehat{p}_d^e(x)}-\frac{\text{PRED}_{Y,d,K}(R,x)\{1-\text{PRED}_S(R,x)\}}{\widehat{p}_{d,K}^o(x)},$$

$$\widehat{\mathbb{E}}\{\widetilde{\Delta}_j^o(d,x)\,|\,R,x\}=\left\{\frac{\text{PRED}_{Y,d,j}(R,x)}{\widehat{p}_{d,j}^o(x)}-\frac{\text{PRED}_{Y,d,K}(R,x)}{\widehat{p}_{d,K}^o(x)}\right\}\{1-\text{PRED}_S(R,x)\},\quad j=1,...,K-1,$$

    and stack $\widehat{\mathbb{E}}\{\widetilde{\Delta}^o(d,x)\,|\,R,x\}=(\widehat{\mathbb{E}}\{\widetilde{\Delta}_1^o(d,x)\,|\,R,x\},...,\widehat{\mathbb{E}}\{\widetilde{\Delta}_{K-1}^o(d,x)\,|\,R,x\})^\top$.

    Compute $\widehat{\mu}_{\text{INIT}}(d,x)$ by least squares regression of $\widehat{\mathbb{E}}\{\widetilde{\Delta}^e(d,x)\,|\,R_i,x\}$ on $\widehat{\mathbb{E}}\{\widetilde{\Delta}^o(d,x)\,|\,R_i,x\}$ over $i\in\text{TRAIN}$ with $X_i=x$.

    Compute, for generic $(R,x)$,

$$\begin{aligned}\widehat{\sigma}_d^2(\widehat{\mu}_{\text{INIT}},R,x)=&\frac{\text{PRED}_{D=d}(R,x)\text{PRED}_S(R,x)}{\widehat{p}_d^e(x)^2}+\frac{\text{PRED}_{Y,d,K}(R,x)\{1-\text{PRED}_S(R,x)\}}{\widehat{p}_{d,K}^o(x)^2}\\&+\sum_{j=1}^{K-1}\widehat{\mu}_{j,\text{INIT}}(d,x)^2\left(\frac{\text{PRED}_{Y,d,j}(R,x)}{\widehat{p}_{d,j}^o(x)^2}+\frac{\text{PRED}_{Y,d,K}(R,x)}{\widehat{p}_{d,K}^o(x)^2}\right)\{1-\text{PRED}_S(R,x)\}\\&+\sum_{\substack{j,k=1\\j\neq k}}^{K-1}\widehat{\mu}_{j,\text{INIT}}(d,x)\widehat{\mu}_{k,\text{INIT}}(d,x)\frac{\text{PRED}_{Y,d,K}(R,x)}{\widehat{p}_{d,K}^o(x)^2}\{1-\text{PRED}_S(R,x)\}.\end{aligned}$$

    Set $\widehat{H}_d(R,x)=\widehat{\mathbb{E}}\{\widetilde{\Delta}^o(d,x)\,|\,R,x\}/\widehat{\sigma}_d^2(\widehat{\mu}_{\text{INIT}},R,x)\in\mathbb{R}^{K-1}$.

  **end for**

**end for**

**for all** $x\in\mathcal{X}$ **do**

  **for all** $d\in\{0,1\}$ **do**

    On TEST, compute $\widehat{p}_d^e(x)=\mathbb{E}_{\text{TEST}}(1\{D_i=d,S_i=e\}\,|\,X_i=x)$ and $\widehat{p}_{d,j}^o(x)=\mathbb{E}_{\text{TEST}}(1\{Y_i=y_j,D_i=d,S_i=o\}\,|\,X_i=x)$ for $j=1,...,K$.

    For each $i\in\text{TEST}$ with $X_i=x$, form the sample analogues:

$$\widehat{\widetilde{\Delta}}_i^e(d,x)=\frac{1\{D_i=d,S_i=e\}}{\widehat{p}_d^e(x)}-\frac{1\{Y_i=y_K,D_i=d,S_i=o\}}{\widehat{p}_{d,K}^o(x)},$$

$$\widehat{\widetilde{\Delta}}_{i,j}^o(d,x)=\frac{1\{Y_i=y_j,D_i=d,S_i=o\}}{\widehat{p}_{d,j}^o(x)}-\frac{1\{Y_i=y_K,D_i=d,S_i=o\}}{\widehat{p}_{d,K}^o(x)},\quad j=1,...,K-1,$$

    and stack $\widehat{\widetilde{\Delta}}_i^o(d,x)=(\widehat{\widetilde{\Delta}}_{i,1}^o(d,x),...,\widehat{\widetilde{\Delta}}_{i,K-1}^o(d,x))^\top$.

    Compute $\widehat{\mu}(d,x)=\left\{\sum_{i\in\text{TEST}:X_i=x}\widehat{H}_d(R_i,x)\widehat{\widetilde{\Delta}}_i^o(d,x)^\top\right\}^{-1}\left\{\sum_{i\in\text{TEST}:X_i=x}\widehat{H}_d(R_i,x)\widehat{\widetilde{\Delta}}_i^e(d,x)\right\}$.

  **end for**

**end for**

For $d\in\{0,1\}$, set $\widehat{\mu}(d)=\mathbb{E}_{\text{TEST}}\{\widehat{\mu}(d,X)\,|\,S=e\}$ (empirical mean over $i\in\text{TEST}$ with $S_i=e$).

Compute arm means $\widehat{m}(d)=\sum_{j=1}^{K-1}y_j\widehat{\mu}_j(d)+y_K\left(1-\sum_{j=1}^{K-1}\widehat{\mu}_j(d)\right)$.

Set $\widehat{\tau}=\widehat{m}(1)-\widehat{m}(0)$.

Bootstrap the standard error of $\widehat{\tau}$ on TEST, holding $\widehat{H}_0,\widehat{H}_1$ fixed, and report a confidence interval.

**Algorithm E.3** Inference with discrete outcomes, continuous covariates, and incomplete observational cases: details

**Require:** Data $\{S_i, 1\{S_i=e\}D_i, 1\{S_i=o\}Y_i, X_i, R_i\}_{i=1}^n$ with $Y \in \mathcal{Y}=\{y_1,...,y_K\}$ and $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$.

**Require:** Basis $p_J : \mathcal{X} \to \mathbb{R}^J$ (set $P_i = p_J(X_i)$) to approximate a function $g(x)$, an initial representation $H_{\text{INIT}}(X,R) \in \mathbb{R}^{K-1}$, and the maps $\psi(\cdot)$ and $\lambda$ such that $\theta(x) = \psi(g(x))$ and $\tau(x) = \lambda^\top \psi(g(x))$.

**Notation.** For any representation $H(X,R) \in \mathbb{R}^{K-1}$ (write $H_i = H(X_i,R_i)$), define $m = K(K+2)$ and the stacked primitive vector $B_i(H) \in \mathbb{R}^m$ by stacking

$$B_i(H) = \Big( H_i 1\{D_i=1,S_i=e\}, H_i 1\{D_i=0,S_i=e\}, 1\{D_i=1,S_i=e\}, 1\{D_i=0,S_i=e\},$$
$$H_i 1\{Y_i=y_1,S_i=o\},...,H_i 1\{Y_i=y_K,S_i=o\}, 1\{Y_i=y_1,S_i=o\},...,1\{Y_i=y_K,S_i=o\} \Big).$$

For a fold $\mathcal{I} \in \{\text{TRAIN},\text{TEST}\}$ and a choice of $H$, define the series-OLS estimator

$$\widehat{\beta}_{\mathcal{I}}(H) = \arg\min_{\beta \in \mathbb{R}^{J \times m}} \frac{1}{n_{\mathcal{I}}} \sum_{i \in \mathcal{I}} \|B_i(H) - P_i^\top \beta\|^2, \qquad \widehat{g}_{\mathcal{I}}(x;H) = p_J(x)^\top \widehat{\beta}_{\mathcal{I}}(H).$$

1: Split indices into training fold TRAIN and test fold TEST.

**Training fold: learn $\widehat{H}(R,X)$.**

2: Fit nuisance predictors on TRAIN: $\text{PRED}_Y(R,X) \approx \Pr(Y=\cdot \mid S=o,X,R)$ using $\{(R_i,X_i,Y_i) : i \in \text{TRAIN}, S_i=o\}$, $\text{PRED}_D(R,X) \approx \Pr(D=1 \mid S=e,X,R)$ using $\{(R_i,X_i,D_i) : i \in \text{TRAIN}, S_i=e\}$, $\text{PRED}_S(R,X) \approx \Pr(S=e \mid X,R)$ using $\{(R_i,X_i,S_i) : i \in \text{TRAIN}\}$.

3: Compute $\widehat{\beta}_{\text{TRAIN}}(H_{\text{INIT}})$ from regressing $B_i(H_{\text{INIT}})$ on $P_i$ over $i \in \text{TRAIN}$, and set $\widehat{g}_{\text{TRAIN}}(x) = \widehat{g}_{\text{TRAIN}}(x;H_{\text{INIT}})$.

4: Set $\widehat{\theta}_{\text{INIT}}(X) = \psi(\widehat{g}_{\text{TRAIN}}(X))$.

5: Let $\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{D=d,S=e\} \mid X)$ and $\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{Y=y_j,S=o\} \mid X)$ denote the corresponding components of $\widehat{g}_{\text{TRAIN}}(X)$ (for $d \in \{0,1\}$ and $j=1,...,K$).

6: Form the plug-in conditional expectations (for generic $(R,X)$):

$$\widehat{\mathbb{E}}\{\Delta^e(X) \mid R,X\} = \left\{ \frac{\text{PRED}_D(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{D=1,S=e\} \mid X)} - \frac{1-\text{PRED}_D(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{D=0,S=e\} \mid X)} \right\} \text{PRED}_S(R,X),$$

$$\widehat{\mathbb{E}}\{\Delta_j^o(X) \mid R,X\} = \left\{ \frac{\text{PRED}_{Y,j}(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{Y=y_j,S=o\} \mid X)} - \frac{\text{PRED}_{Y,K}(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{Y=y_K,S=o\} \mid X)} \right\} \{1-\text{PRED}_S(R,X)\}, \quad j=1,...,K-1,$$

and stack $\widehat{\mathbb{E}}\{\Delta^o(X) \mid R,X\} = (\widehat{\mathbb{E}}\{\Delta_1^o(X) \mid R,X\},...,\widehat{\mathbb{E}}\{\Delta_{K-1}^o(X) \mid R,X\})^\top$.

7: Compute, for generic $(R,X)$,

$$\widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R,X) = \left( \frac{\text{PRED}_D(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{D=1,S=e\} \mid X)^2} + \frac{1-\text{PRED}_D(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{D=0,S=e\} \mid X)^2} \right) \text{PRED}_S(R,X)$$
$$+ \sum_{j=1}^{K-1} \widehat{\theta}_{j,\text{INIT}}(X)^2 \left( \frac{\text{PRED}_{Y,j}(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{Y=y_j,S=o\} \mid X)^2} + \frac{\text{PRED}_{Y,K}(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{Y=y_K,S=o\} \mid X)^2} \right) \{1-\text{PRED}_S(R,X)\}$$
$$+ \sum_{\substack{j,k=1 \\ j \neq k}}^{K-1} \widehat{\theta}_{j,\text{INIT}}(X) \widehat{\theta}_{k,\text{INIT}}(X) \frac{\text{PRED}_{Y,K}(R,X)}{\widehat{\mathbb{E}}_{\text{TRAIN}}(1\{Y=y_K,S=o\} \mid X)^2} \{1-\text{PRED}_S(R,X)\}.$$

8: Set $\widehat{H}(R,X) = \widehat{\mathbb{E}}\{\Delta^o(X) \mid R,X\} / \widehat{\sigma}^2(\widehat{\theta}_{\text{INIT}},R,X) \in \mathbb{R}^{K-1}$.

**Test fold: estimate $\widehat{\theta}(x)$ and $\widehat{\tau}(x)$.**

9: Compute $\widehat{\beta}_{\text{TEST}}(\widehat{H})$ from regressing $B_i(\widehat{H})$ on $P_i$ over $i \in \text{TEST}$, and set $\widehat{g}_{\text{TEST}}(x) = \widehat{g}_{\text{TEST}}(x;\widehat{H})$.

10: Set $\widehat{\theta}(x) = \psi(\widehat{g}_{\text{TEST}}(x))$ and $\widehat{\tau}(x) = \lambda^\top \psi(\widehat{g}_{\text{TEST}}(x))$.

**Aggregate and inference.**

11: Compute $\widehat{\tau} = \mathbb{E}_{\text{TEST}}[\widehat{\tau}(X) \mid S=e]$ (empirical mean over $i \in \text{TEST}$ with $S_i=e$).

12: Bootstrap the standard error of $\widehat{\tau}$ on TEST, holding $\widehat{H}$ fixed, and report $\widehat{\tau} \pm z_{1-\alpha/2} \widehat{v}/\sqrt{n}$.

# F Additional Details on Empirical Settings

## F.1 Forest Cover Details

This appendix provides additional details on the Uganda forest cover setting used in the calibrated simulations in Section 6 of the main text. 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. 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 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, 2–5 km, and 5–10 km); Appendix Figure F.1 provides a map of these regions. In the main text, we focus on the 0–2 km band.

**Construction of the remotely sensed variable:** For each $0.01^\circ$ 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 distinct from the Landsat imagery underlying the Hansen et al. forest-cover measurements. We merge these embeddings to the cell-level forest-cover outcome by geographic coordinates. Altogether, the merged Uganda-wide dataset contains approximately 7 million $0.01^\circ$ grid cells, each with the constructed forest-cover outcome and MOSAIKS features.

We then train a random forest to predict $Y$ from the MOSAIKS embeddings using grid

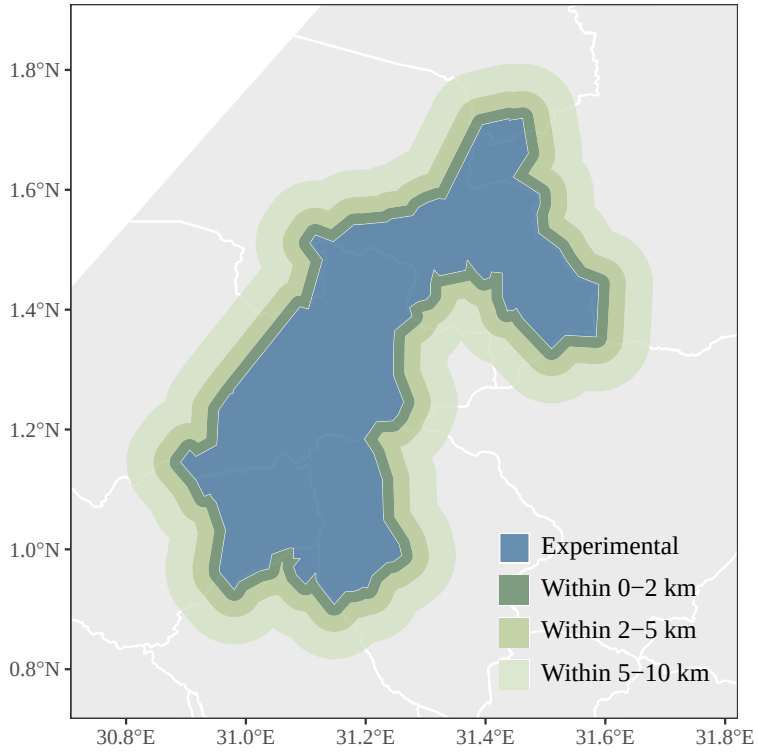



Figure F.1: We illustrate the experimental and observational samples used in our analysis of the Uganda forest cover data in Section 6. The experimental sample is the grid cells covered by the payments for ecosystem services experiment in Jayachandran et al. (2017). The observational sample consists of grid cells in surrounding geographic bands. In the main text, we focus on the 0-2 kilometer geographic band.

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.

## 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;

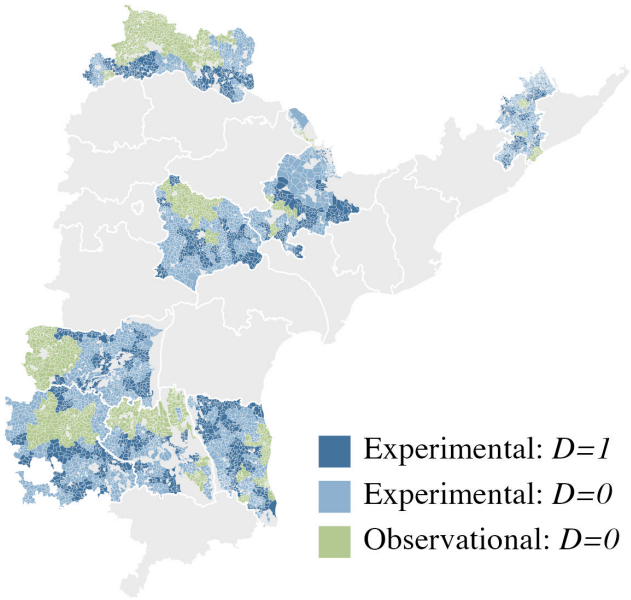



Figure F.2: We illustrate the experimental and observational samples in the Smartcards evaluation (Muralidharan et al., 2023).

- 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 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. The causal parameter is the effect of early adoption (2010 Smartcards) for villages in the experiment.

Appendix Table F.1 summarizes village characteristics. We study 8,320 villages, with an average population exceeding 2,000 individuals per village. Villages are located in rural areas. The populated area within a typical village is geographically concentrated.

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. The consumption outcome variables 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 is 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

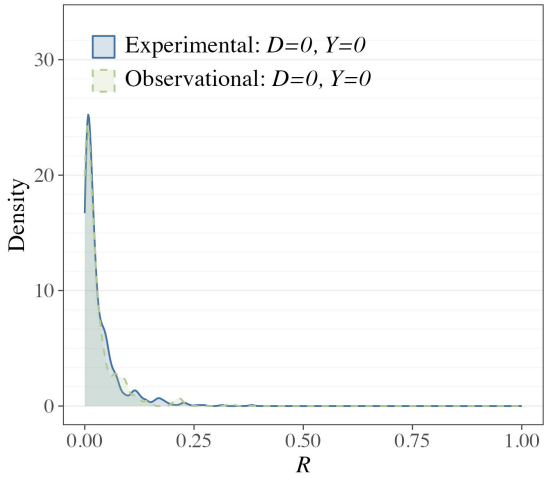
Next, we assess whether stability (Assumption 2) is plausible in each empirical setting. Here, stability requires that four conditional distributions of the remotely sensed variable should be the same across samples: $f_R(r \mid S=e,D=d,Y=y) = f_R(r \mid S=o,D=d,Y=y)$, for each $(d,y)$ stratum. In both settings, we observe outcomes for untreated units in the experimental and observational samples.[24] 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.

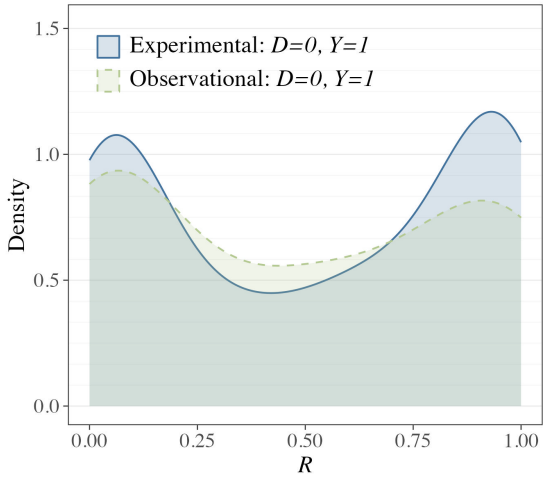
To complement the visual evidence, we conduct Kolmogorov-Smirnov tests for equality of

[24] 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

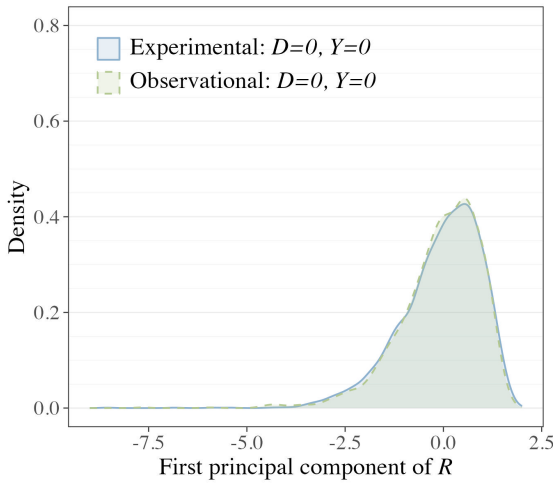



(a) Densities of $R \mid S, D=0, Y=0$.

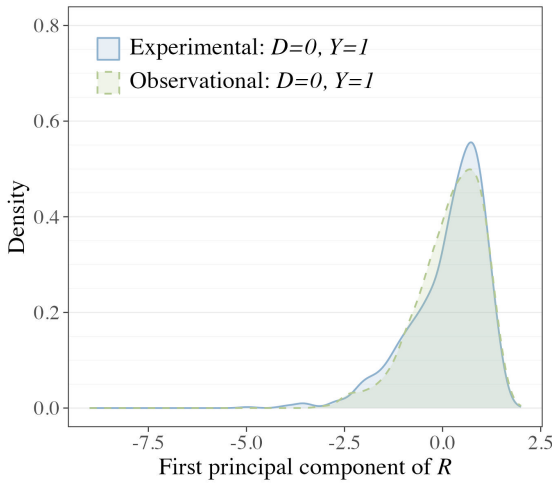



(b) Densities of $R \mid S, D=0, Y=1$.

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.

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.

# G Additional Monte Carlo Simulation Results

## G.1 Incomplete Observational Cases

In Section 6 of the main text, we reported semi-synthetic simulations for the setting with incomplete observational cases in which no experimental outcomes are collected. We provide additional implementation details and results for these simulations.

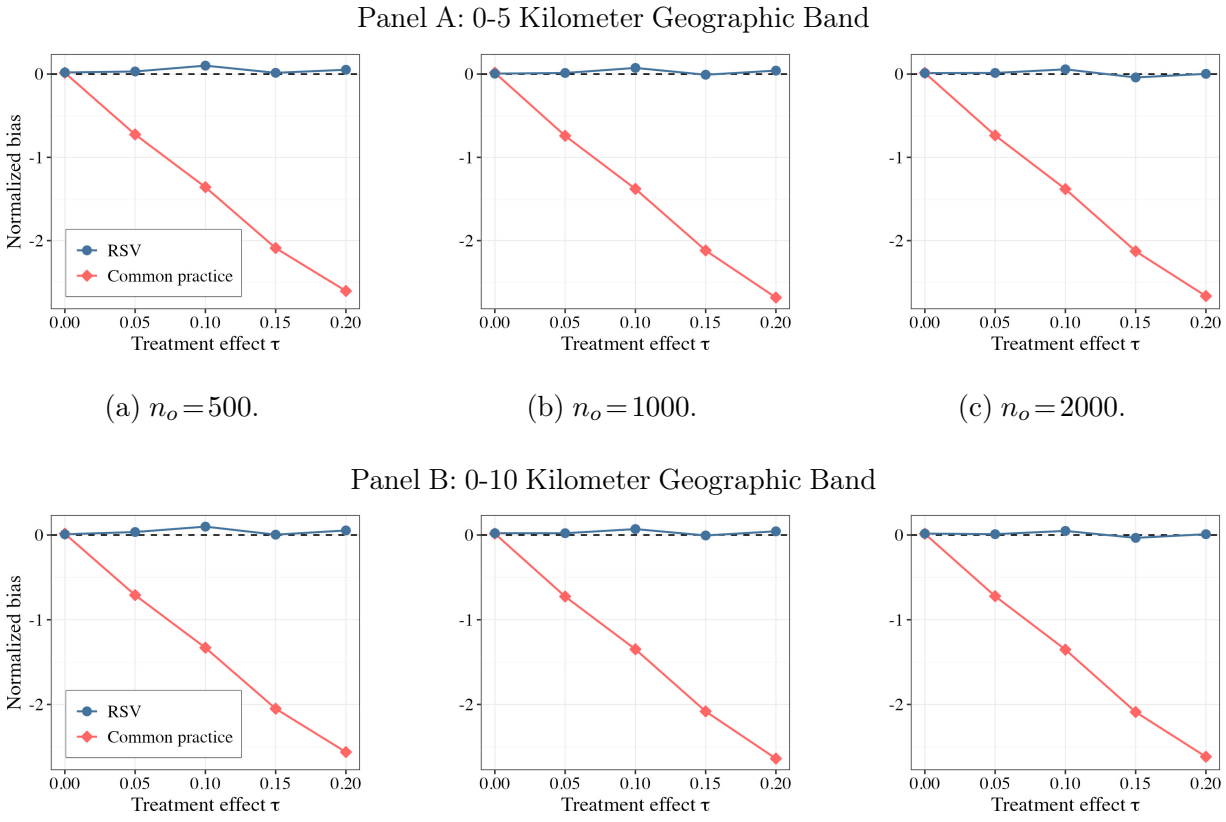
**Additional Simulation Details:** Let $\alpha_e$ denote the baseline outcome probability $P(Y(0) = 1 \mid S = e)$ in the experimental sample and let $\alpha_o$ denote the baseline outcome probability $P(Y(0) = 1 \mid S = o)$ in the observational sample. For experimental sample size $n_e$, observational sample size $n_o$, and treatment effect $\tau$, we perform $B = 500$ replications, following the simulation design described in the main text.

We compare two estimators. The two-step estimator corresponding to common practice trains a predictor of $Y$ given $R$ in the observational sample and regresses predicted outcomes on treatment in the experimental sample. We estimate the prediction of $Y$ given $R$ in the observational sample using logistic regression and random forests in the forest cover and Smartcards setting respectively. For the estimator proposed in Section 4, we use 2-fold cross-fitting. In the forest cover setting, we estimate the representation using logistic regression to fit $\widehat{p}_Y(\cdot \mid r)$, $\widehat{p}_D(r)$, and $\widehat{p}_S(r)$. In the Smartcards setting, we estimate these predictions using random forests. For tractability, we truncate the remotely sensed from $\mathbb{R}^{4050}$ to $\mathbb{R}^{1050}$ for computational tractability. In particular, we use the initial 1,000 features of the pre-trained MOSAIKS embedding and the nighttime luminosity measures. To estimate the representation, we use the `R` package `randomForest` with 100 trees and the package default options to fit $\widehat{p}_Y(\cdot \mid r)$, $\widehat{p}_D(r)$, and $\widehat{p}_S(r)$. When predicting the outcomes, we set the class weights ten-to-one because one outcome value is much more frequent than the other.

We calibrate the baseline outcome probabilities $\alpha_o$, $\alpha_e$ empirically In the forest cover setting, $\alpha_e$ is chosen to match the average outcome $Y$ among grid cells in the experimental sample and $\alpha_o$ is chosen to match the average outcome among grid cells in the observational sample (i.e., the two kilometer geographic band surrounding the experimental sample). In the Smartcards setting, $\alpha_e$ and $\alpha_o$ are chosen to match the average outcome $Y$ among villages in the experimental and observational samples, respectively.

**Additional Simulation Results Varying the Baseline Outcome Probabilities and Sample Sizes:** We now present additional simulation results that compare our estimator

and the two-step estimator for alternative choices of the baseline outcome probabilities and the sample size. In the forest cover setting, Appendix Figure G.1 reports the normalized bias and Appendix Figure G.2 reports the coverage of each estimator as we vary the choice of $\alpha_o$ (calibrating its value using the 0-5 kilometer and 0-10 kilometer geographic band around the experimental sample) and the choice of $n_o = 1000,2000$. In the Smartcard setting, Figure G.3 reports the normalized bias and coverage of each estimator as we vary the choice of $n_o = 500,1500$. For these alternative parameter choices, we find the same results as in the main text.

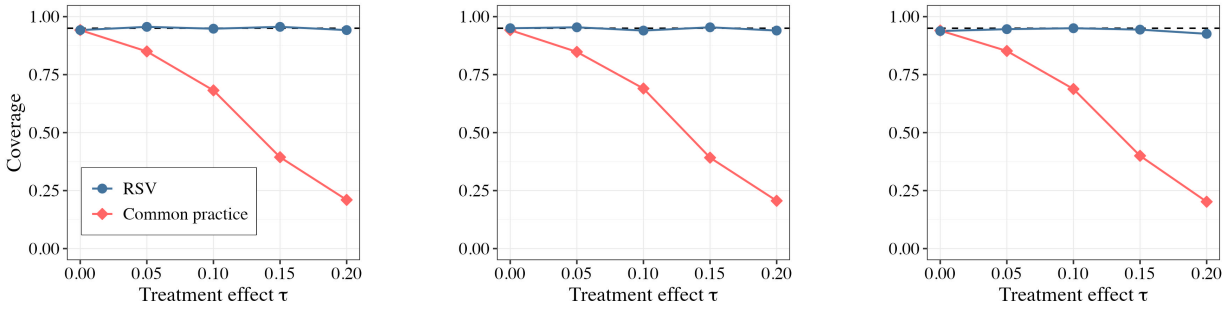



(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 2000$.

(d) $n_o = 500$. (e) $n_o = 1000$. (f) $n_o = 2000$.

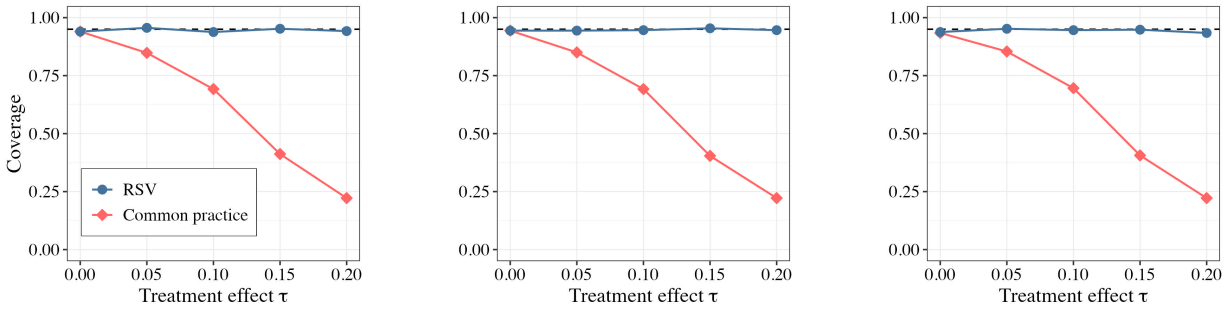
Figure G.1: Comparison of the normalized bias of the commonly-used, two-step estimator and our estimator in the forest cover setting. Each panel varies the choice of the observational sample used to calibrate $\alpha_o$. For each value of the synthetic treatment effect $\tau$, observational outcome probability $\alpha_o$, and observational sample size $n_o \in \{500,1000,2000\}$, we conduct 500 simulations.

Panel A: 0-5 Kilometer Geographic Band



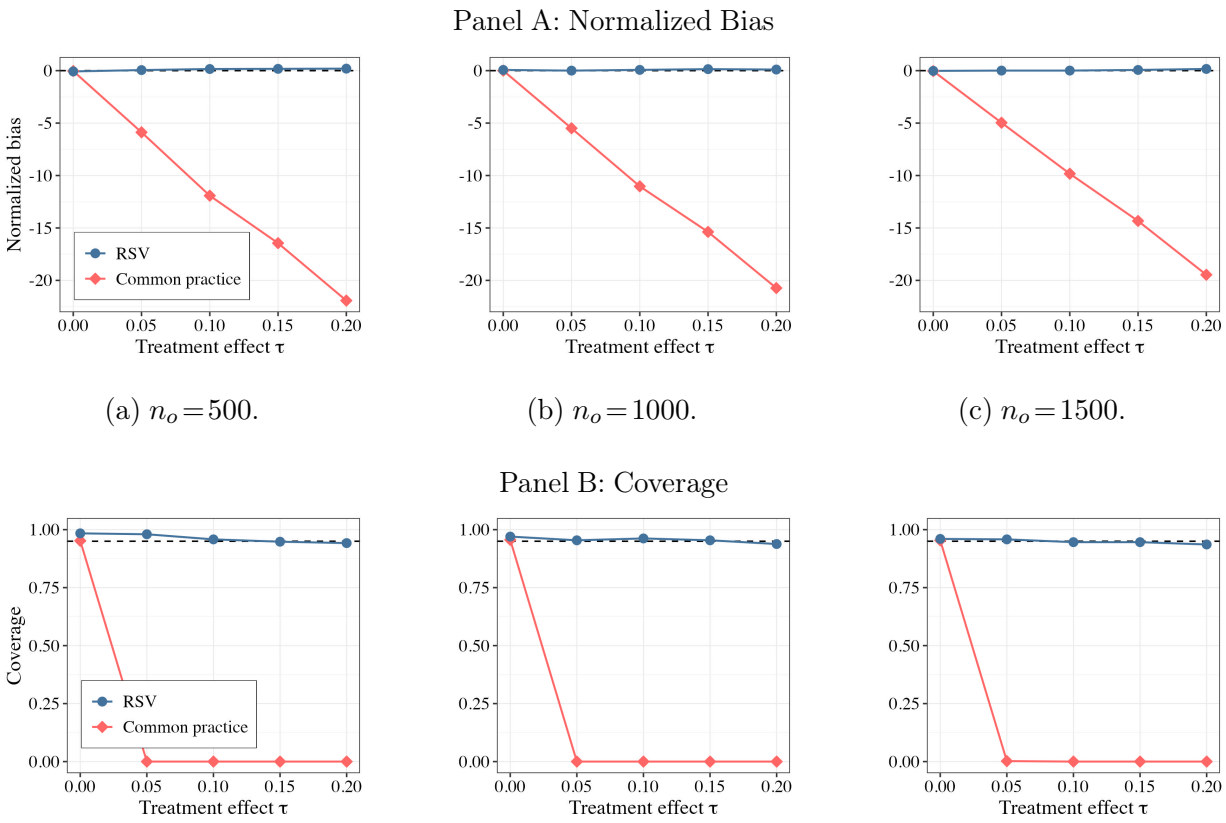

(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 2000$.

Panel B: 0-10 Kilometer Geographic Band




(d) $n_o = 500$. (e) $n_o = 1000$. (f) $n_o = 2000$.

Figure G.2: Comparison of the coverage of the commonly-used, two-step estimator and our estimator in the forest cover setting. Each panel varies the choice of the observational sample used to calibrate $\alpha_o$. For each value of the synthetic treatment effect $\tau$, observational outcome probability $\alpha_o$, and observational sample size $n_o \in \{500,1000,2000\}$, we conduct 500 simulations.




(a) $n_o = 500$. (b) $n_o = 1000$. (c) $n_o = 1500$.

(d) $n_o = 500$. (e) $n_o = 1000$. (f) $n_o = 1500$.

Figure G.3: Comparison of the normalized bias and coverage of the commonly-used, two-step estimator and our estimator in the Smartcards setting, varying the observational sample size $n_o \in \{500, 1000, 1500\}$. For each value of the synthetic treatment effect $\tau$ and observational sample size $n_o$, we conduct 500 simulations.

## G.2 Incomplete Observational Cases with Validation Sample

In Section 6, we reported semi-synthetic simulations for the setting with incomplete observational cases in which no experimental outcomes are collected. Prediction-powered inference methods were not feasible in this setting as there are no observations with jointly observed $(Y,D,R)$ with variation in $D$.

To compare our estimator against PPI methods, we report simulations with incomplete observational cases (Assumption 3(ii)) in which some experimental outcomes are observed, as discussed in Section 5.1. As discussed in Section 3.2.2, PPI methods correct the errors of generic machine learning output, and so we focus on the Ugandan forest cover setting in which the remotely sensed variable $R$ is the scalar output of the pretrained random forest.

**Simulation Details When Some Experimental Outcomes are Observed:** The simulation design follows the same data-generating process as described in Section 6, except we now retain outcomes for a random subset of experimental units.

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

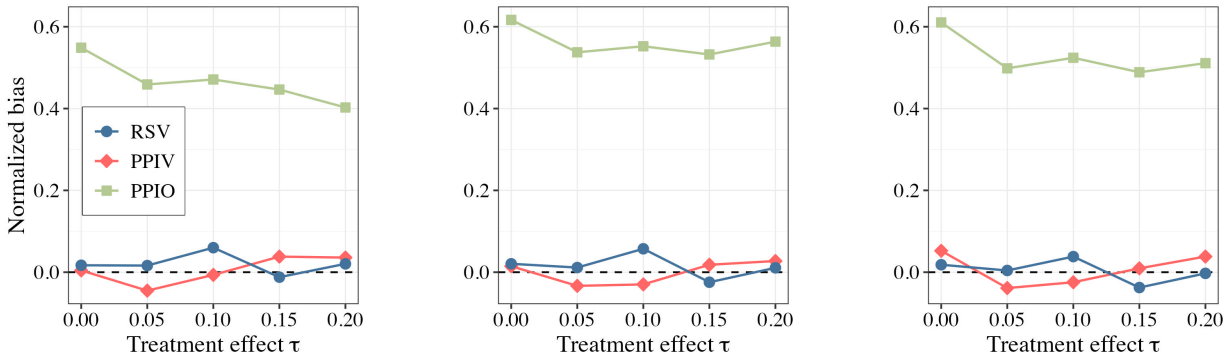
As before, for the estimator proposed in Section 4, we use 2-fold cross-fitting. In the forest cover setting, we estimate the representation using logistic regression to fit $\widehat{p}_Y(\cdot \mid r)$, $\widehat{p}_D(r)$, and $\widehat{p}_S(r)$. This data environment now enables comparison with PPI methods (Angelopoulos et al., 2023) since there is now a subset of units for which $(Y,D,R)$ is observed and $D$ varies. We consider consider two PPI estimators: a PPI estimator using only the $n_{eo}$ validation units (PPIV); and the PPI estimator using both the $n_{eo}$ validation units and the $n_o$ observational units (PPIO). We also consider a benchmark regression that uses only the $n_{eo}$ experimental units with observed outcomes, estimating the treatment effect by regressing $Y_i$ on $D_i$ without using the remotely sensed variable $R_i$ or the observational sample.

**Simulation Results:** Two findings emerge. First, Appendix Figure G.4 compares the normalized bias of our estimator, a PPI estimator using only the validation units, and a PPI estimator using both the validation units and the observational units (PPIO). Since $\alpha_e \neq \alpha_o$ across simulation configurations, the PPI estimator that uses both the validation sample and

the observational sample in the rectifiers is biased, confirming Proposition 2. By contrast, both our estimator and a PPI estimator using only the validation units are approximately unbiased.

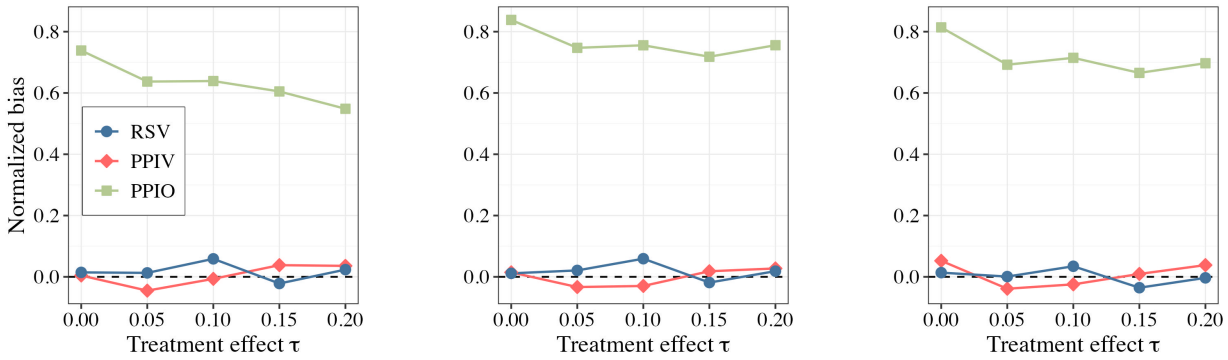
Second, while approximately unbiased, the PPI estimator that only uses the validation sample in the rectifiers (disregarding the observational sample) is unbiased but inefficient. Our estimator yields standard errors that can be up to 40% smaller. Additionally, our estimator yields standard errors that are up to 60% smaller than the benchmark regression using only the validation units. In the presence of a validation sample, 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.

Panel A: 0-2 Kilometer Geographic Band



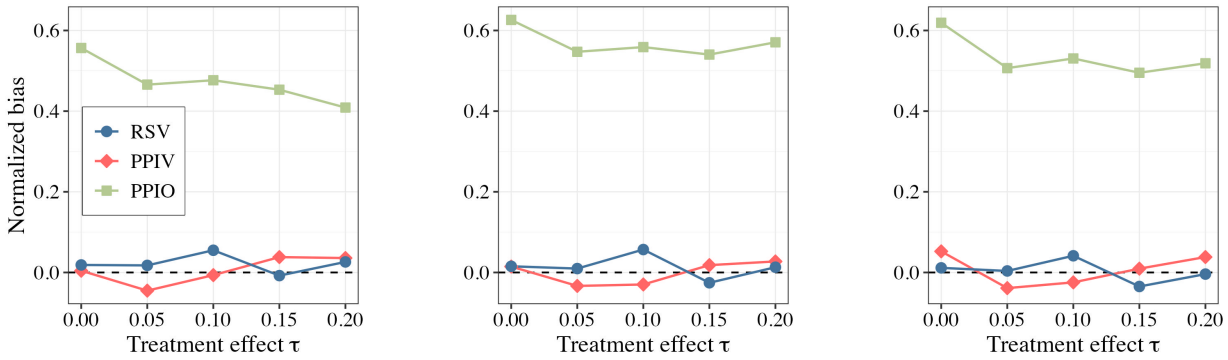

(a) $n_{eo} = 100$. (b) $n_{eo} = 250$. (c) $n_{eo} = 500$.

Panel B: 0-5 Kilometer Geographic Band




(d) $n_{eo} = 100$. (e) $n_{eo} = 250$. (f) $n_{eo} = 500$.

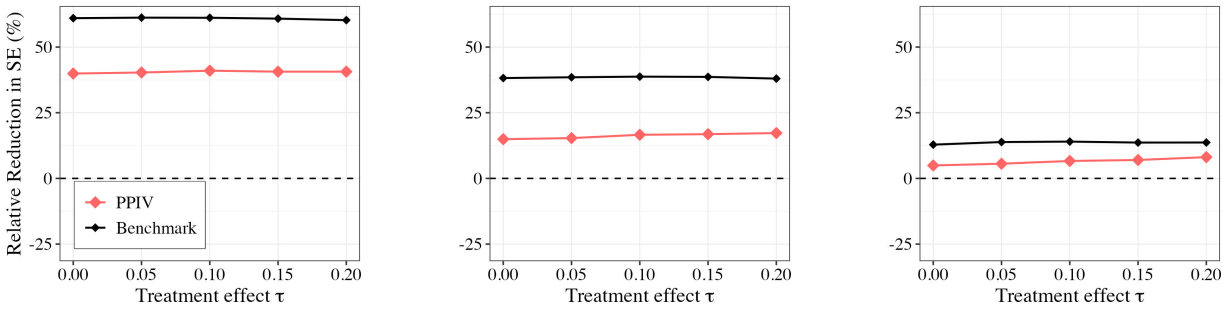
Panel C: 0-10 Kilometer Geographic Band




(g) $n_{eo} = 100$. (h) $n_{eo} = 250$. (i) $n_{eo} = 500$.

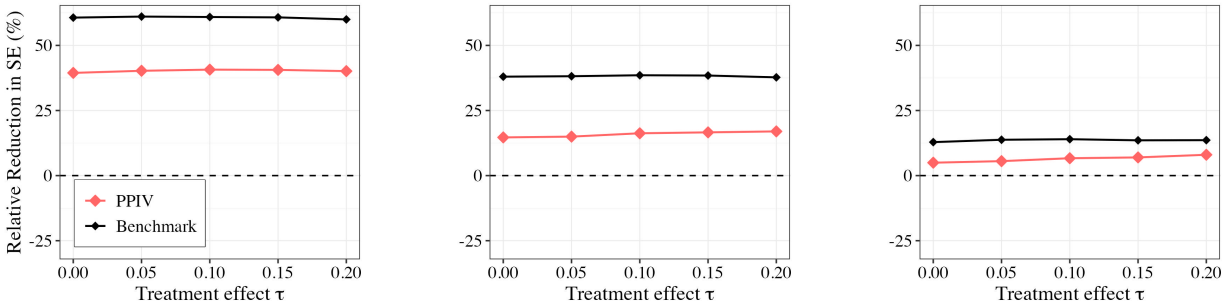
Figure G.4: Comparison of the normalized bias of our estimator, PPI using only the validation units (PPIV), and PPI using both the validation units and the obsevational units (PPIO). Each panel varies the choice of observational sample used to calibrate $\alpha_o$. For each value of the synthetic treatment effect $\tau$ and validation sample size $n_{eo} \in \{100,250,500\}$, we conduct 500 simulations.

Panel A: 0-2 Kilometer Geographic Band



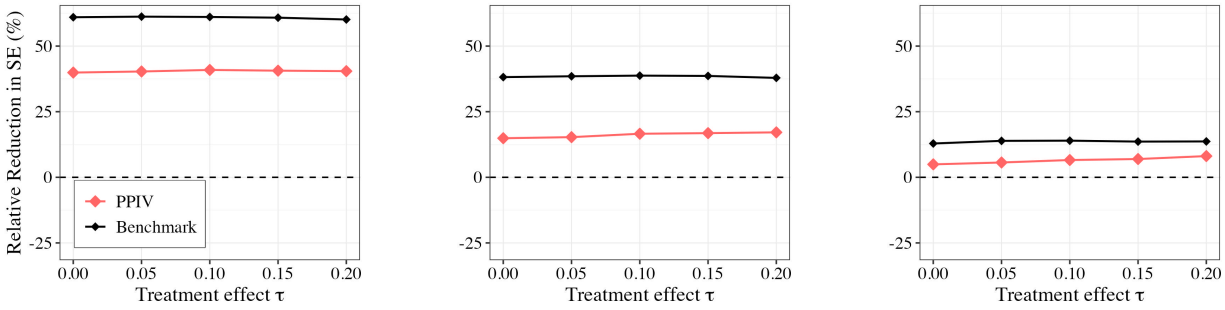

(a) $n_{eo} = 100$. (b) $n_{eo} = 250$. (c) $n_{eo} = 500$.

Panel B: 0-5 Kilometer Geographic Band




(d) $n_{eo} = 100$. (e) $n_{eo} = 250$. (f) $n_{eo} = 500$.

Panel C: 0-10 Kilometer Geographic Band




(g) $n_{eo} = 100$. (h) $n_{eo} = 250$. (i) $n_{eo} = 500$.

Figure G.5: Percentage reduction in standard errors of our estimator relative to PPI using only the validation units (PPIV) and the benchmark regression using only the validation units. Each panel varies the choice of observational sample used to calibrate $\alpha_o$. For each value of the synthetic treatment effect $\tau$ and validation sample size $n_{eo} \in \{100,250,500\}$, we conduct 500 simulations.

# H Additional Empirical Results

## H.1 Smartcards Experiment 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. As mentioned in the main text, we analyze this experiment through the lens of a non-randomly selected validation sample (Section 5.1). The experimental sample consists of treated and untreated villages, the validation sample consists of buffer villages, and the observational sample consists of non-study villages. Consequently, we analyze the experiment with outcomes on only 155 mandals out of 291 mandals (i.e, approximately 50% fewer mandals).

We implement our estimator using Random Forests to estimate the nuisance functions that enter the representation. 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. (2020, Theorem 2). Alternatively, we could use the limited complexity of random forests, along the lines of Chernozhukov et al. (2020, Theorem 3).

For completeness, Appendix Table H.1 reports the point estimates, standard errors and confidence intervals associated with Figure ?? in the main text. We also report the point estimate, standard error and a 90% confidence interval for the difference in the benchmark estimate and our estimator for each poverty outcome.

**Assessing stability and no direct effects:** In the main text, we assessed whether stability and no-direct-effects are plausible for the outcome whether the village has only low- and middle-income households. We now report the same analyses for the other two outcomes: (i) whether the village is in the bottom quartile of per capita consumption, and (ii) whether the village has only low-income households.

First, 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. In both settings, we observe outcomes for untreated units in the experimental and observational samples. Appendix Figure H.1 directly evaluates the conditions for $d=0$ and $y \in \{0,1\}$ by subsetting and visualizing densities for both of these outcomes. As in the main text, to complement the visual evidence, we conduct Kolmogorov-Smirnov tests for equality of the conditional distributions across samples. For the low income outcome, we

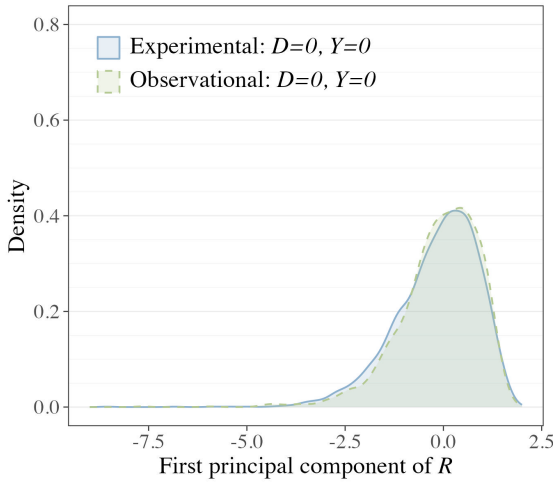
| | Consumption | Low income | Middle income |
|---|---|---|---|
| | (1) | (2) | (3) |
| RSV Estimate | -0.0525 | -0.0261 | -0.0450 |
| | (0.0436) | (0.0220) | (0.0360) |
| Benchmark Estimate | -0.0746 | -0.0235 | -0.0397 |
| | (0.0445) | (0.0148) | (0.0274) |
| Difference between Estimates | -0.0222 | 0.0026 | 0.0053 |
| | (0.0271) | (0.0142) | (0.0246) |

Table H.1: Our method recovers the unbiased benchmark estimate. The benchmark is the difference-in-means an economist would obtain if they could observe treatments and outcomes in the experiment. Our method uses treatments and RSVs in an experiment, and outcomes and RSVs in an observational sample. For each poverty outcome, we report the benchmark estimate and its 90% confidence interval and our estimate and its 90% confidence interval. We also report the difference between the benchmark and our estimate, its standard error and its 90% confidence interval. "Consumption" refers to the outcome whether the village is in the bottom quartile of per capita consumption, "Low Income" refer to the outcome whether the village has only low-income households, and "Middle Income" refers to the outcome whether the village has only low- and middle-income households. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

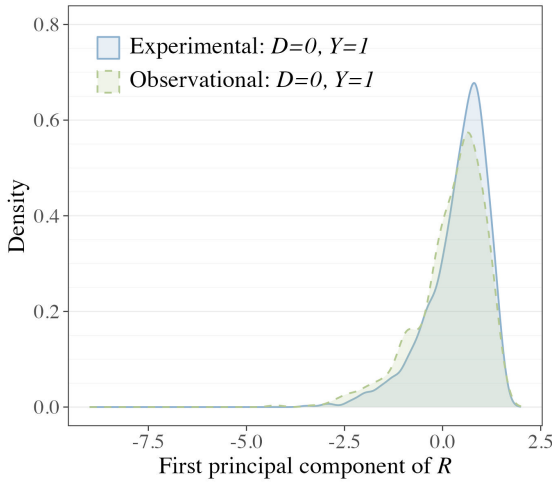
cannot reject the null of equal distributions at the 5% level. For the consumption outcome, we can marginally reject the null of equal distributions at the 5% level for the $Y=0$ case ($p=0.039$) and we can reject at the 1% level for the $Y=1$ case ($p<0.001$).

Second, incomplete observational cases, we require that the treatment affects the RSV only through the outcome (Assumption 3(ii)). As in the main text, since ground-truth outcomes are available in this application, we can directly assess this condition: if no direct effects holds, the distributions of $R \mid S=e, Y=y, D=0$ and $R \mid S=e, Y=y, D=1$ in the experimental sample should coincide. Appendix Figure **??** visualizes this comparison by plotting the density of the standardized first principal component of $R$, conditional on each value of $Y$ and $D$. The distributions overlap closely. Kolmogorov–Smirnov tests cannot reject the null of equality

Panel A: Consumption

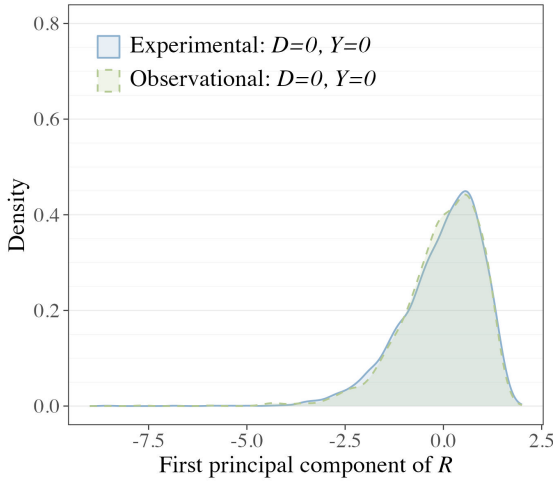



(a) Densities of $R \mid S, D=0, Y=0$.

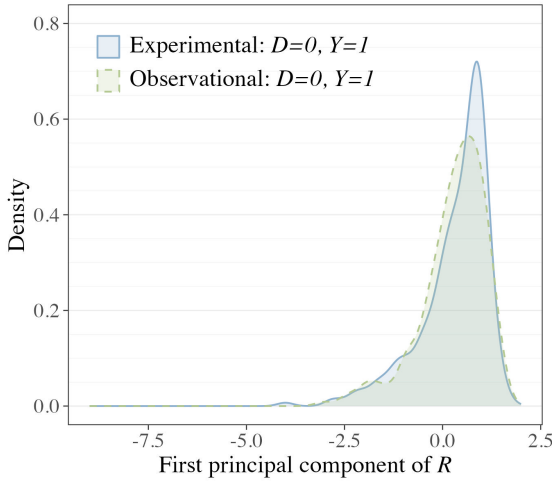



(b) Densities of $R \mid S, D=0, Y=1$.

Panel B: Low Income




(c) Densities of $R \mid S, D=0, Y=0$.




(d) Densities of $R \mid S, D=0, Y=1$.

Figure H.1: Our stability assumption (Assumption 2(i)) is plausible for the two additional outcomes in the Smartcards experiment. 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. "Consumption" refers to the outcome whether the village is in the bottom quartile of per capita consumption, and "Low Income" refer to the outcome whether the village has only low-income households.

at the 5% level for the low income outcome. For the consumption outcome, we can marginally reject the null of equal distributions at the 5% level for the $Y=0$ case ($p=0.048$) and the $Y=1$ case ($p=0.033$).

**Comparing alternative representations:** Appendix Figure H.3 visualizes how our efficient representation $\widehat{H}(R)$ compares to the simple representation $\widehat{P}(Y=1\,|\,R,S=o)$. Even though the RSV is high-dimensional, these representations are scalars. Therefore, we can visualize each village as a point in a plot with $\widehat{H}(R)$ on the vertical axis and $\widehat{P}(Y=1\,|\,R,S=o)$ 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 efficient representation $\widehat{H}(R)$ often takes negative values, while the simple representation $\widehat{P}(Y=1\,|\,R,S=o)$ is bounded in the unit interval $(0,1)$. It is efficient to extrapolate beyond observed villages. Using $\widehat{P}(Y=1\,|\,R,S=o)$ in place of $\widehat{H}(R)$ in Algorithm **??** would only interpolate among observed villages. The representations $\widehat{H}(R)$ and $\textsc{pred}_Y(R)$ are generally correlated, but their relationship is nonlinear. As discussed in the main text, these differences matter for the precision of downstream inference.

**Allowing for 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, Figures **??** and **??** report average global treatment effect estimates in the presence of spillovers.

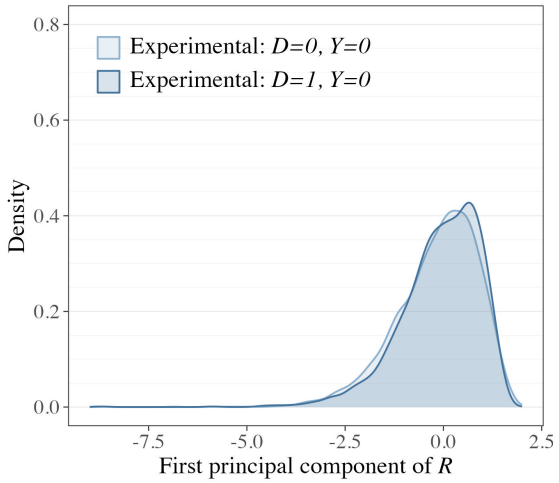
As a robustness check, we now repeat our empirical analysis after subsetting the villages in the manner described above. The subsetting step eliminates about 27% of the total number villages. In particular, about 27% of villages in the full sample have at least one village within a 20 km radius exposed to opposite treatment status. The distance between two villages is measured using their centroids, following Muralidharan et al. (2023).

Under the conditions of Corollary C.4, Figures **??** and **??** report global average treatment effect estimates in the presence of spillovers. We find qualitatively similar results as before. Here the benchmark uses the same subset of villages as our method.

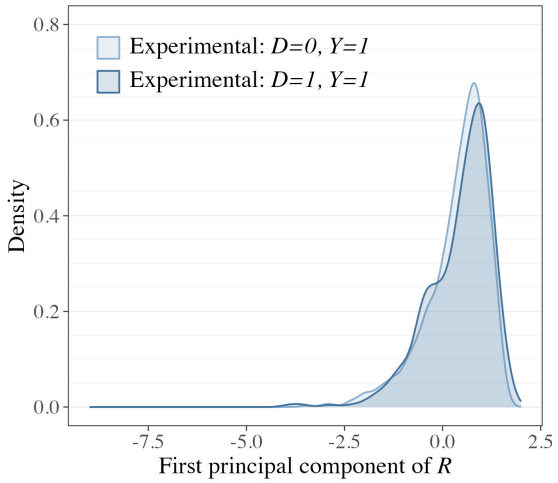
## H.2 Payments for Ecosystem Services Experiment and Crop Burning

We present additional details and results for the empirical application in Section 7.2, based on Jack et al. (2025). In our analysis, we classify a field as experimental ($S=e$) if it did not receive a random spot check. We classify a field as observational ($S=o$) if it received a

Panel A: Consumption

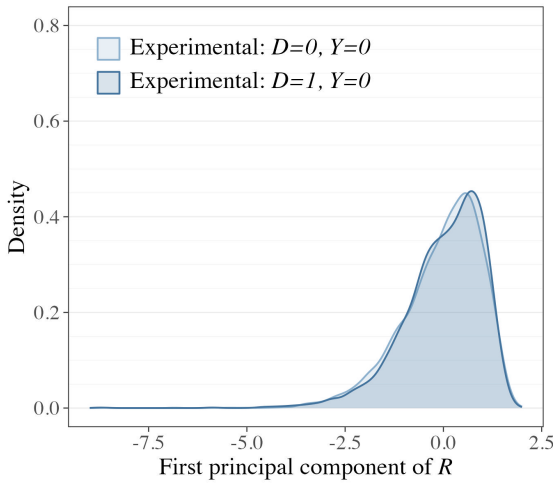



(a) Densities of $R \mid S = e$, $D$, $Y = 0$.

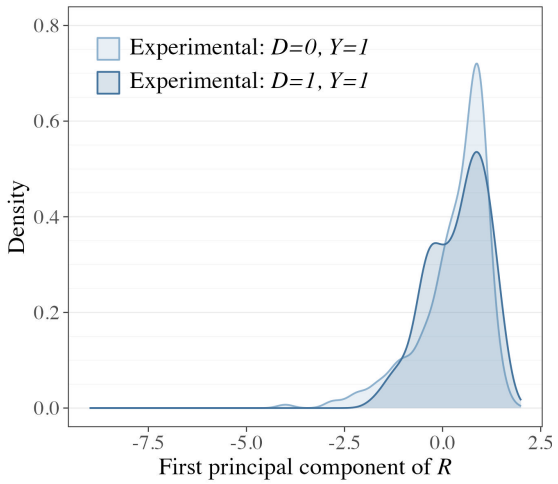



(b) Densities of $R \mid S = e$, $D$, $Y = 1$.

Panel B: Low Income

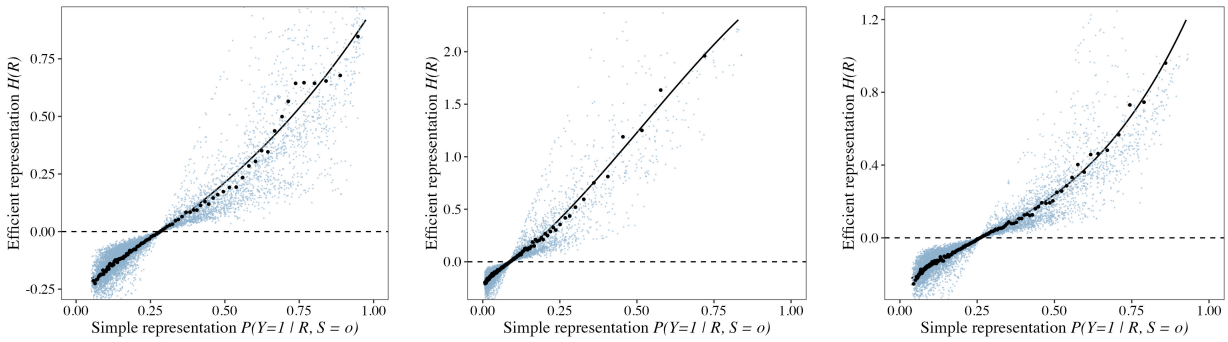



(c) Densities of $R \mid S = e$, $D$, $Y = 0$.




(d) Densities of $R \mid S = e$, $D$, $Y = 1$.

Figure H.2: No direct effects (Assumption 3(ii)) is plausible in the Smartcards experiment. Within each plot, we compare $f_R \mid S = e, D = 0, Y = y)$ against $f_R(R \mid S = e, D = 1, Y = y)$, using data from Muralidharan et al. (2023). Because the RSV $R$ is high-dimensional, we visualize the density of its standardized first principal component. "Consumption" refers to the outcome whether the village is in the bottom quartile of per capita consumption, and "Low Income" refer to the outcome whether the village has only low-income households.

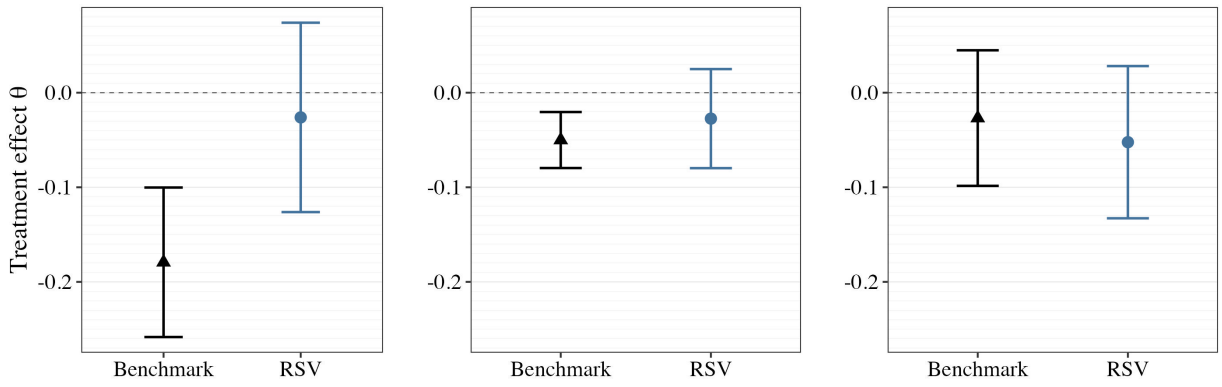



(a) Consumption. (b) Low Income. (c) Middle Income.

Figure H.3: We contrast efficient versus simple representations of the RSV. The efficient representation combines three predictions. The simple representation uses only one prediction. "Consumption" refers to the outcome whether the village is in the bottom quartile of per capita consumption, "Low Income" refer to the outcome whether the village has only low-income households, and "Middle Income" refers to the outcome whether the village has only low- and middle-income households.

| Outcome $Y$ | Consumption | Low Income | Middle Income |
|---|---|---|---|
| | (1) | (2) | (3) |
| Relevance, $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ | 0.2349 | 0.5068 | 0.2145 |
| | (0.0810) | (0.1715) | (0.0708) |
| | [0.1016, 0.3682] | [0.2246, 0.7889] | [0.0981, 0.3309] |

Table H.2: Satellite images are relevant to the poverty outcomes in the presence of spillovers. We report $\mathbb{E}_n\{\widehat{H}(R)\widehat{\Delta}^o\}$ and its 90% confidence interval using our learned representation. The estimates correspond to Corollary C.4. "Consumption" refers to the outcome whether the village is in the bottom quartile of per capita consumption, "Low Income" refer to the outcome whether the village has only low-income households, and "Middle Income" refers to the outcome whether the village has only low- and middle-income households. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.




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

Figure H.4: Our method recovers the unbiased benchmark estimate and its 90% confidence interval in the presence of spillovers. For each poverty outcome, we report the benchmark estimate and its 90% confidence interval and our estimate and its 90% confidence interval. The estimates correspond to Corollary C.4, i.e. using the subsetted sample ("subset"). "Consumption" refers to the outcome whether the village is in the bottom quartile of per capita consumption, "Low Income" refer to the outcome whether the village has only low-income households, and "Middle Income" refers to the outcome whether the village has only low- and middle-income households. Bootstrap standard errors, based on 1000 replications, are clustered at the sub-district level.

random spot check. We analyze this setting with incomplete observational cases (Assumption 3(ii)). Even though the treatment varies in the observational sample, there are no units with $Y=1, D=0$ in the observational sample, and so $\Delta^o(0)$ is undefined.

In Section 7.2, we reported the plug-in estimate, obtained by regressing the authors' "maximum accuracy" classifier for whether a field has not been burned as the outcome (Column 1 of Table 4). Jack et al. (2025) construct another "balanced accuracy" classifier based on an alternative threshold rule. We find similar results in Appendix Table H.3, now using the "balanced accuracy" classifier as the outcome as well as the continuous random forest prediction as the outcome.

| Outcome: | Maximum Accuracy | Balanced Accuracy | Random Forest Predictions |
|---|---|---|---|
| | (1) | (2) | (3) |
| Estimate | 0.075 | 0.067 | 0.030 |
| | (0.038) | (0.046) | (0.017) |

Table H.3: We obtain similar results based on the plug-in estimator that uses the authors' "maximum accuracy" classifier, "balanced accuracy" classifer, and the continuous random forest predictions. We separately regress each of these outcomes on the treatment indicator, controlling for stratum fixed effects. Bootstrap standard errors in parentheses, based on 5000 replications, are clustered at the village level.