Debiasing Machine Learning Predictions for Causal Inference Without Additional Ground Truth Data: "One Map, Many Trials" in Satellite-Driven Poverty Analysis

Markus B. Pettersson[1], **Connor Thomas Jerzak**[2], **Adel Daoud**[3]

[1]Data Science & AI Division, Department of Computer Science, Chalmers University of Technology, Gothenburg, Sweden
[2]Department of Government, The University of Texas at Austin, Austin, TX, USA
[3]Institute for Analytical Sociology, Linköping University, Norrköping, Sweden
markus.pettersson@chalmers.se, connor.jerzak@austin.utexas.edu, adel.daoud@liu.se

Abstract

Machine learning models trained on Earth observation data, such as satellite imagery, have demonstrated significant promise in predicting household-level wealth indices, enabling the creation of high-resolution wealth maps that can be leveraged across multiple causal trials while addressing chronic data scarcity in global development research. However, because standard training objectives prioritize overall predictive accuracy, these predictions often suffer from shrinkage toward the mean, leading to attenuated estimates of causal treatment effects and limiting their utility in policy evaluations. Existing debiasing methods, such as Prediction-Powered Inference (PPI), can handle this attenuation bias but require additional fresh ground-truth data at the downstream stage of causal inference, which restricts their applicability in data-scarce environments. We introduce and evaluate two post-hoc correction methods—Linear Calibration Correction (LCC) and a Tweedie's correction approach—that substantially reduce shrinkage-induced prediction bias without relying on newly collected labeled data. LCC applies a simple linear transformation estimated on a held-out calibration split; Tweedie's method locally de-shrink predictions using density score estimates and a noise scale learned upstream. We provide practical diagnostics for *when* a correction is warranted and discuss practical limitations. Across analytical results, simulations, and experiments with Demographic and Health Surveys (DHS) data, both approaches reduce attenuation; Tweedie's correction yields nearly unbiased treatment-effect estimates, enabling a "one map, many trials" paradigm. Although we demonstrate on EO-ML wealth mapping, the methods are not geospatial-specific: they apply to any setting where imputed outcomes are reused downstream (e.g., pollution indices, population density, or LLM-derived indicators).

Code — https://aidevlab.org/unshrink
Extended version — https://arxiv.org/abs/2508.01341

Introduction

Machine learning (ML) models trained on Earth observation (EO) data have recently demonstrated impressive performance in predicting household-level wealth indices, such as the International Wealth Index (IWI), with reported R^2 values approaching 0.80 (Pettersson et al. 2023; Jean et al. 2016; Daoud et al. 2023; Chi et al. 2022). These EO-ML models offer a much-needed solution to longstanding data scarcity issues in global development, enabling researchers to generate high-resolution, retrospective "wealth maps" in data-poor regions. Indeed, two recurring downstream uses are: (i) evaluating aid or infrastructure programs (*e.g.*, differences in post-intervention wealth across treated vs. control regions), and (ii) tracking regional poverty trends over time.

But while these predictions are promising in the EO-ML upstream task, a key question remains (Jerzak, Johansson, and Daoud 2023a,b): the value of EO-ML-derived data hinges on whether domain researchers, in the downstream phase, can use these data to estimate causal effects for their specific application areas, from evaluating applied environmental to social science questions (Daoud and Dubhashi 2023). However, even under ideal conditions, the predictions of EO-ML models tend to have lower variance than the target values, leading to attenuated treatment effects and, thus, potentially misleading conclusions (Berglund 2012). When upstream predictions shrink toward the global mean, estimated effects are biased toward zero, increasing Type-II error: a true 5% effect can appear as only 2–3%, inviting "no impact" conclusions despite meaningful gains. Figure 1 provides a pipeline overview linking the upstream map to multiple downstream trials without requiring new labels.

Recent research has developed methods that can debias upstream machine-learning model predictions, but they require either fresh data in the downstream stage (Olofsson et al. 2013; Egami et al. 2023; Lu, Bates, and Wang 2024) or changing the way the model is trained (Ratledge et al. 2022). For example, relying on such fresh data, Angelopoulos et al. (2023) propose Prediction-Powered Inference (PPI), a method that corrects any ML model for downstream inference. Similarly, Egami et al. (2023) achieve the same goal but with better sample efficiency (de Pieuchon et al. 2025), but it also requires knowing the probability of observing a sample point. While these methods offer a path to integrate upstream and downstream analyses, their success depends on the availability of newly collected treatment-predictor-outcome tuples in the downstream causal step or knowledge of the data labeling process (Egami et al. 2023). Often, such data are prohibitively costly or unethical to collect.

Ratledge et al. (2022) address the challenge of attenuation bias and the unavailability of data in the downstream by

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

adjusting the model's loss function instead. During training, they penalize quintile-level discrepancies between observed and predicted means, encouraging predictions to match the true outcome distribution more closely across its support. Although their solutions do not need additional downstream data, they require modifying the training pipeline, which is computationally expensive and can degrade predictive performance.

Thus, to enable broader applicability of ML models, the upstream team needs to create a data product that is agnostic to downstream use, requires no downstream-fresh data, produces unbiased causal estimates (due to attenuation), and is computationally efficient (Daoud and Dubhashi 2023).

We address this gap by proposing a linear calibration method and a family of Tweedie corrections, a novel approach that reduces prediction bias without requiring fresh labeled data. We evaluate these two approaches against three others: Prediction-Powered Inference (PPI), the loss-adjustment approach of Ratledge et al. (2022), and naive (uncorrected) predictions. The methods introduced here pave the way for the same wealth map to be reused across multiple trials without needing to refit a model for each trial. This also removes the requirement for interaction between the upstream machine-learning team creating the data product and any other domain-knowledge team in the downstream phase. Hence the concept: *one map, many trials*. We derive analytical proofs and demonstrate the empirical performance of all methods in both simulations and real-world applications, using data from the African Demographic and Health Survey (DHS).

We make these corrections accessible via an open-source software package called `unshrink` (see Appendix F); although developed for EO-ML wealth maps, these same attenuation and remedies arise broadly whenever surrogate outcomes replace direct measurements, including in environmental (air pollution, urban heat), demographic (population density), and text/LLM-annotated constructs.

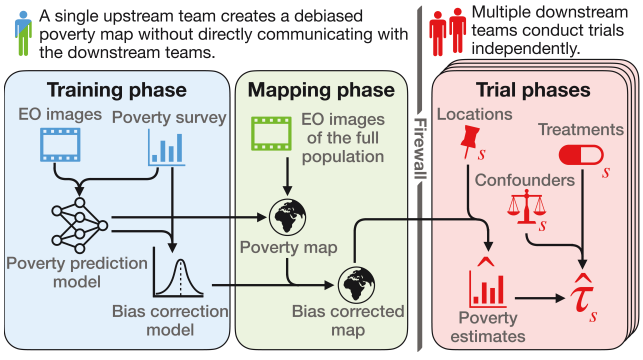

Figure 1: Pipeline for causal inference using ML predictions without fresh ground-truth data. Upstream teams train and calibrate once; downstream teams reuse the same map across multiple trials.

Problem Setup & Related Work

Statistical predictions can systematically shrink estimates toward the central modes of the training distribution (Ting 2024a). This phenomenon is rooted in optimal error-minimization estimation (cf. Stein's paradox), where a model "hedges" predictions in regions of greater uncertainty. When facing noise or limited training data, the optimal strategy, in terms of minimizing residuals, is to pull predictions toward denser regions of the training data. For instance, consider an image that appears with an IWI of 80, but the model has observed this outcome only once, compared to thousands of observations with an IWI of around 70. A perfect prediction would be 80 for this data point, but the model will pull this data point's prediction towards 70 instead to hedge against the risk of overcommitting to a rare value.

When such biased predictions replace true outcomes in downstream causal analyses, the resulting regression-to-the-mean induces attenuation bias, reducing observed treatment group differences even under randomization (Shu and Yi 2019). The difference in mean predicted outcomes between treated and control units shrinks both group means toward the global average, thereby attenuating the estimated effect. In our held-out DHS clusters, the calibration curve exhibits this shrinkage pattern—overpredicting poor locations and underpredicting wealthy ones (Figure 2).

Regression-to-the-mean bias is particularly acute in upstream tasks with scarce ground truth (due to finite samples) or in high-dimensional input data (due to the curse of dimensionality), forcing greater reliance on error-prone predictions—something especially pertinent in development economics, political science, and global health contexts—where household data is often expensive or unavailable—to increasingly substitute upstream predictions (Kino et al. 2021; Kakooei et al. 2024; Daoud and Johansson 2024). Moreover, high out-of-sample R^2 does not eliminate this bias: models can achieve strong accuracy yet yield attenuated causal estimates (Ting 2024b).

Existing remedies are PPI, which subtracts a rectifier estimated from fresh ground-truth pairs at the causal-inference stage (Angelopoulos et al. 2023), or loss-function adjustments that penalize quintile-level errors (Ratledge et al. 2022). However, these approaches require that the upstream team adapts the data for downstream analysis, slowing or limiting altogether "one map, many trials" efforts.

To address these limitations, we propose a Tweedie-inspired approach. This Tweedie approach debiases the shrinkage effect in ML-predicted outcomes, enabling downstream researchers to accurately recover causal effects from the upstream data product. Our Tweedie framework treats the ML model as a black box: we assume access only to its out-of-sample predictions and do not rely on retraining or modifying the model. Under this setting, we demonstrate that shrinkage can be reversed, either by learning a global calibration correction from held-out labeled data or by estimating a local correction using ideas from Tweedie's formula, initially described by Robbins and Tweedie in 1956 in the context of empirical Bayes estimation of means under Gaussian observation noise. Both strategies aim to recover unbiased treatment effect estimates from biased pre-

dictions, without requiring fresh ground-truth labels at the time of intervention. We benchmark our Tweedie adaptation to existing methods, which are compared in Table 5 based on their associated assumptions and data/training requirements.[1] In short, Tweedie's formula is canonical; we adapt it to post-hoc, local de-shrinkage of surrogate outcomes used for causal estimation without new labels, estimating the density score and noise scale from upstream data and applying correction downstream at analysis time.

Setup. Our pipeline assumes access to two datasets. First, from upstream, a labeled dataset with training points (with indices in $\mathcal{I}_{\text{Train}}$): $D_{\text{Train}} = \{(\mathbf{X}_i, Y_i) : i \in \mathcal{I}_{\text{Train}}\}$, where $\mathbf{X}_i \in \mathcal{X}$ denotes covariates (e.g., satellite imagery) and Y_i is ground-truth outcome (e.g., poverty level from the DHS).

Next, in the downstream phase, we assume that different teams have assembled collections of trial datasets indexed by s, corresponding to different randomized controlled trials, quasi-experiments, or geographic partitions, with a specific treatment (or action) of interest indexed by treatment setting: $D_{\text{Trial } s} = \{(\mathbf{X}_i, \hat{Y}_i, A_{is}) : i \in \mathcal{I}_{\text{Trial } s}, s \in \mathcal{S}\}$. The data product associated with the outcome is population-wide (e.g., all-Africa), but investigators may want to use that data in multiple causal settings, s, each with a different treatment, $A_{is} \in \{0, 1\}$. The outcome Y_i is not directly observed for all i. Instead, inference will need to take place through the noisy variable, $\hat{Y}_i$. We henceforth remove dependence on s for simplicity.

A key idea here is that the upstream team—a machine-learning team—creates an *agnostic* dataset (Grimmer, Roberts, and Stewart 2022), meaning that this team's data product should produce unbiased causal estimates for any number of applications (trials) for a variety of (social science) teams downstream. This setup implies that the upstream team versus the downstream teams should be able to execute on their respective research goals without communication—hence the firewall between the two phases.

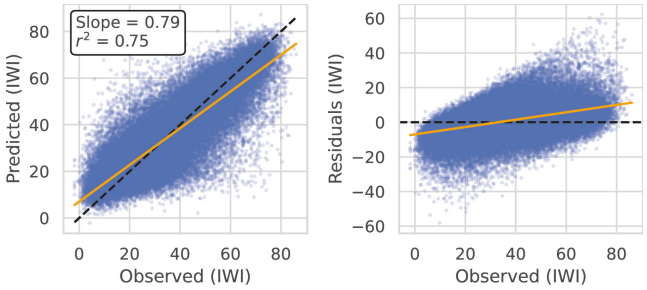

Figure 2: The left is a calibration plot of predicted versus true IWI values at held-out DHS clusters. The orange line shows that the predictions exhibit attenuation bias, resulting in overestimation of wealth for poorer locations and underestimation for wealthier ones. The dashed lines represent the ideal relationship. The right plot shows the residuals.

[1]As previously mentioned, the only method we are unable to deploy in our setting is the DSL method, as the probability of DHS samples depends on unknown country-level DHS sampling strategies (probabilities not publicly provided at the unit level).

Debiasing Methods

In this section, we present a suite of methods designed to debias treatment effect estimates that rely on ML-predicted outcomes. Our overarching goal is to mitigate regression-to-the-mean dynamics in the downstream analysis, improving estimates of causal effects or population-level statistics. We first more formally describe two existing solutions, *PPI* and the method of Ratledge et al. (2022), before introducing two no-new-data debiasing approaches—(1) Linear Calibration Correction (LCC) and (2) Tweedie's Corrections—that aim to correct shrinkage bias without requiring fresh labels in the causal step.

Existing Debiasing Methods (PPI, Ratledge, DSL)

PPI proposes a framework for valid statistical inference when a large set of unlabeled observations is paired with a predictive model's outputs (Angelopoulos et al. 2023). Concretely, letting θ^* be the target parameter (e.g., a population mean) and f denote the model that maps covariates $\mathbf{X}_i$ to predicted outcomes $f(\mathbf{X}_i)$, PPI forms an estimator by adding a *rectifier* term—an adjustment for the average discrepancy between Y_i and $f(\mathbf{X}_i)$ observed in a smaller, labeled dataset. An estimate for θ^* is computed using an "imputed" version based on unlabeled data ($\{\mathbf{X}_i\}_{i=n+1}^{N}$), then corrected by subtracting the bias estimated on the labeled subset. For mean estimates, this adjustment is:

$$\hat{\theta}^{\text{PPI}} = \frac{1}{N-n}\sum_{i=n+1}^{N} f(\mathbf{X}_i) - \frac{1}{n}\sum_{i=1}^{n}\left(f(\mathbf{X}_i) - Y_i\right)$$

Because the rectifier is estimated on ground-truth pairs $\{\mathbf{X}_i, Y_i\}_{i=1}^{n}$, the procedure maintains valid coverage for θ^* in large samples, provided that the predictive model f is fixed (i.e., not retrained on the labeled data). Unlike our approach, however, PPI crucially relies on gathering labeled observations in the downstream phase. Such labels come from high-quality surveys coordinated by USAID across the African continent and cost millions of US dollars to execute. Given the 2025 funding cuts to USAID, it is unlikely that new and fresh DHS data points will be collected. Hence, there is an acute need for unbiased data creation, which should be fully handled in the upstream phase. We will nonetheless include PPI correction in our analysis with the goal of showing *what would have happened to our estimates had fresh data actually been available downstream.*

Another approach to remedy attenuation bias is to adjust the loss function: Ratledge et al. (2022) propose a specialized upstream training procedure that directly penalizes quintile-level discrepancies between observed and predicted means, thereby encouraging $\hat{Y}_i$ to match the true distribution of Y_i more closely across the support of the outcome distribution. Specifically, the training objective augments the mean squared error loss with an ℓ_2 penalty and a quintile-bias term:

$$\mathcal{L}(\theta) \;=\; \text{MSE}(\theta) \;+\; \lambda_r \,\|\theta\|_2^2 \;+\; \lambda_b \,\hat{E}_b(\theta),$$

where $\text{MSE}(\theta)$ is the mean squared error given parameters, θ, $\lambda_r \,\|\theta\|_2^2$ is a regularization term, and $\lambda_b \hat{E}_b(\theta)$ is the

quintile-specific bias loss, with $\hat{E}_b(\theta) = \max_j\{\hat{B}_j(\hat{f}_\theta)^2\}$ and $\hat{B}_j(\hat{f}_\theta) = \mathbb{E}[\hat{f}_\theta(\mathbf{X}_i) - Y_i \mid Y_i \in Q_j]$ penalizing quintile-level discrepancies (where quantiles are indexed by j). Here, λ_r and λ_b denote the weighting terms for the regularization and quantile-specific loss, respectively.[2]

Through the quintile-specific bias loss, this approach reduces the likelihood that group-level estimates of causal effects are attenuated when $\hat{Y}_i$ is used in downstream analyses by directly operating on model parameters, θ. In contrast, the two new methods introduced next operate as post-hoc corrections on out-of-sample predictions from a black-box model, requiring no specialized retraining or customized loss function that might reduce predictive performance.

Linear Calibration Correction (LCC)

For LCC, we approximate the relationship between EO-ML predictions and ground-truth wealth on a held-out calibration set by a linear transformation:

$$\mathbb{E}[\hat{Y}_i \mid Y_i,\, A_i] = \mathbb{E}[\hat{Y}_i \mid Y_i] = k\,Y_i + m, k > 0,\, m \in \mathbb{R}. (A1)$$

When $0 < k \leq 1$, shrinkage toward the mean is indicated; $m \in \mathbb{R}$ is the intercept, representing the mean fixed bias. The first equality essentially indicates that there is now conditional independence of $\hat{Y}_i$ from group/treatment variable $A_i \in \{0, 1\}$ given Y_i; the second equality indicates a global affine (linear) calibration of predictions to truth, i.e., $\mathbb{E}[\hat{Y}_i \mid Y_i] = kY_i + m$ with constant slope and intercept. This linearity might not always be empirically satisfied, but appears to occur approximately in our data (see, e.g., Figure 9). When these predictions are used in place of true outcomes to estimate causal effects, this shrinkage attenuates estimated treatment effects.

Let $\tau = \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]$ denote the true Average Treatment Effect (ATE). Using $\hat{Y}_i$ instead of Y_i for ATE estimation yields the bias shown in Proposition 1.

Proposition 1 (Attenuation bias under linear shrinkage) *Let Y_i be the true outcome and $\hat{Y}_i$ an ML prediction obeying the linear calibration model (A1) with k and m constants. Let $A_i \in \{0, 1\}$ denote a binary treatment that is randomly assigned ($A_i \perp Y_i(a)$ for $a \in \mathcal{A}$). Define the true average treatment effect (ATE) $\tau := \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$ and the estimate, $\hat{\tau}$, from the difference in predicted expected outcomes:*

$$\hat{\tau} := \mathbb{E}[\hat{Y}_i \mid A_i = 1] - \mathbb{E}[\hat{Y}_i \mid A_i = 0].$$

Then $\mathbb{E}[\hat{\tau}] = k\,\tau$. That is, the ATE estimated on the predictions is multiplied by the shrinkage factor k and is therefore attenuated toward 0 whenever $k < 1$.

Proof. Take expectations and use linearity:

$$\begin{aligned}\hat{\tau} &= \mathbb{E}[\hat{Y}_i \mid A_i = 1] - \mathbb{E}[\hat{Y}_i \mid A_i = 0] \\ &= \mathbb{E}[\mathbb{E}[\hat{Y}_i \mid Y_i, A_i = 1] \mid A_i = 1] \\ &\quad - \mathbb{E}[\mathbb{E}[\hat{Y}_i \mid Y_i, A_i = 0] \mid A_i = 0] \\ &= \mathbb{E}[kY_i + m \mid A_i = 1] - \mathbb{E}[kY_i + m \mid A_i = 0] \\ &= k\{\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]\} = k\tau.\end{aligned}$$

Replacing observed with potential outcomes follows from randomization; ATE estimates with $\hat{Y}_i$ are biased downward by a factor of k.

Now, to adjust for this attenuation bias, we will estimate k and m on a held-out calibration set from the training data by regressing $\hat{Y}_i$ on Y_i (i.e., fitting the linear model with functional form, $\hat{Y}_i = \hat{k}Y_i + \hat{m} + e_i$, where e_i are mean 0 residuals). We then invert this fitted relationship (solving for Y_i, which gives $Y_i = \frac{(\hat{Y}_i - \hat{m} - e_i)}{\hat{k}}$) to obtain a calibrated estimate, given by $\hat{Y}_i^L = \frac{\hat{Y}_i - \hat{m}}{\hat{k}}$, assuming conditional mean zero residuals e_i. Here, $\hat{Y}_i^L$ reflects the calibrated estimate of the true outcome, Y_i. This calibrated estimator restores the scale of the outcome variable, yielding Proposition 2:

Proposition 2 (Consistency of LCC correction) *Under the conditions of Proposition 1, estimate (k, m) on an independent calibration sample and set*[3]*:*

$$\hat{Y}_i^L := (\hat{Y}_i - \hat{m})/\hat{k};\ \hat{\tau}^L := \mathbb{E}[\hat{Y}_i^L | A_i = 1] - \mathbb{E}[\hat{Y}_i^L | A_i = 0].$$

If $(\hat{k}, \hat{m}) \xrightarrow{p} (k, m)$, and $k > 0$, then $\hat{\tau}^L \xrightarrow{p} \tau$; i.e., the linear calibration correction removes the attenuation bias.

Proof sketch. By Slutsky's theorem, since $(\hat{k}, \hat{m}) \xrightarrow{p} (k, m)$, it follows that:

$$\begin{aligned}\hat{Y}_i^L &= [[kY_i + m + e_i] - \widehat{m}]/\hat{k} \\ &= \underbrace{(k/\hat{k})Y_i}_{\to Y_i} + \underbrace{(m - \hat{m})/\hat{k}}_{\to 0} + \underbrace{e_i/\hat{k}}_{e_i/k} \to Y_i + e_i/k.\end{aligned}$$

Since $\mathbb{E}[e_i | A_i] = 0$ under the conditional mean 0 assumption outlined earlier,

$$\begin{aligned}\hat{\tau}^L &\to \mathbb{E}[Y_i + e_i/k \mid A_i = 1] - \mathbb{E}[Y_i + e_i/k \mid A_i = 0] \\ &= \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0] = \tau.\end{aligned}$$

A drawback is that we assume a linear model for the shrinkage mechanism, which enters through an assumed functional form for the expected value of $\hat{Y}_i$ given Y_i (cf. A1); estimating $\hat{m}$ and $\hat{k}$ introduces another potential source of error.

Tweedie's Correction

To go beyond the linearity assumption in LCC in a way that respects the structure of ML predictions, we now turn to Tweedie's identity, which provides a local, density–score–based way to reverse shrinkage without assuming global linearity of the measurement error process.

[2] In our experiments, we use AdamW with weight decay for ℓ_2 regularization; to avoid double-counting, we set $\lambda_r = 0$ whenever weight decay is non-zero.

[3] To be clear, $(\hat{k}, \hat{m})$ are estimated on a random validation-style calibration split carved from the upstream labeled data.

A first instinct might be to formalize shrinkage in a way that generalizes beyond the linearity assumption of LCC by using a measurement-error model, in which the prediction is a noisy version of the truth:

$$\hat{Y}_i = Y_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

However, in ML predictions scenarios, models approximate $\mathbb{E}[Y_i \mid \mathbf{X}_i]$ for covariates $\mathbf{X}_i$ and thus predictions tend to exhibit systematically lower variance than targets due to shrinkage toward the mean (the opposite of what is implied in the baseline classical case, where $\text{Var}(\hat{Y}_i) > \text{Var}(Y_i)$). Indeed, upstream ML models are trained to minimize expected loss over $\mathbf{X}_i$ (Shalev-Shwartz and Ben-David 2014):

$$Y_i = \hat{Y}_i + \epsilon_i, \quad \hat{Y}_i = f_\theta(\mathbf{X}_i)$$

where θ are neural parameters optimized to reduce MSE. Any residual noise $Y_i - \hat{Y}_i$ thus arises *after* the prediction, approximately independent of $\hat{Y}_i$ itself in distribution (as the model would otherwise exploit correlations to reduce loss). This reasoning therefore inverts the classical view, motivating adoption of the Berkson error model (Carroll et al. 2006; Heid et al. 2004):

$$Y_i = \hat{Y}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \perp \hat{Y}_i,$$

as a more natural framing for debiasing ML surrogates.

Adapting Tweedie's Formula. To reverse shrinkage under this model without fresh labels, we propose an approach using *Tweedie's formula*, introduced by Robbins (1956) and commonly encountered in Gaussian diffusion models:

—Tweedie's Formula:—

For any random variable, Z, with $Z = \omega + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$,

$$\mathbb{E}[\omega \mid Z = z] = z + \sigma^2 \tfrac{d}{dz} \log p_Z(z) \quad \text{(Efron 2011)}.$$

In its native form, Tweedie's formula under the Berkson model gives us an expression for $\mathbb{E}[\hat{Y}_i \mid Y_i = y]$, i.e., recovering model predictions under attenuation bias from true labels. Our aim is to derive a novel debiasing method that works in the opposite direction, i.e., recovering estimates of the true labels without attenuation bias in downstream applications. We achieve this goal by introducing a new-label-free corrected pseudo-outcome, $\widetilde{Y}_i$, whose conditional mean recovers the truth, $\mathbb{E}[\widetilde{Y}_i \mid Y_i] = Y_i$. To obtain this end, we need to deploy the score-swapping relation:

$$\frac{d}{dy} \log p_Y(y) = \mathbb{E}\left[\frac{d}{d\hat{y}} \log p_{\hat{Y}}(\hat{Y}_i) \mid Y_i = y \right],$$

as derived in Appendix I, to the Tweedie identity:

$$\begin{aligned} \mathbb{E}\left[\hat{Y}_i \mid Y_i = y\right] &= y + \sigma^2 \frac{d}{dy} \log p_Y(y) \\ &= y + \sigma^2 \, \mathbb{E}\left[\frac{d}{d\hat{y}} \log p_{\hat{Y}}(\hat{Y}_i) \mid Y_i = y \right] \\ \implies y &= \mathbb{E}\left[\hat{Y}_i - \sigma^2 \frac{d}{d\hat{y}} \log p_{\hat{Y}}(\hat{Y}_i) \mid Y_i = y \right], \end{aligned}$$

where the move from the second to the third line follows from linearity of expectations.

Intuitively, subtraction in the correction arises from inverting the forward application of Tweedie's formula under the Berkson model. Unconditionally, $\mathbb{E}[\hat{Y}_i] = \mathbb{E}[Y_i]$, so predictions are centered around the true values on average (as ensured by MSE minimization during training). Conditionally, however, $\mathbb{E}[\hat{Y}_i \mid Y_i = y] = y + \sigma^2 \frac{d}{dy} \log p_Y(y)$, where the score term attenuates extreme predictions toward the population mean—for instance, when y exceeds the mode of p_Y, the score is negative and pulls the conditional expectation below y and towards the mode. In other words, while the overall distribution of predictions matches that of the truths in expectation, a given prediction (absent adjustment) shrinks toward the mean conditional on the realized true value. The inversion thus requires subtracting the score—now estimated via the swapped relation on $p_{\hat{Y}}$—to reverse this attenuation and expand the tails.

This discussion then yields the Tweedie corrected pseudo-outcome under the Berkson model:

$$\widetilde{Y}_i = \hat{Y}_i - \sigma^2 \frac{d}{d\hat{y}} \log p_{\hat{Y}}(\hat{Y}_i). \tag{1}$$

The score term expands both tails: when the score is positive (where we evaluate the score rising to the left of a local mode), it pushes $\hat{Y}_i$ downward; when the score is negative (score evaluated to the right of a local mode) it pushes $\hat{Y}_i$ upward; around a mode, where the score is near zero, adjustment is negligible. In practice, σ and the score function will need to be estimated on data: σ is estimated from residuals; the score function with, e.g., a KDE on $\{\hat{Y}_i\}_{i=1}^n$.

Proposition 3 (Tweedie calibration identity) *The pseudo-outcome satisfies:* $\mathbb{E}\left[\widetilde{Y}_i \,\middle|\, Y_i\right] = Y_i$.

Proof sketch. By Tweedie's identity, $\mathbb{E}[\hat{Y}_i \mid Y_i = y] = y + \sigma^2 \frac{d}{dy} \log p_Y(y)$. By score-swapping (Appendix I), $\frac{d}{dy} \log p_Y(y) = \mathbb{E}[\frac{d}{d\widetilde{Y}} \log p_{\hat{Y}}(\hat{Y}_i) \mid Y_i = y]$. Take conditional expectations of $\widetilde{Y}_i$ given $Y_i = y$ to obtain y.

Remark. This is the key statement giving rise to well-calibrated downstream applications of the "one-map" ML prediction. Note that the variance of $\tilde{Y}_i$ will almost certainly be larger than the variance of $\hat{Y}_i$, so for purely predictive purposes, the overall MSE of pseudo-outcomes may also be larger that of $\hat{Y}_i$; however, downstream teams will for causal inference purposes be concerned with well-calibrated group conditional means—discussed later.

Proposition 4 (Local Reduction to LCC) *If $p_{\hat{Y}}$ is locally Gaussian with mean μ and variance γ^2, then the score is $-(\hat{y} - \mu)/\gamma^2$ and the pseudo-outcome is the affine map:*

$$\widetilde{Y}_i \;=\; \hat{Y}_i - \sigma^2 \left(-\frac{\hat{Y}_i - \mu}{\gamma^2} \right) = \mu + \left(1 + \tfrac{\sigma^2}{\gamma^2}\right)(\hat{Y}_i - \mu),$$

i.e., a linear calibration with slope $(1 + \sigma^2/\gamma^2)$ increasing in the ratio of measurement error variance to local signal variance that expands extremes away from μ.

Proof. Plug the Gaussian score into $\widetilde{Y}_i$.

Further Relations to LCC. When the score term is negligible (e.g., flat density), $\widetilde{Y}_i \approx \hat{Y}_i$, reverting to uncorrected predictions. While LCC provides a single global scaling adjustment for errors that operates in a uniform manner, Tweedie adjustment here can handle more complex distributional patterns, rescaling depending on the position of predictions relative to local (possibly multiple) modes. Thus, LCC captures a single regression-to-the-mean dynamic, while Tweedie adjustment captures a broader range of error processes, adjusting for regression to local modes.

Proposition 5 (Unbiasedness of Tweedie means) *Under the assumptions of Proposition 3, for any partitioning variable A_i such that $\hat{Y}_i \perp A_i \mid Y_i$ (which implies $\tilde{Y}_i \perp A_i \mid Y_i$ [because $\tilde{Y}_i$ is a deterministic function of $\hat{Y}_i$]):*

$$\mathbb{E}[\widetilde{Y}_i \mid A_i = a] \;=\; \mathbb{E}[Y_i \mid A_i = a].$$

Consequently, the difference-in-means estimator $\hat{\tau} := \hat{\mathbb{E}}[\widetilde{Y}_i \mid A_i = 1] - \hat{\mathbb{E}}[\widetilde{Y}_i \mid A_i = 0]$ satisfies $\mathbb{E}[\hat{\tau}] = \mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]$. Under random assignment, $\mathbb{E}[\hat{\tau}] = \tau$.

Proof sketch. Iterated expectations using Proposition 3 yield $\mathbb{E}[\widetilde{Y}_i \mid A_i = a] = \mathbb{E}\{\mathbb{E}[\widetilde{Y}_i \mid Y_i, A_i = a] \mid A_i = a\} = \mathbb{E}[Y_i \mid A_i = a]$. Independence of A_i and ϵ_i is key here, since it implies that there is no additional relationship between $\tilde{Y}_i$ and Y_i once Y_i is known, so $\mathbb{E}[\tilde{Y}_i|Y_i] = \mathbb{E}[\tilde{Y}_i|Y_i, A]$. Randomization identifies $\mathbb{E}[Y_i \mid A_i = a] = \mathbb{E}[Y_i(a)]$ and thus enables unbiased ATE estimates in RCT scenarios.

Remark. The key assumption here, $\hat{Y}_i \perp A_i \mid Y_i$, indicates that, conditional on the true value of Y_i, there is no additional signal relating A_i and $\hat{Y}_i$. For example, if A_i is gender and Y_i is income, the assumption states that, among truly high-income individuals, the model error does not predict gender. This unbiasedness result does therefore not mean all forms of bias are eliminated, simply that the bias added from attenuation is eliminated. See Appendix I for a discussion of plug-in consistency, which is, in practice, complicated by how overfitting in the training set can lead to distribution shifts affecting performance.

Viewed through an optimization lens, the Tweedie pseudo-outcome $\widetilde{Y}_i = \hat{Y}_i - \sigma^2 \frac{d}{d\hat{Y}} \log p_{\hat{Y}}(\hat{Y}_i)$ is exactly a single gradient-descent step on the log-density $\log p_{\hat{Y}}$, with learning rate $\eta = \sigma^2$, that nudges predictions away from the over-shrunk high-density region toward the tails (see Appendix I). This perspective naturally points to a broader class of Tweedie-style updates obtained by varying the step size, number of steps, and implicit error model. In the Appendix, we discuss how one might organize the resulting *Tweedie corrections* as a family indexed by the implicit error model that links the latent outcome Y_i and the ML prediction $\hat{Y}_i$. The variants differ by the structure one is willing to assume and by how one wishes to correct global scaling (via a single calibration slope k and number of update steps) in addition to local shrinkage. In our experiments below, we focus on the Berkson variant we have been analyzing—leaving to future work elucidation of the full range of bias-variance tradeoffs in the Tweedie family.

Experiments with Simulated Data

Simulation Design. Our simulation study benchmarks all debiasing methods and monitors the upstream and downstream analyses while maintaining the firewall between them. The simulation uses a canonical directed acyclic graph (DAG), shown in Figure 3. It is *canonical* in the sense that the EO-ML literature tends to assume this structure (Sakamoto, Jerzak, and Daoud 2025), while illustrating the various research phases and corresponding data needs.

In phase 0, the data-generation process (approximating *reality*) unfolds in the direction of the arrows. In the upstream phase, only the blue-nodes data are available, with finite samples of Y_i, to produce $\hat{Y}_i$ for the entire population N. In the downstream, only red nodes are available: $\hat{Y}_i$ is now available for N, jointly with C_i and A_{is}. They enable inference on the estimand τ_s. (We suppress the index s henceforth.) Appendix D delineates our simulation further.

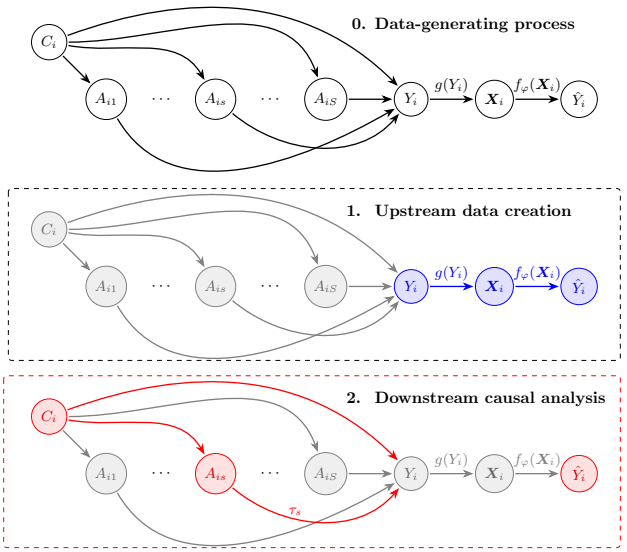

Figure 3: A panel illustrating the problem setup.
0. Reality (data generation) shows the complete causal DAG.
1. Upstream phase: Y_i, $\mathbf{X}_i$, and $\hat{Y}_i$ are observed (blue).
2. Downstream phase: A_{is}, C_i, and $\hat{Y}_i$ are available (red.)

Simulation Results. We focus on two key objectives in the downstream task: minimizing error and minimizing causal bias. To assess error, we measure how closely the predicted values match the true values of τ, using Mean Absolute Error (MAE). A lower MAE indicates a more accurate recovery of the true treatment effects. To assess bias, we examine the calibration slope from regressing the predicted effects $\hat{\tau}$ on the true τ values and analyzing the calibration slope. If the estimates are unbiased, the slope should be close to one. Intuitively, a slope of one means predicted and true scales align (no global shrinkage) so downstream treatment effects are not attenuated purely by scaling. Deviations from one indicate systematic over-/underestimation

across the range of τ. We report both the slope and its standard error, checking whether the 95% confidence interval includes one. Figure 7 visualizes these slope results.

Results (Table 1) show that Tweedie's correction achieves the lowest MAE and yields slope estimates whose confidence interval contains one, indicating no detectable attenuation bias. In contrast, both the Ratledge and naive approaches yield calibration slopes significantly below one, indicating systematic underestimation of effects.

Method	MAE	Slope $\pm$ SE	$1 \in$ 95% CI?
Tweedie's	**0.04**	**0.995 $\pm$ 0.006**	✓
LCC	0.05	1.008 $\pm$ 0.007	✓
PPI (10%)	0.19	0.985 $\pm$ 0.028	✓
Ratledge	0.37	0.641 $\pm$ 0.024	×
Naive	0.48	0.535 $\pm$ 0.004	×

Table 1: Performance on simulated treatment-effect estimation. Tweedie's correction yields the lowest error and an unbiased slope; Ratledge and the naive approach show substantial attenuation bias.

Experiments with Real Data

In this section, we conduct two types of downstream-real experiments. First, we will evaluate how well our debiasing methods succeed in descriptive tasks, especially those involving group aggregation. Second, we will apply our procedure to causal tasks. These two tasks cover the main use cases of our EO-ML-driven poverty data (see Figure 5).

The full dataset consists of $n \approx 69{,}000$ geo-located survey clusters from the Demographic and Health Surveys (DHS) across 30 African countries (2009–2020), where each cluster comprises 20–30 households sampled via stratified two-stage design (primary sampling units selected with probability proportional to population size, followed by random household selection). We partitioned the data into five random folds for cross-validation (CV). For each run, four folds were used as upstream data (three for training, one for calibration): $D_{\text{Train}} = \{(\mathbf{X}_i, Y_i) : i \in \mathcal{I}_{\text{Train}}\}$. Downstream trial datasets $D_{\text{Trial } s}$ were drawn from the remaining, fifth fold and geo-referenced aid interventions, s, ensuring no overlap with training data.

Here, Y_i denotes the International Wealth Index (IWI), a continuous asset-based wealth score (0–100) aggregated at the cluster level to mitigate noise (Burke et al. 2021; Zhu, Jerzak, and Daoud 2025). The covariate array $\mathbf{X}_i$ consists of Landsat satellite imagery (30 m resolution, six multispectral bands) centered on each DHS cluster (6.72 km × 6.72 km tiles), processed by taking the three-year median over cloud-free pixels leading up to the survey date.

Two models were fitted for each fold, one using MSE and one using the loss introduced by Ratledge et al. (2022). The models shared the same ResNet-50 architecture and were initialized with weights pretrained on the SSL4EO-L dataset (Stewart et al. 2023). For more details about model training, see Appendix B. The resulting MSE models all have an R^2 score between 0.74 and 0.76 on their hold-out fold, and an average slope of 0.79 . Their combined predictions can be seen in Figure 2. The Ratledge models have a slope of 0.99, but at the cost of reduced R^2 scores between 0.67 and 0.70.

Estimating Region Means with Bias Adjustment. In our descriptive task, we estimate conditional means, averaging wealth by local second-level administrative region (ADM2), roughly corresponding to municipalities or counties. We group the clusters in the held-out downstream fold by ADM2 region and compare the mean observed IWI with the predicted IWI with and without the different methods of bias correction. These quantities are subject to the same shrinkage bias as causal estimates, and our framework applies equally well in this setting, enabling us to assess how effectively we correct biased group-level estimates.

Descriptive Results. As expected, the MSE of the naive model exhibits shrinkage behavior: it tends to overpredict the wealth of poorer regions and underpredict the wealth of richer regions. As shown in Figure 4, Tweedie's correction accurately compensates for this, not only resulting in less biased estimates (regression slope of 0.83 to 0.90), but also decreases the MAE from 2.67 to 2.39. In fact, if we consider only regions with at least 100 clusters, the Tweedie-corrected estimates exhibit virtually no shrinkage bias, with a slope of 0.99. Appendix C explicates.

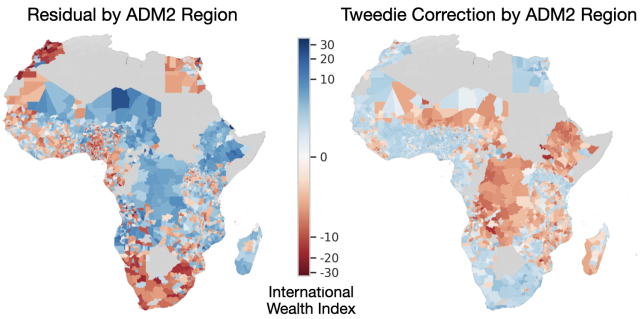

Figure 4: Mean residuals (LEFT) and adjustments from Tweedie's method (RIGHT) for each ADM2 region in the dataset. Regions that are overestimated in the left plot (blue) are shown to be correctly adjusted down in the right plot (red), and vice versa.

Empirical Analysis of Aid Interventions. To assess the practical utility of our correction methods in a setting where the true treatment effect is unknown, we construct an evaluation framework based on real-world development projects. Specifically, we leverage the same dataset of geo-referenced interventions funded by the World Bank and the Chinese government as was used in (Malik et al. 2021; Conlin 2024). Each project is tagged with an associated aid sector $s \in \mathcal{S}$ (e.g., *Health*, *Water Supply and Sanitation*, or *Women in Development*), which defines a downstream trial $D_{\text{Trial } s} = \{(\mathbf{X}_i, \hat{Y}_i, A_{is}) : i \in \mathcal{I}_{\text{Trial } s}\}$. Here, $A_{is} \in \{0, 1\}$ indicates exposure to a sector-s intervention at location i, $\mathbf{X}_i$ are satellite features, and $\hat{Y}_i$ are IWI predictions from the upstream-trained model. Figure 8 depicts the spatial distribution of treatments across the continent.

Since the treatment effects of these interventions are unknown, we deem this analysis a test of external validity by comparing ML predicted wealth values with observed ones. Our goal is to assess whether bias-corrected predictions can reflect causal estimates similar to those from groundtruth.

To avoid label leakage and emulate real-world constraints, we perform a two-stage evaluation. We begin by randomly partitioning the available DHS survey data into two disjoint sets: an upstream set for training and calibrating the EO-ML model, and a downstream set used solely for evaluation. The model is trained on the upstream data, without any exposure to the intervention data or knowledge of future evaluation criteria. Shrinkage corrections (e.g., LCC or Tweedie's correction) are also estimated on this upstream set.

We match project locations to the corresponding ADM2-level administrative regions, roughly equivalent to districts or municipalities. For a given funder-sector pair (e.g., Health projects funded by the World Bank), we identify all ADM2 regions containing at least one relevant intervention site. We then consider villages surveyed in the downstream DHS data within those regions, 3 to 8 years after the intervention, as the treated group. Villages in the same surveys, but outside the treated ADM2 regions, serve as controls.

For each funder-sector combination, we estimate the average treatment effect as the difference in mean outcomes between treated/control villages, for simplicity. We compute this using both observed IWI values and the model-predicted (and correction-adjusted) IWI values. The former provides a benchmark estimate using labeled data, while the latter evaluates the debiasing methods applied to predictions.

Causal Results. See Figure 5 for results. Each point is a causal study. While the x-axis shows what would have happened when using the label Y_i in these studies, the y-axis shows the corresponding causal effect using $\hat{Y}_i$ with its corresponding adjustment approach. Orange lines check the correlation among these data. A score of one is perfect. The analysis shows that Tweedie's correction (panel *d*) comes close to this desiderata, with a correlation of 0.998; and it does so with the lowest MAE of 0.57. Had we used the naive predictions (panel *a*), the correlation would be 0.958 and the MAE would be 0.69. Assuming the PPI set up, with 10% fresh data available, we would have achieved a 0.981; and it does so with the lowest MAE of 0.73. The PPI performance is, however, a function of available data in the downstream, and panel *f* shows that with increasing fresh data, the PPI correction will outperform our Tweedie. However, since PPI demands substantial amounts of such data, this method would necessitate costly new surveys, resulting in large expenditures for downstream teams (de Pieuchon et al. 2025).

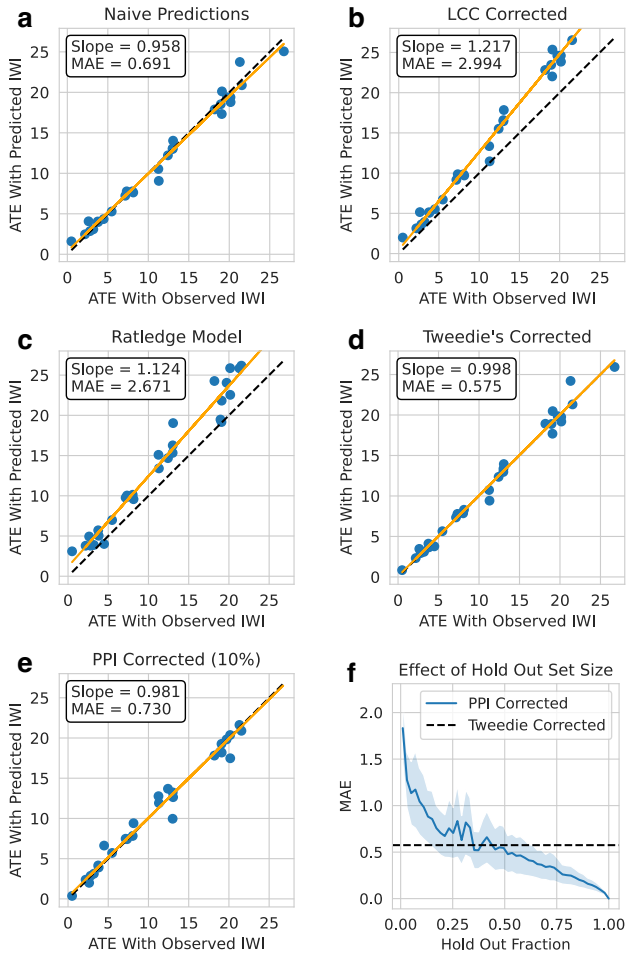

Figure 5: *(a–e)* Estimated treatment effects of aid interventions using unadjusted predictions, which display attenuation bias due to shrinkage. *(f)* PPI outperforms Tweedie's correction in MAE bias reduction on larger hold-out sets, but requires many ground-truth labels.

Discussion and Conclusion

Our experiments show that *post-hoc*, data–free corrections can eliminate the lion's share of the shrinkage-induced attenuation that plagues causal analyses based on machine–learning predictions. In controlled simulations where the true ATE is known, Tweedie's correction reduces mean–absolute error by an order of magnitude and, crucially, restores calibration: the fitted slope of $\hat{\tau}$ on τ is indistinguishable from one. The linear calibration correction also performs nearly as well despite its simplicity, suggesting that much of the bias arises from a first-order scaling distortion that can be identified on a small held-out calibration set. When applied to real DHS imagery, both corrections markedly narrow the gap between predicted and observed country-level means, and in a quasi-experimental evaluation of large-scale aid projects, they bring estimated ATE into line with those obtained from ground-truth survey data.

That said, bias can remain or be reintroduced when (a) the joint distribution of $(\mathbf{X}, Y)$ shifts between up and downstream, (b) errors are strongly heteroskedastic or non-Gaussian so that a single σ over- or under-corrects in parts of the support, or (c) $\hat{\sigma}$ is misestimated (e.g., from overfitting late in training). We therefore (i) recommend estimating σ on the training split to reduce over-correction, (ii) report block-bootstrap uncertainty for spatial analyses, and (iii) monitor calibration after correction. □

References

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

Berglund, L. 2012. Regression dilution bias: tools for correction methods and sample size calculation. *Upsala journal of medical sciences*, 117(3): 279–283.

Burke, M.; Driscoll, A.; Lobell, D. B.; and Ermon, S. 2021. Using Satellite Imagery to Understand and Promote Sustainable Development. 371(6535): eabe8628.

Carroll, R. J.; Ruppert, D.; Stefanski, L. A.; and Crainiceanu, C. M. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC.

Chi, G.; Fang, H.; Chatterjee, S.; and Blumenstock, J. E. 2022. Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3): e2113658119.

Conlin, C. 2024. *Using Machine Learning and Daytime Satellite Imagery to Estimate Aid's Effect on Wealth: Comparing China and World Bank Programs in Africa*. Master's thesis, Linkoping University, Department of Management and Engineering, The Institute for Analytical Sociology (IAS), Linkoping, Sweden. URN: urn:nbn:se:liu:diva-205256, ISRN: LIU-IEI-FIL-A–24/04656–SE.

Daoud, A.; and Dubhashi, D. 2023. Statistical Modeling: The Three Cultures. 5(1).

Daoud, A.; and Johansson, F. D. 2024. The Impact of Austerity on Children: Uncovering Effect Heterogeneity by Political, Economic, and Family Factors in Low- and Middle-Income Countries. 118: 102973.

Daoud, A.; Jordán, F.; Sharma, M.; Johansson, F.; Dubhashi, D.; Paul, S.; and Banerjee, S. 2023. Using satellite images and deep learning to measure health and living standards in India. *Social Indicators Research*, 167(1): 475–505.

de Pieuchon, N. A.; Daoud, A.; Jerzak, C. T.; Johansson, M.; and Johansson, R. 2025. Benchmarking Debiasing Methods for LLM-based Parameter Estimates. arXiv:2506.09627.

Devroye, L.; Györfi, L.; and Lugosi, G. 2013. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media.

Efron, B. 2011. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496): 1602–1614.

Egami, N.; Hinck, M.; Stewart, B.; and Wei, H. 2023. Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 68589–68601. Curran Associates, Inc.

Grimmer, J.; Roberts, M. E.; and Stewart, B. M. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Heid, I.; Küchenhoff, H.; Miles, J.; Kreienbrock, L.; and Wichmann, H. 2004. Two Dimensions of Measurement Error: Classical and Berkson Error in Residential Radon Exposure Assessment. *Journal of Exposure Science & Environmental Epidemiology*, 14(5): 365–377.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining Satellite Imagery and Machine Learning to Predict Poverty. *Science*, 353(6301): 790–794.

Jerzak, C. T.; Johansson, F.; and Daoud, A. 2023a. Integrating earth observation data into causal inference: challenges and opportunities. *arXiv preprint arXiv:2301.12985*.

Jerzak, C. T.; Johansson, F. D.; and Daoud, A. 2023b. Image-based Treatment Effect Heterogeneity. In *Conference on Causal Learning and Reasoning*, 531–552. PMLR.

Kakooei, M.; Bailie, J.; Söderberg, A.; Becevic, A.; and Daoud, A. 2024. Mapping Africa Settlements: High Resolution Urban and Rural Map by Deep Learning and Satellite Imagery. *arXiv preprint arXiv:2411.02935*.

Kino, S.; Hsu, Y.-T.; Shiba, K.; Chien, Y.-S.; Mita, C.; Kawachi, I.; and Daoud, A. 2021. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM-population Health*, 15: 100836.

Lu, K.; Bates, S.; and Wang, S. 2024. Quantifying Uncertainty in Area and Regression Coefficient Estimation from Remote Sensing Maps. *arXiv preprint arXiv:2407.13659*.

Malik, A.; Parks, B.; Russell, B.; Lin, J.; Walsh, K.; Solomon, K.; Zhang, S.; Elston, T.; and Goodman, S. 2021. Banking on the Belt and Road: Insights from a new global dataset of 13,427 Chinese development projects. *Williamsburg, VA: AidData at William & Mary*, 23–36.

Olofsson, P.; Foody, G. M.; Stehman, S. V.; and Woodcock, C. E. 2013. Making Better Use of Accuracy Data in Land Change Studies: Estimating Accuracy and Area and Quantifying Uncertainty Using Stratified Estimation. *Remote sensing of environment*, 129: 122–131.

Pettersson, M. B.; Kakooei, M.; Ortheden, J.; Johansson, F. D.; and Daoud, A. 2023. Time Series of Satellite Imagery Improve Deep Learning Estimates of Neighborhood-Level Poverty in Africa. In *IJCAI*, 6165–6173.

Ratledge, N.; Cadamuro, G.; de la Cuesta, B.; Stigler, M.; and Burke, M. 2022. Using Machine Learning to Assess the Livelihood Impact of Electricity Access. *Nature*, 611(7936): 491–495. Number: 7936 Publisher: Nature Publishing Group.

Robbins, H. 1956. An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, volume I, 157–163. Berkeley and Los Angeles: University of California Press.

Sakamoto, K.; Jerzak, C. T.; and Daoud, A. 2025. A Scoping Review of Earth Observation and Machine Learning for Causal Inference: Implications for the Geography of Poverty. In Hall, O.; and Wahab, I., eds., *Geography of Poverty*.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press.

Shu, D.; and Yi, G. Y. 2019. Causal Inference with Measurement Error in Outcomes: Bias Analysis and Estimation Methods. *Statistical methods in medical research*, 28(7): 2049–2068.

Stewart, A. J.; Lehmann, N.; Corley, I. A.; Wang, Y.; Chang, Y.-C.; Braham, N. A. A.; Sehgal, S.; Robinson, C.; and Banerjee, A. 2023. SSL4EO-L: Datasets and Foundation Models for Landsat Imagery. arXiv:2306.09424.

Ting, Y.-S. 2024a. Why Machine Learning Models Systematically Underestimate Extreme Values. *arXiv preprint arXiv:2412.05806*.

Ting, Y.-S. 2024b. Why Machine Learning Models Systematically Underestimate Extreme Values. arXiv:2412.05806.

Zhu, F. W.; Jerzak, C. T.; and Daoud, A. 2025. Optimizing Multi-Scale Representations to Detect Effect Heterogeneity Using Earth Observation and Computer Vision: Applications to Two Anti-Poverty RCTs. In Huang, B.; and Drton, M., eds., *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, 894–919. PMLR.