Remote Control: Debiasing Machine Learning Predictions for Causal Inference

By Matthew Gordon, Megan Ayers, Eliana Stone, and Luke Sanford*

Draft: January 3, 2026

ABSTRACT

Advances in machine learning and the increasing availability of high-dimensional data have led to the proliferation of social science research that uses the predictions of machine learning models as proxies for measures of human activity or environmental outcomes. However, prediction errors can lead to bias when estimating regression coefficients. In this paper, we show how this bias can arise, and demonstrate the use of an adversarial machine learning algorithm in order to debias predictions. These methods are applicable to any setting where machine learned predictions are the dependent variable in a regression. We conduct simulations and empirical exercises using ground-truth and satellite data on forest cover in Africa. Using the predictions from a standard machine learning model leads to biased parameter estimates, while the predictions from the adversarial model give precise estimates of the true effects. Finally, we replicate a study of the effects of artisanal gold mining on deforestation in Africa (Girard, Molina-Millán and Vic, 2025). We find that after correcting for bias using a novel sample of hand-labeled points, standard confidence intervals can not rule out a null effect, even though our confidence intervals are 19% smaller than those obtained using alternative bias correction methods.

* Thanks to Eli Fenichel, Philipp Ketz, the Berkeley Political Methodology Seminar, the LSE/Imperial College Workshop, and participants at the AGU, ICLR, TWEEDS, and AERE conferences for helpful feedback and support. We also thank Refine.ink for help with manuscript preparation. Gordon: Paris School of Economics; matthew.gordon@psemail.eu. Ayers: Yale Department of Statistics and Data Science; m.ayers@yale.edu. Stone: Yale School of the Environment; eliana.stone@yale.edu. Sanford: Yale School of the Environment; luke.sanford@yale.edu.

I. Introduction

Advances in machine learning and the increasing availability of satellite imagery and other high-dimensional datasets have led to the proliferation of social science research that uses the predictions of machine learning models as proxies for measures of human activity or environmental outcomes. However, when the machine learning models used to measure these outcomes minimize a standard loss function, the resulting predictions can produce biased estimates when estimating regression coefficients. If the measurement error in the outcome variable is correlated with policy variables or important confounders, as is the case for many widely used remote sensing data sets, estimates of the causal impacts of interventions will be biased. This bias can occur even in cases when researchers have a good instrument or an experimental research design.

In this paper, we show how this bias can arise, and we propose a method to generate unbiased predictions using adversarial debiasing algorithms (Zhang, Lemoine and Mitchell, 2018). We adapt this procedure from the algorithmic fairness literature, where it was originally developed to ensure that machine learning algorithms do not encode racial or other undesirable biases for decisions like hiring, admissions, or bail. The method uses a machine learning algorithm with a modified loss function that ensures that predictions are unbiased with respect to 'protected characteristics'. We directly borrow their approach show that it has broad applications to any setting in which researchers are using machine-learned outcome variables as the dependent variable in regressions.

We demonstrate the usefulness of the approach in the context of satellite derived measures of deforestation. Previous research has derived measures of economic output, air pollution, land-use change, and other variables at high resolution by using algorithms trained on satellite imagery and some ground-truth data (Henderson, Storeygard and Weil, 2011; Hansen et al., 2013; Meng et al., 2019).

A typical approach is to train a machine learning model on satellite data using a limited number of ground-truth observations and then using the model predictions to impute outcomes for a larger population of interest. These measurements have been widely used as a dependent variable in regressions to estimate the effects of various policies on deforestation, GDP, pollution, and other variables (see e.g. Burgess et al. 2012; Alix-Garcia et al. 2013; Meng et al. 2019; Asher, Garg and Novosad 2020; Wren-Lewis, Becerra-Valbuena and Houngbedji 2020; Slough 2021; Sanford 2021; Jack et al. 2022). For example, the Hansen et al. (2013) estimates of deforestation have been cited more than 10,000 times.

While machine learning models can obtain high accuracy, there are widely documented biases in the predictions that can pose problems when they are used in regressions. Tropek et al. (2014) showed that the Hansen et al. (2013) predictions confuse tree plantations for forests, for example. Various biases have been shown to exist in satellite measures of air pollution and economic activity as well Fowlie, Rubin and Walker (2019); Bluhm and McCord (2022). Although often ignored in early applications, this non-classical measurement error can violate the assumptions that are required for consistent estimation of regression coefficients.

We show analytically and intuitively why this can be a problem, starting with the well-known result that the bias in a regression coefficient depends on the covariance between the covariate and the machine learning model's prediction error. Crucially, this bias can occur even in a randomized control trial or quasi-experimental setting where standard assumptions for causal identification typically hold, since treatment can induce differential measurement error. We give a number of examples of how this can occur in practice.

Given the formula for the bias of a regression coefficient, a simple test for bias is to regress prediction errors on the independent variables of interest using a subsample of ground-truth data. This test can also be used to 'correct' regression

coefficients estimated using the full sample, a technique known as prediction-powered inference (PPI), or predict-then-debias (PTD) (Angelopoulos et al., 2023; Kluger et al., 2025). A power analysis of the regression can be used to determine how many observations researchers would need to label to detect bias of a given size. In many cases, researchers may be able to collect this ground-truth data by visually interpreting high-resolution imagery.

Our primary contribution is to then demonstrate how the use of machine learning models with modified loss functions can eliminate biases in coefficient estimates. Intuitively adversarial debiasing can be understood as follows — a primary model attempts to minimize prediction error for the outcome of interest. The measurement errors are then passed to a secondary model (the adversary), that tries to predict the treatment status of an observation. When tuning the first model, a penalty term is added to the loss function that increases if the adversary's predictions improve. Thus the primary model attempts to minimize prediction error while also making errors uninformative about treatment status. In the special case where the adversary is a linear regression of prediction errors on regression covariates, we show that this loss function penalizes the covariance between prediction errors and the treatment variable. This can reduce or eliminate biases in regression coefficients estimated using these predictions.

We then demonstrate the effectiveness of these approaches by applying them to measurements of forest loss in Africa. We conduct simulations and a simple descriptive exercise to measure the cross-sectional relationship between forest cover and distance to the nearest road, using ground-truth data on forest cover (Bastin, 2017). In our simulations, the adversarial debiasing approach allows us to estimate the true parameter without requiring any knowledge of the sources of prediction error, while standard machine learning model predictions lead to biased estimates.

We then turn to a setting with a causal research design, replicating a study of the effects of artisanal gold-mining on deforestation Girard, Molina-Millán and Vic (2025). Given the previously found tradeoffs between human health and ecological outcomes (Benshaul-Tolonen, 2019), careful measurement is critical in this context. Unfortunately there are not good publicly available sources of labeled forest data that show changes over time. We therefore generate a novel dataset of ground-truth deforestation data in Africa using optimal sampling methods to improve efficiency (Neyman, 1934; Gordon and Papp, Forthcoming; Zrnic and Candès, 2024). Using this data to correct for biases enlarges the confidence intervals in Girard, Molina-Millán and Vic (2025). Ultimately, we cannot rule out a null effect, though the confidence intervals produced using the debiased model are 19% shorter than those produce using the prediction-powered-inference method (Kluger et al., 2025).

This paper contributes to a growing literature that has begun to document the problem of non-random measurement error in machine-learning models (see Jain 2020 for a review[1]). A number of new papers propose econometric estimators that can correct for the non-classical measurement error in some cases (Zhang, 2021; Fong and Grimmer, 2021; Ratledge et al., 2021; Proctor, Carleton and Sum, 2023; Angelopoulos et al., 2023; Torchiana et al., 2023; Kluger et al., 2025; Rambachan, Singh and Viviano, 2025; Carlson and Dell, 2025). Most of these methods use a subset of data where the ground-truth is known to debias predictions[2]. Our method has two key advantages: 1) unlike some of the above methods, we require no assumptions about sources of bias or functional forms beyond those standard for causal inference, and 2) rather than correcting a biased set of predictions, we show how to generate predictions without biases. While more demanding,

[1]For topic specific reviews, see Balboni et al. 2022 on deforestation, Gibson et al. 2021 and Bluhm and McCord 2022 on night lights, and Fowlie, Rubin and Walker 2019 on air pollution.

[2]Torchiana et al. (2023) is an exception, they make assumptions about the data generating process that avoid the need for labeled data.

in that it requires researchers build a custom machine-learning model for any given analysis, this method is widely applicable, and it also improves efficiency in many of our experiments. We show the conditions under which predictions from a debiased model will result in smaller confidence intervals than alternative methods.

The ability to build customized machine learning models for outcomes of interest may become increasingly important as researchers learn more about the shortcomings of off-the-shelf remotely sensed data products in certain contexts. Our results show that very simple machine learning models can be sufficient to obtain consistent parameter estimates, as long as the prediction errors are balanced with respect to the policy variable. While we are not the first to customize machine learning predictions for use in social science research (Ratledge et al., 2021), our adversarial approach is both simpler and more general, in that it can be applied to attenuation bias as well as many other types of non-classical measurement errors. Both methods have applicability beyond satellite data. The same approach can be applied to machine learning predictions on text data, such patents or tweets for example.

Finally, we build connections between previous work on machine learning measurement error described above, and the literature on algorithmic bias and adversarial debiasing (Kleinberg, Mullainathan and Raghavan, 2016; Zhang, Lemoine and Mitchell, 2018; Kleinberg et al., 2018; Kim et al., 2022; Liang, Lu and Mu, 2023; Arnold, Dobbie and Hull, 2024). We directly adapt some of the results regarding algorithmic bias in decision making to solve a common estimation problem. Closely related to our work Chernozhukov et al. (2020) studied the use of an adversarial model to debiased estimates of heterogeneous treatment effects, when the treatment effect is modeled as a function of observables. Our paper shows how to use an adversarial model to debias measurement error in an outcome vari-

able that is not observed, but can be predicted as a function of observables. The resulting predictions can then be used to estimate treatment effects.

In the next section, we present an analytical framework that shows why machine learned predictions can result in biased estimates of regression coefficients, and we show how bias correction approaches and adversarial debiasing can solve these problems. In section III we show that the methods work with simulated data and in a simple cross-sectional descriptive regression to recover the relationship between roads and forest cover. In IV we apply these methods to a time-series, causal application: the effect of a gold mining on deforestation in Africa.

II. Framework

To fix ideas, consider a regression with one independent variable X[3]. We seek to estimate the relationship between changes in X and an outcome Y according to the linear model:

$$Y_i = \alpha + \beta X_i + e_i. \tag{1}$$

In this case, our parameter of interest is β, the marginal effect of X on Y. However we do not have access to the true Y_i for all observations. Instead we have predictions $\widehat{Y_i}$ from a machine learning model that generates predictions $\widehat{Y_i} = Y_i + \nu_i$, where ν_i is the measurement error for a given observation.

These predictions are generated based on a subset of the data for which the true values of Y_i are known. Let $j \in J$ index observations in this labeled data. While a researcher could estimate equation (1) using only labeled data, using the predictions from the unlabeled data can add power (Proctor, Carleton and Sum, 2023; Carlson and Dell, 2025).

The predictions in the unlabeled data are generated as $\widehat{Y_i} = f(k_i, \omega)$, where

[3]This section presents an expanded and refined version of the ideas in Sanford et al. (2025)

k_i are some predictors of Y_i that are observed for all i. Model weights ω are chosen by minimizing some function of ν_j, for example, the sum of squared errors: $\sum_j \nu_j^2 = \sum_j (\widehat{Y}_j(k_j, \omega) - Y_j)^2$.

While this may be a sensible approach for minimizing prediction errors, it can generate bias when $\widehat{Y}$ is used as a proxy for Y in policy evaluation. When we estimate β from (1) using OLS with $\widehat{Y}_i$ as the dependent variable instead of Y_i, our estimate of β will equal in expectation:

$$\mathbb{E}[\widehat{\beta}] = \beta + \frac{Cov(e, X)}{Var(X)} + \frac{Cov(\nu, X)}{Var(X)}. \tag{2}$$

The middle term in the equation is standard endogeneity bias, and must equal zero in expectation to estimate the true parameter without bias. For what follows, we ignore this source of potential bias as it is not the focus of our analysis. The third term is the result of using machine learned proxies for Y_i instead of the true values, and is the focus of this paper. In words, the measurement error from our machine learning model will bias our estimate if it is correlated with the treatment variable.

This situation can arise in several ways. First, measurement error can be correlated with the true values, Y_i. Fowlie, Rubin and Walker (2019) show this type of measurement error in satellite derived estimates of air pollution. The satellite measures show attenuation bias at higher concentrations, underestimating true concentrations. This is depicted in a Directed-Acyclic-Graph (DAG) in Figure 1.b, which we have called outcome-induced bias. In this case, we can model the measurement error as a function of Y_i plus an idiosyncratic component ϵ_i:

$$\begin{aligned} \nu_i &= g(Y_i) + \epsilon_i \\ \nu_i &= g(\beta X_i + e_i) + \epsilon_i. \end{aligned} \tag{3}$$

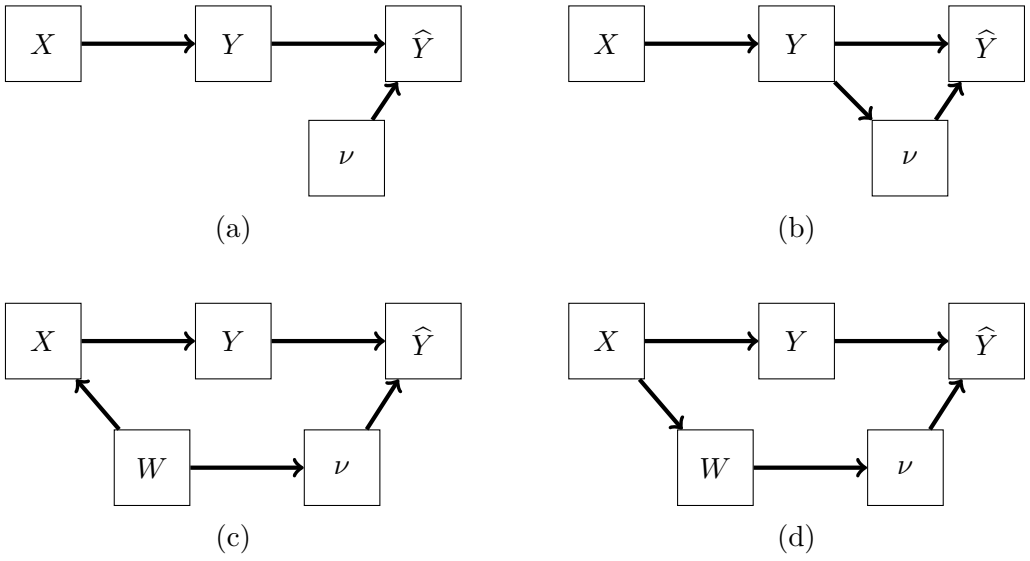


Figure 1. : Four Directed-Acyclic-Graphs illustrating potential relationships between treatment (X), outcomes (Y), measurement error (ν), machine learning model predictions ($\widehat{Y}$), and unobserved variables (W). (a) is classical measurement error, (b) shows outcome-induced bias, (c) shows confounder-induced bias, and (d) shows treatment-induced bias.

Thus, when $g(\beta X_i + e_i)$ covaries with respect to X_i, $\widehat{\beta}$ may be biased, even in an RCT setting where $Cov(e, X) = 0$. Intuitively, if $\widehat{Y}$ doesn't change over some domain of the ground truth variable, estimates of the treatment effect cannot learn about effects in that part of the distribution of Y. Another example of this type of measurement error occurs when Y is binary, since in this case, errors are always negatively correlated with the true value of Y (Aigner, 1973).

Alternatively, assume measurement error is a function of some other variable W_i that is correlated with both treatment and outcomes, and some idiosyncratic component. For example, if we want to estimate the effect of a payment-for-ecosystem-services program on deforestation, some unobserved W (e.g. soil moisture) might both affect the probability that a parcel is enrolled in the program and cause forest cover to be over-estimated because of the increased ambient veg-

etation — in other words there is selection into treatment that is correlated with measurement error. See Figure 1.c for a graphical representation. In this case, measurement error can be modelled as a function of W:

$$\nu_i = g(W_i) + \epsilon_i. \tag{4}$$

If $Cov(X, g(W)) \neq 0$, again our estimates will be biased. Note W does not operate like a traditional confounder here, in that it does not affect the true values of Y, except through it's effect on X. It only affects $\widehat{Y}$ through its effect on ν.

Finally, in certain cases, treatment can induce measurement error. For example, take a researcher studying the effects of a cash transfer program on deforestation. The treatment causes recipients to invest in irrigation and high-yielding crops, which may be more often confused for forests than the previous landcover. This would result in overestimation of the post-treatment forest cover for the treated group. We refer to this as treatment-induced bias. See Figure 1.d for a graphical representation.

These examples show that even if researchers have an experimental or quasi-experimental source of exogenous variation, estimates may still be biased when the dependent variable contains prediction error. Instrumental variables can be useful for measurement error in X. Even a valid and relevant instrument will not guarantee an unbiased estimate when $\widehat{Y}$ contains measurement error, however.

A. Biased Predictions: Why do they occur?

If measurement error is systematic with respect to treatment, as in the cases described above, this raises the question of why the machine learning model didn't generate better measurements in the first place? Fong and Tyler (2021) claim that these types of errors are unlikely to occur with machine learning, since if the

measurement error correlates with X, it does so at the expense of predicting Y. Despite this claim, biased measurements can arise for several reasons.

As described above, machine learning models are typically trained by choosing a set of model parameters, ω, that minimize a loss function in a training data set. When choosing ω, a more complex model will better fit the training data, however, out of sample predictions will have greater variance. On the other hand, a simple model may be right on average, but biased in certain regions of the feature space. Navigating this bias-variance tradeoff is at the core of modern machine learning methods.

As a result, bias can arise for at least two reasons. First, limited or unrepresentative training data means that certain regions of the feature space can be given less weight in the training process. Similarly, if data in certain parts of the feature space contains less information about the target variable of interest, then that will have less influence on the ω. This could occur if some of the data is of poorer resolution, for example. Liang, Lu and Mu (2023) formalize these ideas, showing that unless an algorithm's inputs satisfy a particular type of balance, the algorithm faces a tradeoff between accuracy and fairness (equivalent to unbiasedness in our context). In some cases they show that adding the group variable (or treatment variable X in our case) can even increase biases in predictions.

B. Detecting and Correcting Bias

Unlike for omitted variable bias, an estimate of measurement error bias is directly obtainable if researchers have access to or can generate some ground-truth values of Y. This is the approach followed by the prediction-powered inference literature (Angelopoulos et al., 2023) and the predict-then-debias methods (Kluger et al., 2025). Consider the regression:

$$\nu_j = \gamma X_J + u_j, \tag{5}$$

with X as a vector of independent variables, and $\nu_j = \widehat{Y} - Y$ is the prediction error. Our estimate of γ will be $\widehat{\gamma} = (X'X)^{-1}X'\nu$. This is exactly the multivariate analog of the bias term in equation 2. This shows that a very simple regression coefficient can be used to estimate bias, under the assumption that the labeled set is representative of the population of interest.

In practice, researchers may be able to obtain a number of such labels by visually interpreting high-resolution satellite imagery, for example. In this case, it should be easy to make sure that J is representative of the broader population, in which case the estimator is consistent. In cases where the labelled data suffers from selection bias or is non-representative in some other way, selection on observables techniques using the machine learning model inputs (k) may be a promising approach (Imbens, 2004; Carlson and Dell, 2025; Rambachan, Singh and Viviano, 2025). Estimates of the standard errors of $\widehat{\gamma}$ can be used to test whether the bias is significantly different from zero, or to rule out biases greater than a certain size, though standard errors may need to be adjusted for spatial or serial correlation.

It is also simple to adapt standard power calculations to estimate a 'minimum detectable bias' given a certain number of observations, and an estimate of the standard error of γ. Researchers can then estimate the number of labeled observations which will likely be necessary to rule out some amount of measurement error bias. We demonstrate this procedure in our applications.

Estimating the bias in this way can be a useful diagnostic, but it can also be used to perform a 'bias correction' on estimates of $\widehat{\beta}$ from equation (2). This bias correction, first shown by Angelopoulos et al. (2023), is:

$$\widehat{\beta}_c = \widehat{\beta} - \widehat{\gamma}. \tag{6}$$

In expectation, $\widehat{\beta}_c$ is a consistent estimator for β when the labeled sample

is representative. An estimate of the standard error of $\widehat{\beta}_c$ can be produced by a normal approximation as described in Angelopoulos et al. (2023) and Kluger et al. (2025) or with a bootstrap procedure. Note that $\widehat{\beta}$ is estimated using all of the points, while $\widehat{\gamma}$ is estimated only using the labeled points.

The advantage of this approach is its relative simplicity. It can be implemented with off-the-shelf predictions and a small amount of labeled data, as long as the labeled data is representative. There are some shortcomings of this estimator, however. The estimation uncertainty around $\widehat{\gamma}$ inflates the standard errors around $\widehat{\beta}_c$. Furthermore, this approach takes a set of $\widehat{Y}$ predictions as given. The joint distributions of $\widehat{Y}, Y$, and X potentially limit the precision of $\widehat{\beta}_c$. In the next section, we describe how to obtain better predictions of $\widehat{Y}$ for a given estimation problem.

C. Adversarial Debiasing

Recall that the machine learning model predictions are a function of k, input features, and ω, model weights, that are chosen to minimize some loss function. If the model is a linear regression, for example, then the model weights are the regression coefficients. Given the potential bias from prediction errors in equation (2), we can formulate the model's objective function as a constrained optimization problem. In particular, we seek a model that directly avoids differential measurement error by construction with parameters ω^* that satisfy:

$$\begin{aligned} \omega^* &= \arg\min_{\omega} L_p(\widehat{Y}(\omega), Y, k) \\ &\text{such that } Cov(X, Y - \widehat{Y}(\omega)) = 0. \end{aligned} \tag{7}$$

where L_p is a standard loss function, such as mean squared error. The constraint on the loss function ensures that the measurement errors will not be biased with respect to X.

This objective function is nearly isomorphic to that of the adversarial debiasing approach proposed by Zhang, Lemoine and Mitchell (2018), which was proposed as a way to debias machine learning model predictions with respect to race or gender. Zhang, Lemoine and Mitchell (2018) introduce a secondary model, called the adversary, with model weights γ and loss function $L_a(\widehat{X}(\gamma), X, Y, \widehat{Y}(\omega))$. The adversary chooses γ to try to predict a 'protected' variable using the measurement errors from the primary model. In Zhang, Lemoine and Mitchell (2018), the protected variable is race or gender; in our setup, the protected variable is X, an observation's treatment status.

An adversarial debiasing model is trained to minimize L_p, while maximizing L_a, subject to the adversary choosing γ in such a way as to minimize L_a. Formally, this can be written as:

$$\min_{\omega} \max_{\gamma} \left\{ L_p\left(\hat{Y}(\omega), Y, k\right) - \alpha \cdot L_a\left(X, \hat{X}\left(\gamma\right), Y, \hat{Y}(\omega), k\right) \right\} \quad \text{such that } \gamma \in \operatorname{argmin}\ L_a(X, Y, \widehat{Y}(\omega), \gamma, k) \tag{8}$$

where α is a researcher-specified parameter that controls the weight on the adversary's loss function, and must be tuned (e.g. by cross fitting). Intuitively, by maximizing the adversary's loss function, the primary model tries to make sure that the measurement errors contain as little 'information' as possible about X. When linear regression is used as the downstream estimation model, we are specifically concerned with 'information' in the form of the covariance between X and ν. Now consider an adversary model that is a linear model of the exact form of the bias test above in equation (5). The adversary loss function is the mean squared prediction error:

$$L_a = \frac{1}{N} \sum_j (\nu_j - \gamma X_j)^2.$$

The loss function is of this adversary is minimized with respect to γ when $\gamma = (X'X)^{-1}X'\nu$, which is exactly the bias in equation (2). The primary model will try to choose ω such that the prediction errors maximize the adversary's loss function.

Consider prediction errors at step t and $t+1$ of training: ν_t and ν_{t+1}. Holding accuracy of the primary model predictions constant, moving in the opposite direction of the adversarial loss gradient means $|\gamma_{t+1}| < |\gamma_t|$ (Proof in Appendix A). This can easily be applied to versions of equation (1) control variables or instrumental variables as well by using the Frisch-Waugh-Lovell theorem (see details in Appendix B).

Furthermore, this is exactly isomorphic to the constrained optimization problem in equation (7), if α is equal to the Lagrangian multiplier on the equality constraint. In practice, training a model with a constraint on the loss function can be difficult, which is why Zhang, Lemoine and Mitchell (2018) preferred a user specified α, chosen by cross fitting.

In practice, we find a similar approach that balances prediction accuracy with approximating the constraint in equation (7) is to penalize the covariance of X and ν directly in the loss function:

$$\min_{\omega} L_p(\hat{Y}(\omega), Y, k) - \alpha\Big|\sum_j (x_j - \bar{x_j})\nu_j\Big|. \tag{9}$$

For all methods, the choice of α is important. With too low of an α the model does not effectively debias the results, minimizing squared prediction error instead of maximizing the adversarial loss. However, when α is too high the model may produce random measurements to inflate the adversary's loss. While ideally α would be the Lagrangian multiplier on the constraint in equation (7), the typical approach is to use cross fitting within the labelled data, such that

overall prediction error and bias can be examined for different choices of α.

One downside of the approach detailed here is that it requires a unique machine learning model for each downstream estimation task. This suggests that, for example, measuring tree cover to estimate the effect of property rights on deforestation is different from measuring tree cover to estimate the effect of wealth shocks on deforestation. While this may be taxing for researchers, it also suggests that failing to build a measurement strategy for any individual task risks biasing that task.

Furthermore, Zhang, Lemoine and Mitchell (2018), show that the adversary is only guaranteed to produced unbiased measurements under strong assumptions on the loss function. Therefore in practice we recommend combining adversarial debiasing with a post-prediction method of debiasing, e.g. Angelopoulos et al. (2023) or Kluger et al. (2025).

For the researchers that are willing to build these models, adversarial debiasing has a few advantages. It does not require a researcher to know the source of the measurement error – the debiasing procedure will eliminate differential measurement error without specifying the precise source. Secondly, researchers may be able to use more sophisticated adversaries than a linear regression, for example, modeling ν non-parametrically as a function of both X and k. This may improve out-of-sample performance when representative training data is unavailable.

Secondly, adversarial debiasing may allow researchers to achieve unbiased estimates of treatment effects using very simple machine learning models. State of the art models that aim to maximize accuracy can be computationally demanding especially when used over large areas. In our examples below, we are able to recover accurate treatment effects using simple models that train in seconds on a standard CPU.

Finally, as we discuss in the next section, the main advantage of this approach

is that it may improve the precision of estimates of β in equation (1).

Inference and Efficiency

Call $\widehat{Y}_D$ the predictions of a model trained using one of the adversarial approaches outlined in Section II.C, ν_D the corresponding prediction errors, and $\widehat{\beta}_D$ the estimate of β using these predictions as a proxy for Y in equation (1). We wish to show under what conditions is $\text{Var}(\widehat{\beta}_D) < \text{Var}(\widehat{\beta}_C)$, the variance of the predict-then-debias estimator.

Recall e are the structural errors in equation (1), ν are the prediction errors from the standard machine learning model which can be decomposed into a systematic bias term γX, and a residual u that is uncorrelated with X by construction. Given these definitions, the variance of $\widehat{\beta}_D$ will be:

$$
\begin{aligned}
(10)\quad \text{Var}(\widehat{\beta}_D) &= \text{Var}\left((X'X)^{-1}X'\widehat{Y}_D\right) \\
&= (X'X)^{-1}X'\mathbb{E}[ee']X(X'X)^{-1} + (X'X)^{-1}X'\mathbb{E}[\nu_D\nu_D']X(X'X)^{-1} \\
&\quad + 2(X'X)^{-1}X'\mathbb{E}[e\nu_D']X(X'X)^{-1}
\end{aligned}
$$

In contrast, the variance of the predict-then-debias estimator from equation (6) is:

$$
\begin{aligned}
(11)\quad \text{Var}(\widehat{\beta}_c) &= \text{Var}\left(\widehat{\beta} - \widehat{\gamma}\right) \\
&= (X'X)^{-1}X'\left(\mathbb{E}[ee'] + \mathbb{E}[\nu\nu'] + 2\mathbb{E}[e\nu']\right)X(X'X)^{-1} + \\
&\quad (X_J'X_J)^{-1}X_J'\mathbb{E}[uu']X_J(X_J'X_J)^{-1} - 2(X'X)^{-1}X'\mathbb{E}[eu']X_J(X_J'X_J)^{-1} \\
&\quad - 2(X'X)^{-1}X'\mathbb{E}[\nu u']X_J(X_J'X_J)^{-1}
\end{aligned}
$$

where X_J is the matrix of covariates for the labeled points, with zeros in rows corresponding to unlabeled points. Thus it contains a subset of the rows of X.

While it could be possible that the structural errors e are correlated with the prediction error residuals u, such that $\mathbb{E}[eu'] \neq 0$ and $\mathbb{E}[e\nu_D'] \neq 0$, in the case where X is exogenously assigned, we believe these terms are likely small in practice and dominated by the other sources of error. Assuming one is likely not much greater than the other, we ignore them. Conditional on that assumption, and that J is a representative subset of the population, we can plug in equation (5) for ν and simplify to see that $\text{Var}(\widehat{\beta}_D) < \text{Var}(\widehat{\beta}_C) \iff$:

$$(12) \quad (X'X)^{-1}X'\mathbb{E}[\nu_D\nu_D']X(X'X)^{-1} < (X'X)^{-1}X'\mathbb{E}[uu']X(X'X)^{-1} + \left[(X_J'X_J)^{-1}X_J' - 2(X'X)^{-1}X'\right]\mathbb{E}[uu']X_J(X_J'X_J)^{-1}$$

We can make a few observations. All else equal, more accurate predictions will result in smaller standard errors for either estimator. We should generally expect $\mathbb{E}[\nu_D\nu_D'] > \mathbb{E}[uu']$, given that u comes from a model that is minimized with no constraints. This is not always the case in practice, however, as we will see in the following section.

Secondly, X_J is a subset of the rows of X. When X is univariate, $(X'X)^{-1} = 1/\sum_i X_i^2$, which is strictly less than $(X_J'X_J)^{-1} = 1/\sum_j X_j^2$. Thus when the labeled set is small relative to the total sample, we expect the second term on the right hand side to be large and positive. This second term represents the additional variance coming from the fact that the bias in the predictions is estimated using a small sample. Note that when $X_J = X$, the whole RHS of this inequality collapses to zero, the only remaining source of variance in $\widehat{\beta}_C$ comes from the structural error e. This makes sense, since in that case we observe the true values of Y for the whole population.

We can use the typical OLS standard errors or heteroskedasticity-consistent standard errors to estimate the variance of $\widehat{\beta}$ and $\widehat{\beta}_D$ under-sampling uncertainty

after making predictions. However, this practice implicitly assumes that the machine learning model's predictions are fixed — it doesn't take into account uncertainty from the prediction model itself. In essence ω (and γ in the adversarial models) can be considered random variables that might have a different realization if our training data came from a different sample. This would lead to a different set of prediction errors, and a different estimate of $\widehat{\beta}$.

Most studies using machine-learning measurements as outcome variables do not account for this type of model uncertainty in their standard errors. One approach to correct for this uncertainty is to bootstrap the entire estimation procedure, including training of the machine learning model, drawing different samples from the training set to train the model, and then to estimate of β. This approach is computationally intensive, since it requires bootstrapping the training of a potentially complex machine learning model, however, our simulations show that it can be important in practice.

Lastly, Kluger et al. (2025) show that a convex combination of the coefficient generated using the predicted values, and an estimate of the coefficient based on only the true values of Y estimated in the labeled ground-truth data can reduce the variance of the resulting estimate. We use their 'tuning' procedure to obtain all of our estimates in our applications.

In the sections that follow, we use hand-labeled datasets of forest cover in Africa to explore the biases generated by machine learning measurement error. We start with a simulation study of the cross-sectional relationship between roads and forest cover, and then estimate the relationship with real data. In section IV, we apply the above techniques to panel data to determine how important these considerations are in a setting with a causal research design. We use these methods to replicate the Girard, Molina-Millán and Vic (2025) study of the African gold mining boom on local forest cover.

III. Cross Sectional Simulations: Roads and Deforestation

To demonstrate the efficacy of our approach, we conduct simulations and empirical exercises on a common remotely-sensed outcome – forest cover – in settings where we also have access to ground truth data. For our first application, we use a hand-labeled dataset of 20,621 points in West Africa (Bastin, 2017) that are labeled with their percent tree cover as our "true" measures of forest cover. This data is a cross-sectional sample of points from a grid covering an enormous region of dryland forest across much of West Africa. This data was collected in part to show the biases of the Hansen et al. (2013) data in dryland areas. Researchers used high-resolution imagery from 2011-2015 to label the data. We use a sample of the data in West Africa within 30 km of a road and 100 km of a DHS cluster (see Figure 2 for the study area and an example of how the data was labeled).

As inputs to our machine learning models, we use data from the Landsat 7 ETM sensor. This sensor records the surface reflectance of light at several visible, near-infrared, and infrared wavelengths (called 'bands' in the remote sensing literature) at a 30-meter resolution. We generate three popular indices from these bands: Normalized Differenced Vegetation Index, Normalized Differenced Built Index, and Enhanced Vegetation Index. Over the course of a year, each location is observed up to 28 times (cloudiness obscures locations in some areas at some times). We take the 25th, 50th and 75th percentiles of each of the first five bands and three indices, and use those 24 variables as inputs. This feature-engineering strategy mirrors the approach in Hansen et al. (2013). With a simple 1-layer neural network (a logistic regression) we are able to predict forest cover using these variables with 75% accuracy.

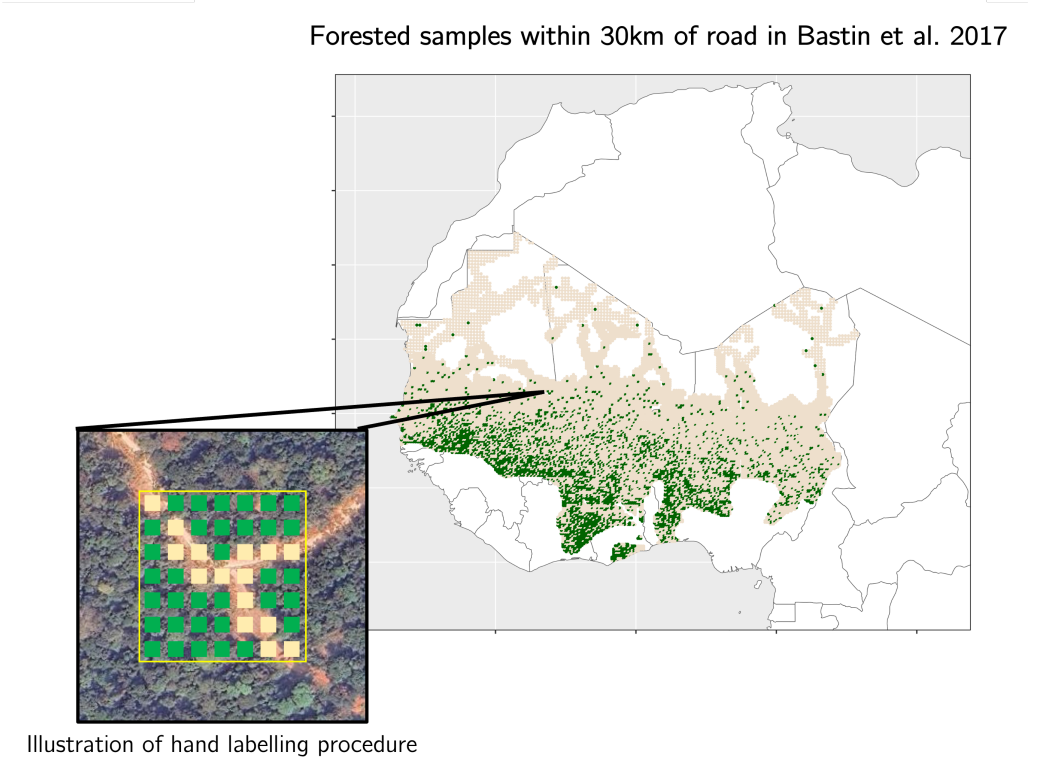


Figure 2. : Colored areas on the map show labeled pixels from Bastin (2017) — green for forested and beige for non-forested. Inset shows an example of how percent forested labels were generated from high-resolution satellite data.

A. Simulation

All of the following empirical exercises take the following structure. First, we divide the data into three equal folds. We train a model on two folds, and predict on the third for each of the folds so that we have a ground-truth value of Y and a prediction $\widehat{Y}$ for each point.

Then we repeat the procedure with a model with the same structure, but adding the constraint in equation (9). Finally, we estimate our regression of interest using the ground truth data, the baseline predictions, and adversarial model predictions.

We also test alternative bias correction methods, including multiple imputation (Proctor, Carleton and Sum, 2023), the tuned predict-then-debias (PTD) method (Kluger et al., 2025), prediction-powered inference (Angelopoulos et al., 2023), and PostPI (Wang, McCormick and Leek, 2020). For all approaches, we bootstrap standard errors than include the uncertainty from model training.

The regression of interest is the cross-sectional relationship between roads and forest cover. Previous work has found that roads are an important driver of deforestation (Asher, Garg and Novosad, 2020). Roads and other infrastructure are non-randomly placed, so this cross sectional relationship is likely to generate confounder-induced bias (Figure 1.c). Consider X to be the presence of a road, Y to be forest cover, and W to be some omitted geographic variable, like slope, that influences both measurement errors and the placement of roads.

Our first simulation follows this procedure:

1) Draw 20,000 observations of W from a Poisson distribution with shape parameter of 1.

2) Assign each observation $X \sim$ Bernoulli with $p = \max(1 - W/4, 0)$, so that treatment is more likely when W is lower.

3) Assign each observation a random forest cover Y and associated satellite data k from the Bastin points.

4) If $W > 0$, make the satellite data artifically 'greener' without changing the label. In practice this is done by replacing the satellite data with the satellite data from a different point with a higher percent forest cover.

5) We then train both a standard and debiased model on this simulated data as described above, and then we estimate a regression using the model's predictions as dependent variables.

The true treatment effect of X is zero, since the forest labels Y are assigned randomly. However, the last step mimics a real source of bias — remotely sensed forest cover tends to be overestimated on steeper slopes since images are taken from above and tend to capture more trees in a smaller spatial area when on a slope. Because of this bias, and selection into treatment, it will appear that roads are associated with lower forest cover. Note also that there are no "traditional" confounders in this simulation — nothing is associated with both road proximity and true forest cover.

Figure 3 shows the distribution of coefficient estimates from 100 bootstrapped runs of each of the models with $\alpha = 1$ and regressions using 10,000 training points and 10,000 unlabeled points. As predicted, using the baseline machine learning model results in a negative and significant estimate of the effect of X on Y. Both the bias correction methods and the adversarial model result in coefficient distributions correctly centered around 0.

The performance of all of these methods likely depends on the sample size of labeled points. Given this, we also run each of the models using a progressively increasing sample of labeled point. For each given sample size J, we use the J labels to train the model, and then predict on the remaining points so that the sample size for the regression is always $N = 20,000$. For each J we bootstrap 100 different versions of the model to estimate standard errors.

Figure 4 shows the results of this exercise. The baseline model generates biased estimates of the relationship between roads and forest cover across all sample sizes. The other methods, aside from PostPI, are centered on $\beta = 0$, and precision increases as the sample size of labeled points increases.

In this setting, PPI and PTD perform best. This situation is a difficult case for the adversarial debiasing method because the satellite data contains no information about the source of the bias, since it has been replaced by imagery from

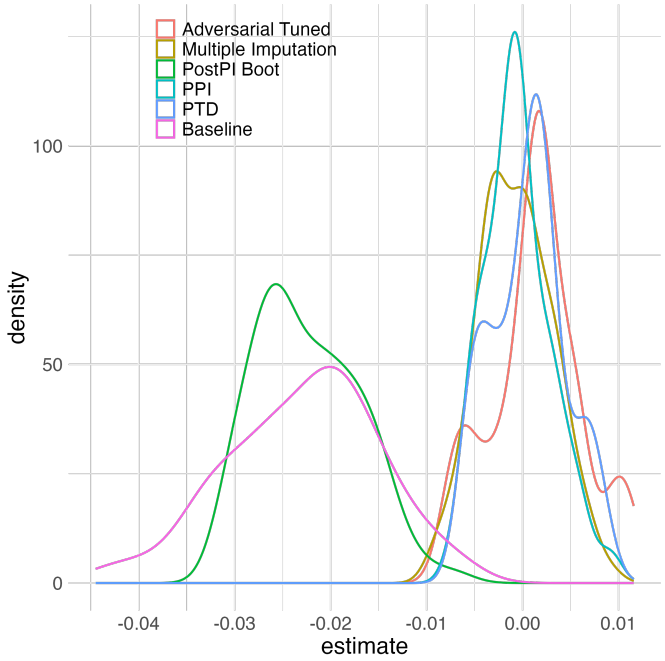


Figure 3. : Simulation experiment estimated distributions from our proposed adversarial model ("Adversarial Tuned"), a baseline neural network model, and alternative bias correction methods, each trained with 10,000 labeled observations. Each distribution shows the coefficients from each model after 100 runs on bootstrapped training data.

randomly drawn pixels with higher forest cover. This requires the adversarial model to sacrifice accuracy in predicting the low W, high Y observations so that the measurement error is balanced across X. Note that the adversary never has access to W, yet is still able to adjust for W-induced measurement error. In real-world cases, the adversarial model may be able to outperform the PTD and PPI methods by learning measurable representations that predict bias and adjusting for them.

Finally, we conduct a power analyses of the bias test in equation (5) to estimate the minimum detectable bias (MDB) at different sample sizes. This could be crucial for a researcher using an off-the-shelf satellite data source deciding how

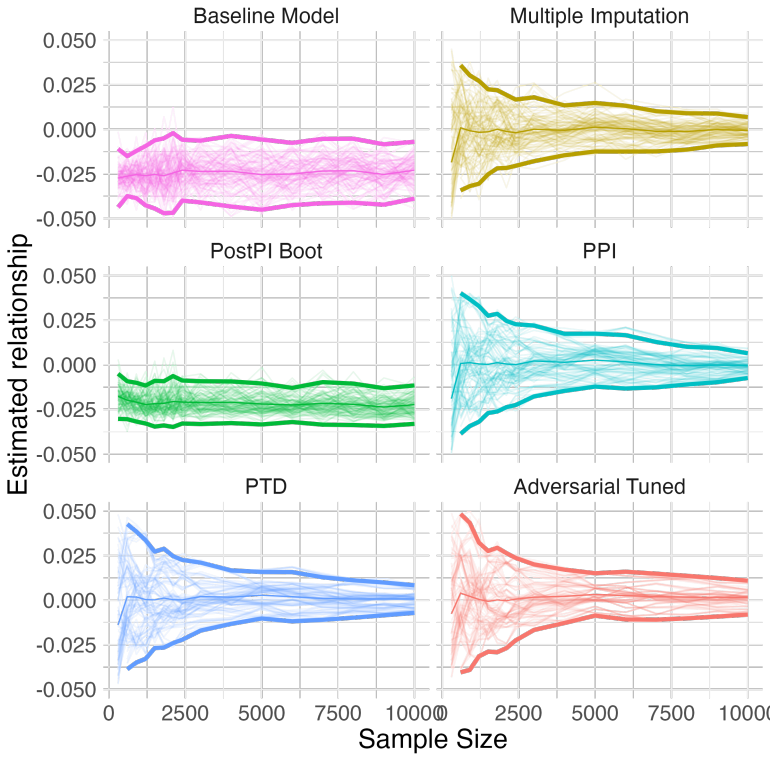


Figure 4. : Estimates from our proposed adversarial model ("Adversarial Tuned"), a baseline neural network model, and competing methods across training sample sizes. Each light colored line is an individual training run where researchers label progressively more observations. The thick lines represent the mean and 2σ intervals.

many points to label in order to rule out large biases in their estimates. The results are presented in Figure 5. Each line represents a different random draw of J labeled points. At each sample size, for each set of points, we estimate the standard error of γ and use that to perform a standard power calculation using 0.8 power and 95% statistical significance. Given that the true magnitude of the bias is 0.025 in this simulation, researchers would need to label more than 2,500 points to detect this bias as statistically different from zero 80% of the time.

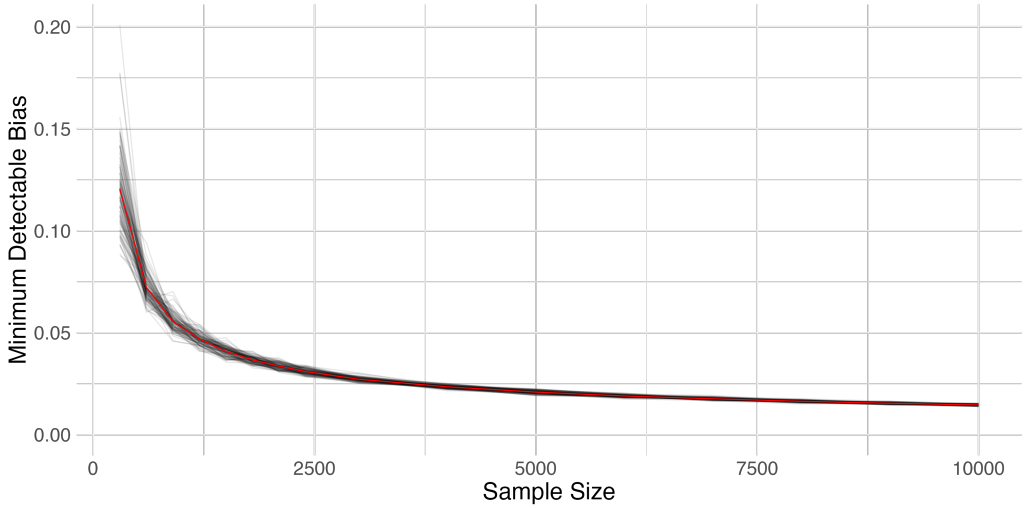


Figure 5. : Minimum detectable bias (MDB) estimates across sample sizes at power of 0.8 and $\alpha = 0.05$. Each black line represents an estimate of MDB using a different random sample of labeled data, and the red line is the true MDB using standard errors estimated with the whole dataset.

B. Descriptive Exercise: Roads and Forest Cover

Next we use the true Bastin (2017) data, and data on the African road network from Meijer et al. (2018), to estimate the gradient of forest cover with respect to distance to the nearest road. Whereas before the independent variable was binary and the outcome was continuous, now our independent variable is continuous, log distance to the nearest road, and our outcome is binary (forested or not). We apply the same cross-fitting procedure as above for both a standard model and an adversarial model using a 3-layer neural network that gives then estimated probability of forest cover as output. We then run the same regressions as in Section III.A. The results are shown in Figure 6.

In this real-world setting with non-manipulated satellite data, we see that the standard machine-learning model overestimates the negative relationship between proximity to roads and forest cover. Clearly, there are omitted variables in this

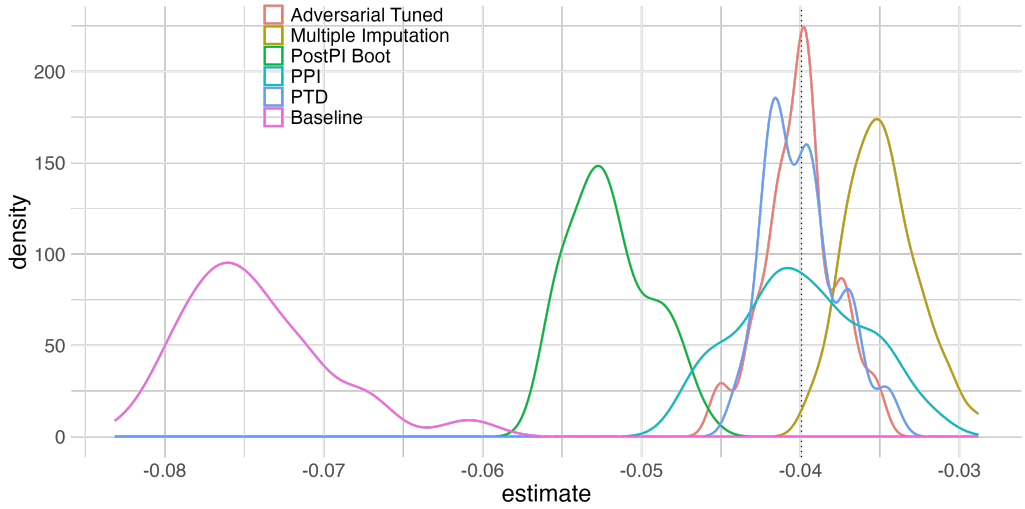


Figure 6. : Distributions of estimated effects of roads on forest cover from our proposed adversarial model ("Adversarial Tuned"), a baseline neural network model, and alternative methods, each trained with 10,000 labeled observations. Each distribution shows coefficients from each model after 100 runs on bootstrapped training data. The dashed vertical line represents the "true" estimate using all 20,000 ground truth labels.

context – topography, aridity, and others – that influence both the prediction errors and the location of roads. Once again, however, the adversarial model and some of the other proposed bias correction methods, including PTD and PPI, are able to generate measurements that recover accurate estimates without any knowledge of these omitted variables. The multiple imputation and PostPI methods reduce bias compared to the baseline machine learning model, but do not perform as well as the other methods. Furthermore, the estimates using the adversarial model have the greatest precision of all methods, though the PTD method is comparable.

In Figure 7 we plot the mean measurement error at each decile of distance from a road for both the adversarial model and the standard model. While the average error for both models is close to zero, the standard model exhibits a strong

measurement error-distance gradient that results in the biases in the coefficient estimates. The positive prediction error at close distances indicates the model is more likely to generate false positives (i.e., predict forest where there is no forest) close to roads, and the negative mean prediction error at far distances indicates the model tends to generate more false negatives at that distance. In contrast, the mean error is close to zero at every decile of distance for the adversarial model, indicating a balance of false positives and false negatives at all deciles of distance.

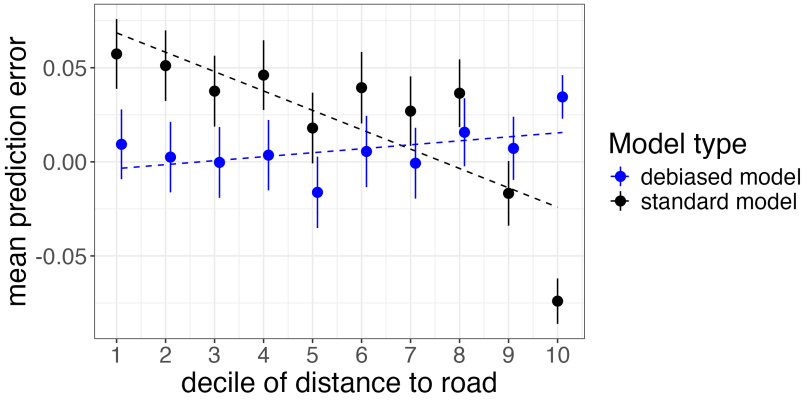
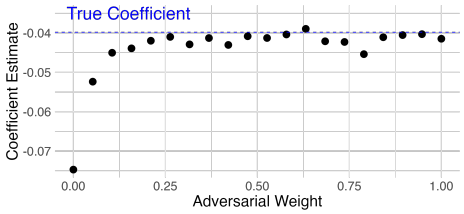


Figure 7. : Measurement error across deciles of distance to road

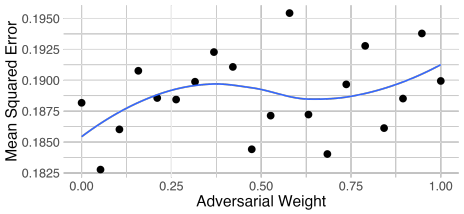
We also use this setting to run experiments on the α parameter: the weight on the adversary in the loss function. The results are summarized in Figure 8 for two different model architectures: a logistic regression and a deep neural net (DNN). The left column shows that, for both models, increasing the weight on the adversary from zero (standard model) to 1 quickly eliminates bias in the estimated coefficients.

Naively, we might think that this increasing weight would come at the cost of predictive accuracy, as described in the simulation, since an unconstrained model should be able to minimize MSE at least as well as a constrained model. This seems to be the case for logistic regression, which shows a trend towards higher

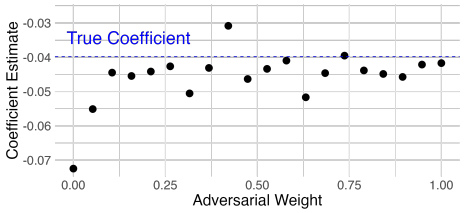
mean-squared error on predictions as the adversary's weight increases. For the DNN, however, increasing the adversary's weight actually improves prediction accuracy. This result can be understood as a kind of regularization effect. In some settings with many model parameters, for example when using LASSO with many features, it is well known that a regularization penalty can reduce overfitting and improve out-of-sample prediction performance (Tibshirani, 1996). While it is difficult to say precisely when adversarial models will improve prediction accuracy, it seems to have done so in this example.



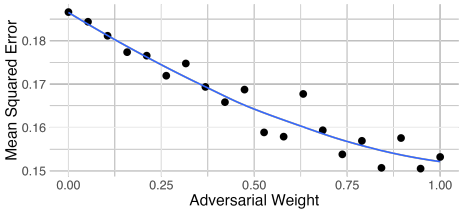
(a) Bias: Logistic Regression Primary Model



(b) Accuracy: Logistic Regression Primary Model



(c) Bias: DNN Primary Model



(d) Accuracy: DNN Primary Model

Figure 8. : Tuning alpha: Tradeoffs between bias and accuracy. Graphs show how coefficient bias (left column) and overall predictive accuracy (MSE - right column) change as the weight on the adversary increases. Top row shows results for a primary model that is a logistic regression. Bottom row shows results for a Deep Neural Net (DNN - 3 layers). Blue lines show Loess smoothed best fit curves.

This application demonstrates that estimates of a simple descriptive relationship can be biased by measurement error. These biases can be mitigated by

incorporating an adversarial debiaser into the model, however, potentially with greater precision compared to other proposed debiasing methods. In the following section, we study whether these methods remain important to study causal relationships using a research design that can rule out many possible confounders.

IV. Artisanal Gold Mining and Deforestation

We now turn to a causal relationship of interest - the effect of artisanal gold mining on deforestation in Africa. We conduct a replication analysis of Girard, Molina-Millán and Vic (2025), which found large effects of artisanal gold mining on deforestation in Africa. This is a topic in which precise measurement is important for policy, due to the sharp tradeoffs between environmental degradation and economic benefits. In the same paper, Girard, Molina-Millán and Vic (2025) find substantial impacts of gold price shocks on wealth in gold-suitable regions. In an earlier analysis, Benshaul-Tolonen (2019) found that increases in mining activity reduce infant mortality. At the same time, Africa is the continent with the highest rates of forest loss. Forest preservation may provide a range of benefits, including carbon capture and promoting biodiversity.

There are also theoretical mechanisms that push the effect of mining on forest cover in different directions. While presumably the direct effect of the mines is to displace some forest, in a context where agricultural expansion is the primary source of forest loss, it is possible that the indirect effects from the mines could reduce deforestation through other channels. Indeed, Foster and Rosenzweig (2003) find that increased economic growth increased forest cover in India.

A. Data Collection and Estimation Strategy

The main specification in Girard, Molina-Millán and Vic (2025) is:

$$\widehat{Y}_{cjt} = \beta_1 G_c \times P_{t-1} + \mu_c + \lambda_{jt} + \epsilon_{ct}, \tag{13}$$

Where $\widehat{\widehat{Y}}_{cjt}$ is the sum of predicted deforestation in a 0.5 x 0.5 degree grid cell c in country j at time t. G_c is a geologic measure of gold-mining suitability in the grid cell, and P_{t-1} is the annual international price of gold. They include grid cell and country-year fixed effects.

We use the replication package for Girard, Molina-Millán and Vic (2025) which contains their data on gold-mining suitability and international gold-prices, as well as their measure of deforestation, which is derived from Hansen et al. (2013)[4]. In particular, their dependent variable is the number of predicted deforested pixels in a grid cell-year, multiplied by 10,000. We replicate their estimates with this measure, but we also rescale this variable by the approximate number of pixels in a grid cell[5], so that effects can be interpreted as changes in the probability of a pixel being deforested.

As discussed, the Hansen et al. (2013) deforestation data is based on machine-learning model predictions about whether a given pixel has been deforested. To enable us to check these predictions for bias, and also train our own models, we randomly sample 50,000 pixels from across the continent of Africa. We extract the Hansen et al. (2013) predictions of deforestation for these pixels, and we also pull the corresponding Landsat 7 data, and use it to construct the same 24 features (k_i) as in the previous section[6].

Ground Truth Data: Optimal Sampling

To debias the Hansen et al. (2013) predictions, we need ground-truth data on deforestation for a representative sample of points. There is a lack of high-quality labeled time-series data on deforestation that creates a serious obstacle for

[4]Technically the Hansen et al. (2013) data measures tree-cover loss, not deforestation. The former may include tree losses due to natural causes as well as human disturbance. Keeping with the usage in Girard, Molina-Millán and Vic (2025), we use the terms interchangeably below.

[5]Given 0.5 degree resolution and 30 meter pixels, this comes out to 3.24 million pixels per cell at the equator, which we divide by 10,000 to correspond to their measure.

[6]The annual 25th, 50th, and 75th percentiles of the first five ETM bands, plus NDVI, NDBI, and EVI.

this type of research, and unfortunately, Hansen et al. (2013) do not make their training data publcily available. The Bastin (2017) data utilized in the previous section is cross-sectional and thus cannot be used for time-series applications. One source of time-series data is Guo, Zhu and Gong (2022). Their hand-labeled data tracks a panel of pixels globally from 2000-2020, and crucially, they over sample areas where they believe forest change has occurred. Unfortunately, they do not record the sample inclusion probabilities of their observations making it difficult to determine the representativeness of the sample without further assumptions. Furthermore, there are only 149 pixels on the continent of Africa in their dataset that are labeled as having been deforested during this time period, which severely limits the power of our bias tests and corrections.

To fill this void, we collect our own data on forest change using optimal sampling methods proposed by Gordon and Papp (Forthcoming) and Zrnic and Candès (2024). The basic idea is to first train a preliminary machine learning model on whatever non-representative ground-truth data is available, in our case, we use the Guo, Zhu and Gong (2022) data, and train a 3-layer LSTM model, which is a type of neural network that has been shown to work efficiently with time-series data (Hochreiter and Schmidhuber, 1997). The model takes the entire time series of features for a pixel, and returns the probability of change in each year.

Next we use this model, which, given the minimal training data, has fairly low out-of-sample-accuracy and is likely biased, to make a preliminary set of predictions on our representative sample of unlabeled points. We use these predictions in combination with Neyman (1934) optimal sampling methods to choose a representative set of these points to label. We sample points with probability $\pi_i \propto \sqrt{p_i(1-p_i)}$, where p_i is our LSTM model's predicted probability of deforestation for pixel i. This sampling strategy leads us to over sample points that the model is most uncertain about. The usefulness of this procedure comes from

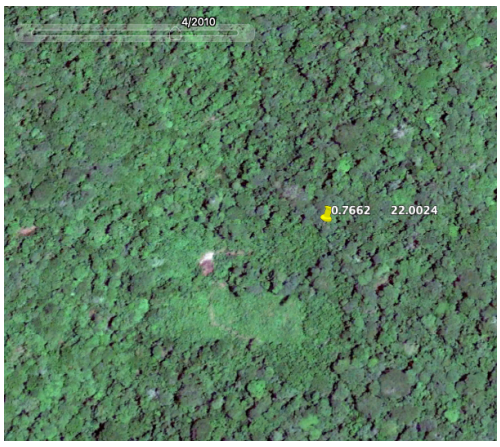
the fact that, even if the p_i are miscalibrated or biased, we will still end up with a representative sample (once labeled observations are weighted by $1/\pi_j$). Any biases in the preliminary model thus only affect the efficiency of our resulting estimates, not the bias.

In order to ensure that we have a sufficiently large sample of labeled points that undergo deforestation, we perform Neyman (1934) optimal sampling within two strata: the first contains the points that Hansen et al. (2013) predicts have been deforested, and the second contains points that are predicted to have no deforestation. We select 400 points from the predicted deforested points, and 1600 points from the much larger set of predicted no-change points. We then re-weight our final sample to be representative of the population.

One we have selected our $J = 2000$ pixels, we need to obtain ground truth measures of deforestation. For this, we turn to visual interpretation of historical high-resolution satellite imagery from Google Earth Pro (Google, 2025). For each point, we have two annotators independently inspect all available imagery, and mark the first year between 2000-2025 in which deforestation is observed. To mimic the structure of the Hansen et al. (2013) data, we define a forested pixel as a having 30% tree cover (vegetation $>$ 5m tall), and deforestation as the loss of more than half of pre-existing tree-cover. As with their data, we do not record multiple instances of deforestation in the same location or regrowth.

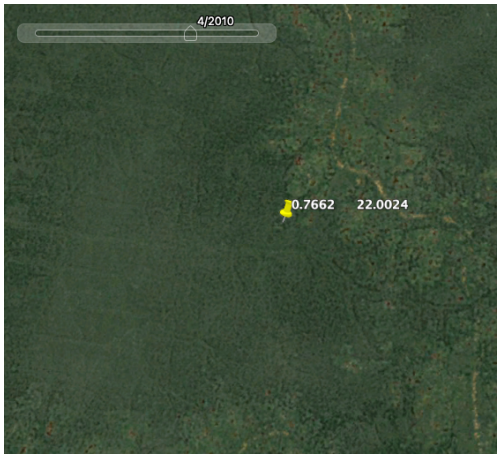
The quality and frequency of available imagery varies across locations, however all locations have, at minimum an annual Landsat composite image available at 30 meter resolution. While not without some difficulty, deforestation can be estimated at this scale by observing when a pixel in a forested region changes from dark green to another color — typically a lighter shade of green or brown. Figure 9 below shows a sample of this imagery for a point that has seen deforestation occur. The top two panels show high-resolution imagery from before and after the

event, while the lower two panels show the same point using the lower-resolution Landsat composite. Looking closely, one can see a light-green patch in the bottom right image corresponding to the deforested area.



(a) High Resolution Imagery: Before

(b) High Resolution Imagery: After



(c) Low Resolution Imagery: Before

(d) Low Resolution Imagery: After

Figure 9. : Google Earth Pro Imagery of Deforestation

Google Earth Pro historical imagery is used to generate ground-truth data on deforestation over time. The top two images show a point in the highest resolution available in 2010 and 2017. The bottom two images show the same point in the same years using the Landsat annual composite.

This process unavoidably introduces its own measurement error, especially in the earlier years of our study period when high-resolution data is less frequently available. It can be challenging in some cases to determine whether vegetation is over 5m, forest cover exceeds 30%, and whether half of existing forest cover is lost. We attempt to mitigate these issues by having a third annotator review the classification in cases where the first two annotators disagree. Despite its remaining shortcomings, we believe this data offers a substantial improvement over alternative sources of ground-truth data, and should prove valuable to researchers in stimulating further investigation when bias is detected.

We then use our labeled data to define Hansen et al. (2013) prediction errors ν, and we merge our labeled and unlabeled points to the grid cell-level data in Girard, Molina-Millán and Vic (2025). This allows us to estimate estimate a version of equation (13) with ν as the dependent variable in order to estimate bias in the causal effect of the gold price shock on deforestation in gold-suitable grid cells and debias the original estimates using prediction-powered inference. To compare approaches, we also use our newly labeled data to train a adversarially debiased machine learning model. We then use the model to make predictions on the remaining 48,000 unlabeled points and estimate equation (13) with these predictions as the dependent variable.

B. Results

Overall we find that the Hansen et al. (2013) is generally accurate, however we find slightly more deforestation. Figure 10 shows a map of our study area. Gold suitable areas are colored in yellow. We also overlay our labelled points. Note that our optimal sampling strategy leads us to oversample in the forested regions near the equator, which is where the most deforestation occurs. In contrast, we sample relatively fewer points in desert regions, where our model was fairly certain that

no deforestation had occurred. The points are colored by the Hansen et al. (2013) prediction errors — there appear to be clear geographic trends. Green points are false positives, points predicted to have deforestation that were not deforested. These seem to be concentrated in the most densely forested regions of central Africa. Red points show false negatives — points that saw deforestation but were missed by Hansen et al. (2013). These appear to be more common in dryland, less forested areas, which are also the areas that have seen the most deforestation. A plausible explanation for these biases is that the Hansen et al. (2013) model is better at identifying deforestation when a large chunk of previously pristine forest is cut down, but struggles in regions with lower baseline forest cover. Their model missing this type of tree loss may also explain why we find more deforestation overall.

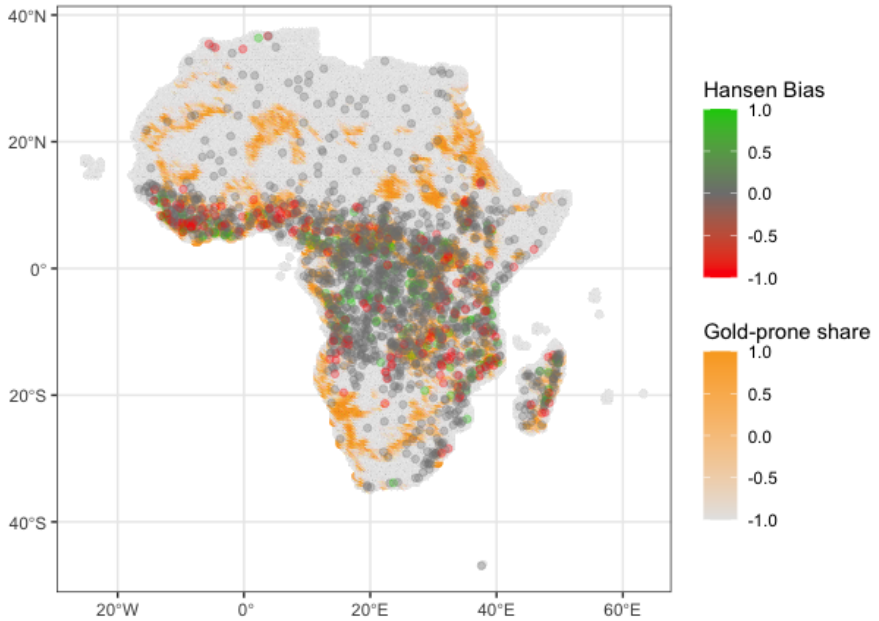


Figure 10. : Prediction Errors in the Hansen Data. Green points are false positives. Red points are false negatives. Yellow shading shows gold-suitability from Girard, Molina-Millán and Vic (2025).

Level effects in prediction error might not be a concern for the time-series application in Girard, Molina-Millán and Vic (2025). If the amount of prediction error is uncorrelated with changes in international gold prices, these differences could be controlled for by their fixed effects. If changes in deforestation generate prediction errors, however, then we have outcome-induced bias, illustrated in Figure 1.b.

Table 1 shows the results of our analysis. Column 1 replicates the exact result from Table 2 in Girard, Molina-Millán and Vic (2025) using the same data and variables. Column 2 uses the same data, but scales their dependent variable by the approximate number of pixels in a grid cell to make effects interpretable as changes in the probability of deforestation. Column 3 estimates the same regression, but at the pixel level, using our random sample of 50,000 unlabeled points. The magnitude is quite close to the estimate in Column 2, but slightly larger, possibly due to our approximation errors in the number of pixels per grid cell. Standard errors are also a bit larger than in Column 2, but the estimate remains very precise.

In Column 4, we estimate the same regression, but using measurement error as the dependent variable. We estimate this regression in the weighted labelled sample of 20,000 points. The coefficient on treatment, which is an estimate of the bias on the estimate in Column 3, is not significantly different from 0, but it is also not very precise, indicating that we cannot rule out large biases. A simple bias-correction can be performed by subtracting the point estimate in Column 4 from Column 3, which would give a point estimate for the effect on deforestation of 0.0005, which is 3.6 times lower than the estimate in Column 3. Finally, in Column 5, we train a debiased LSTM model on the labeled sample, with α chosen by cross fitting, and find a very precise, but much smaller effect of changes in the Girard, Molina-Millán and Vic (2025) treatment variable on deforestation. The

magnitude is reduced by 18 times relative to Column 3.

	$\widehat{\widehat{Y}}_g$ (1)	$\widehat{Y}_g$ (2)	$\widehat{Y}_i$ (3)	ν (4)	$\widehat{Y}_i^D$ (5)
gold suitable × price	0.3033***	9.362×10^{-4}***	0.0018***	0.0013	1.01×10^{-4}***
	(0.0448)	(1.384×10^{-4})	(4.109×10^{-4})	(0.0050)	(1.77×10^{-5})
FE: grid cell	X	X	X	X	X
FE: yearˆcountry	X	X	X	X	X
Num.Obs.	192 348	192 348	900 000	36 000	899 442
Kluger et al. (2025) 95% CI			[-0.0070, 0.0083]		[-0.0049, 0.0080]

Table 1—: Regression results of equation 13. Column 1 replicates an estimate from Table 2 in Girard, Molina-Millán and Vic (2025). Column 2 uses the same data, but converts the units of the dependent variable so that coefficients can be interpreted as changes in the probability of deforestation. Column 3 uses Hansen et al. (2013) predictions to estimate the same regression on a sample of 50,000 randomly chosen pixels. Column 4 estimates biases in the subsample of 2,000 labeled pixels. Column 5 estimates the same regression using predictions from our model trained using adversarial debiasing. Column 5 drops a few observations that are missing satellite data. All regressions contain grid cell and country-year fixed effects, and we cluster standard errors by grid cell following Girard, Molina-Millán and Vic (2025).

Finally, we use the prediction-powered-inference methods from Kluger et al. (2025) to optimally combine estimates from the ground-truth and predicted data. When we use the Hansen et al. (2013) predicted deforestation, as in Column 3, combined with the labeled subsample, we get a point estimate of the coefficient on the treatment variable of 0.00078 with a 95% confidence interval of [-0.0070, 0.0083]. When we use the same optimal tuning method, but with our debiased predictions rather than the Hansen et al. (2013) predictions, we get a point estimate of 0.00068 with a confidence interval of [-0.0049, 0.0080]. Both point estimates are close to each other, and both confidence intervals include zero, but they also include the original point estimate from using the Hansen et al. (2013) data.

The fact that the estimates in Column 5 are overly precise illustrates the value of post-prediction bias correction methods, and accounting for model uncertainty, even when adversarial debiasing is used. The confidence intervals constructed using prediction-powered inference with the debiased predictions are 19% smaller than the confidence intervals constructed using the Hansen et al. (2013) predic-

tions, however, illustrating the value of our method in improving efficiency.

Ultimately we cannot make firm conclusions about the effect of artisanal gold-mining, giving our small labeled sample. This exercise ultimately highlights the importance of collecting better quality labels on forest cover data to improve the precision of our estimates.

V. Conclusion

Advances in machine learning represent a tremendous opportunity for social science research. Satellite data now makes it possible to measure land use changes at unprecedented scale and resolution. Beyond satellite data, machine learning techniques can be used to measure difficult to quantify concepts from text and other unstructured data. As this field advances, however, researchers need to be alert to the possibility of non-classical measurement error generated by these techniques.

In this paper, we demonstrate how measurement error from machine learning algorithms can bias coefficient estimates. We also demonstrate several general and widely-applicable techniques to test for biases and correct these issues.

We demonstrate the usefulness of these techniques in several simulations and empirical exercises studying forest cover in Africa. We find that across applications, standard machine learning models produce measurements which bias the downstream estimation tasks, and that both the bias correction and adversarial debiasing methods are able to recover the true parameters estimated with ground-truth data.

In addition to the many practical applications of this technique, several theoretical questions present themselves for future work. In particular, the question of how to correct standard errors in regressions using machine learned proxies to account for model weight uncertainty appears to be very important. Progress

on these questions will allow researchers to be better prepared to exploit improvements in the availability of data and machine learning algorithms for causal research.

REFERENCES

Aigner, Dennis J. 1973. "Regression with a binary independent variable subject to errors of observation." *Journal of Econometrics*, 1(1): 49–59.

Alix-Garcia, Jennifer, Craig McIntosh, Katharine R. E. Sims, and Jarrod R. Welch. 2013. "The Ecological Footprint of Poverty Alleviation: Evidence from Mexico's Oportunidades Program." *The Review of Economics and Statistics*, 95(2): 417–435.

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-Powered Inference." arXiv:2301.09633.

Arnold, David, Will S. Dobbie, and Peter Hull. 2024. "Building Non-Discriminatory Algorithms in Selected Data."

Asher, Sam, Teevrat Garg, and Paul Novosad. 2020. "The Ecological Impact of Transportation Infrastructure." *The Economic Journal*, 130(629): 1173–1199.

Balboni, Claire, Aaron Berman, Robin Burgess, and Benjamin Olken. 2022. "The Economics of Tropical Deforestation." *Working Paper*.

Bastin, Jean-François et al. 2017. "The extent of forest in dryland biomes." *Science*, 356(6338): 635–638. Publisher: American Association for the Advancement of Science.

Benshaul-Tolonen, Anja. 2019. "Local Industrial Shocks and Infant Mortality." *The Economic Journal*, 129(620): 1561–1592.

Bluhm, Richard, and Gordon C. McCord. 2022. "What Can We Learn from Nighttime Lights for Small Geographies? Measurement Errors and Heterogeneous Elasticities." *Remote Sensing*, 14(5): 1190. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Burgess, Robin, Matthew Hansen, Benjamin A. Olken, Peter Potapov, and Stefanie Sieber. 2012. "The Political Economy of Deforestation in the Tropics*." *The Quarterly Journal of Economics*, 127(4): 1707–1754.

Carlson, Jacob, and Melissa Dell. 2025. "A Unifying Framework for Robust and Efficient Inference with Unstructured Data." arXiv:2505.00282 [econ].

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. 2020. "Adversarial Estimation of Riesz Representers." arXiv:2101.00009 [cs, econ, stat].

Fong, Christian, and Justin Grimmer. 2021. "Causal Inference with Latent Treatments." *American Journal of Political Science*, n/a(n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12649.

Fong, Christian, and Matthew Tyler. 2021. "Machine Learning Predictions as Regression Covariates." *Political Analysis*, 29(4): 467–484. Publisher: Cambridge University Press.

Foster, Andrew D., and Mark R. Rosenzweig. 2003. "Economic Growth and the Rise of Forests." *The Quarterly Journal of Economics*, 118(2): 601–637. Publisher: Oxford University Press.

Fowlie, Meredith, Edward Rubin, and Reed Walker. 2019. "Bringing Satellite-Based Air Quality Estimates Down to Earth." *AEA Papers and Proceedings*, 109: 283–288.

Gibson, John, Susan Olivia, Geua Boe-Gibson, and Chao Li. 2021. "Which night lights data should we use in economics, and where?" *Journal of Development Economics*, 149: 102602.

Girard, Victoire, Teresa Molina-Millán, and Guillaume Vic. 2025. "Artisanal mining in Africa. Green for Gold?" *The Economic Journal*, 135(672): 2578–2597.

Google. 2025. "Google Earth Pro."

Gordon, Matthew, and Anna Papp. Forthcoming. "Open Dumps: Measurement and Trade." *Working Paper*.

Guo, Jing, Zhiliang Zhu, and Peng Gong. 2022. "A global forest reference set with time series annual change information from 2000 to 2020." *International Journal of Remote Sensing*, 43(9): 3152–3162.

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, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science*, 342(6160): 850–853. Publisher: American Association for the Advancement of Science.

Henderson, Vernon, Adam Storeygard, and David N. Weil. 2011. "A Bright Idea for Measuring Economic Growth." *American Economic Review*, 101(3): 194–199.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation*, 9(8): 1735–1780.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics*, 86(1): 4–29.

Jack, B. Kelsey, Seema Jayachandran, Namrata Kala, and Rohini Pande. 2022. "Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning."

Jain, Meha. 2020. "The Benefits and Pitfalls of Using Satellite Data for Causal Inference." *Review of Environmental Economics and Policy*, 14(1): 157–169. Publisher: Oxford Academic.

Kim, Michael P., Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. 2022. "Universal adaptability: Target-independent inference that competes with propensity scoring." *Proceedings of the National Academy of Sciences*, 119(4): e2108097119. Publisher: Proceedings of the National Academy of Sciences.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. "Algorithmic Fairness." *AEA Papers and Proceedings*, 108: 22–27.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." arXiv:1609.05807 [cs, stat].

Kluger, Dan M., Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates. 2025. "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling."

Liang, Annie, Jay Lu, and Xiaosheng Mu. 2023. "Algorithm Design: A Fairness-Accuracy Frontier." arXiv:2112.09975 [econ].

Meijer, Johan R., Mark A. J. Huijbregts, Kees C. G. J. Schotten, and Aafke M. Schipper. 2018. "Global Patterns of Current and Future Road Infrastructure." *Environmental Research Letters*, 13(6): 064006.

Meng, Jun, Chi Li, Randall V. Martin, Aaron van Donkelaar, Perry Hystad, and Michael Brauer. 2019. "Estimated Long-Term (1981–2016) Concentrations of Ambient Fine Particulate Matter across North America from Chemical Transport Modeling, Satellite Remote Sensing, and Ground-Based Measurements." *Environmental Science & Technology*, 53(9): 5071–5079. Publisher: American Chemical Society.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society*, 97(4): 558.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum. 2023. "Parameter Recovery Using Remotely Sensed Variables." *NBER Working Paper*.

Rambachan, Ashesh, Rahul Singh, and Davide Viviano. 2025. "Program Evaluation with Remotely Sensed Outcomes." arXiv:2411.10959 [econ].

Ratledge, Nathan, Gabriel Cadamuro, Brandon De la Cuesta, Matthieu Stigler, and Marshall Burke. 2021. "Using Satellite Imagery and Machine Learning to Estimate the Livelihood Impact of Electricity Access."

Sanford, Luke. 2021. "Democratization, Elections, and Public Goods: The Evidence from Deforestation." *American Journal of Political Science*, n/a(n/a).

Sanford, Luke C., Megan Ayers, Matthew Gordon, and Eliana Stone. 2025. "Adversarial Debiasing for Unbiased Parameter Recovery." arXiv:2502.12323 [cs].

Slough, Tara et al. 2021. "Adoption of community monitoring improves common pool resource management across contexts." *Proceedings of the National Academy of Sciences*, 118(29): e2015367118. Publisher: Proceedings of the National Academy of Sciences.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288. Publisher: [Royal Statistical Society, Wiley].

Torchiana, Adrian L., Ted Rosenbaum, Paul T. Scott, and Eduardo Souza-Rodrigues. 2023. "Improving Estimates of Transitions from Satellite Data: A Hidden Markov Model Approach." *The Review of Economics and*

Statistics, 1–45.

Tropek, Robert, Ondřej Sedláček, Jan Beck, Petr Keil, Zuzana Musilová, Irena Šímová, and David Storch. 2014. "Comment on "High-resolution global maps of 21st-century forest cover change"." *Science*, 344(6187): 981–981. Publisher: American Association for the Advancement of Science.

Wang, Siruo, Tyler H. McCormick, and Jeffrey T. Leek. 2020. "Methods for Correcting Inference Based on Outcomes Predicted by Machine Learning." *Proceedings of the National Academy of Sciences*, 117(48): 30266–30275.

Wren-Lewis, Liam, Luis Becerra-Valbuena, and Kenneth Houngbedji. 2020. "Formalizing land rights can reduce forest loss: Experimental evidence from Benin." *Science Advances*, 6(26): eabb6914.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. "Mitigating Unwanted Biases with Adversarial Learning." arXiv:1801.07593 [cs].

Zhang, Han. 2021. "How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It." SocArXiv preprint.

Zrnic, Tijana, and Emmanuel J. Candès. 2024. "Active Statistical Inference." arXiv:2403.03208 [stat].

Appendix A. Proof that adversarial debiasing penalizes the absolute value of γ

Define $P = X(X'X)^{-1}X'$ as the n x n symmetric and idempotent projection matrix, and $\mathbb{I}$ as the n x n identity matrix. The Adversary's loss function is:

$$\begin{aligned} L_a &= (\nu - P\nu)'(\nu - P\nu) \\ &= v'(\mathbb{I} - P)'(\mathbb{I} - P)\nu = v'(\mathbb{I} - P)\nu \end{aligned} \tag{A1}$$

by the properties of the projection matrix. Take two different vectors of prediction errors, ν and $\tilde{\nu}$, such that $|\gamma| = |(X'X)^{-1}X'\nu| > |\tilde{\gamma}| = |(X'X)^{-1}X'\tilde{\nu}|$. The difference in L_a for these two vectors is:

$$\nu'\nu - \nu'P\nu - [\tilde{\nu}'\tilde{\nu} - \tilde{\nu}'P\tilde{\nu}]. \tag{A2}$$

Assume $\nu'\nu = \tilde{\nu}'\tilde{\nu}$, i.e. the overall prediction error is the same. Then we want that the primary model will choose $\tilde{\nu}$, since $|\tilde{\gamma}|$ is smaller. Since we are minimizing $L_p(\cdot) - \alpha L_a(\cdot)$, we want that $L_a(\nu) < L_a(\tilde{\nu}) \iff \tilde{\nu}'P\tilde{\nu} < \nu'P\nu$. We know:

$$\begin{aligned} |(X'X)^{-1}X'\nu| > |(X'X)^{-1}X'\tilde{\nu}| \iff \\ \nu'X(X'X)^{-1}(X'X)^{-1}X'\nu > \tilde{\nu}'X(X'X)^{-1}(X'X)^{-1}X'\tilde{\nu} \end{aligned} \tag{A3}$$

Since X is univariate, both sides are scalars. Multiply both sides by $X'X$ which is a positive scalar, maintaining the inequality:

$$\begin{aligned} \nu'X(X'X)^{-1}(X'X)(X'X)^{-1}X'\nu > \tilde{\nu}'X(X'X)^{-1}(X'X)(X'X)^{-1}X'\tilde{\nu} \\ \nu'P'P\nu > \tilde{\nu}'P'P\tilde{\nu} \\ \nu'P\nu > \tilde{\nu}'P\tilde{\nu} \end{aligned}$$

Concluding the proof.

Appendix B. Debiasing with Control Variables and Instruments

Adding Control Variables

Assume we want to estimate β_1 in the regression

$$\widehat{Y}_i = \beta_1 x_1 + x_2\beta_2 + e_i \tag{B1}$$

where x_1 is an n x 1 vector of the treatment variable, and x_2 is an n x k matrix of control variables. By the Frisch-Waugh-Lovell theorem, we can write $\widehat{\beta}_1$ as:

$$\widehat{\beta}_1 = (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{\widehat{Y}} \tag{B2}$$

where $\tilde{X}_1$ are the residuals of the regression of X_1 on X_2, and $\tilde{\widehat{Y}}$ are the residuals of the regression of $\widehat{Y}$ on X_2. $\widehat{\beta}_1$ can thus be rewritten as follows:

$$\begin{aligned}\widehat{\beta}_1 &= (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'(\mathbb{I} - X_2(X_2'X_2)^{-1}X_2')(Y+\nu) \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'(\mathbb{I} - X_2(X_2'X_2)^{-1}X_2')Y + \\ &\quad (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'(\mathbb{I} - X_2(X_2'X_2)^{-1}X_2')\nu \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{Y} + (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{\nu}\end{aligned} \tag{B3}$$

where $\tilde{\nu}$ are the residuals of the regression of ν on X_2. The expectation of this estimate is

$$\mathbb{E}[\beta_1] = \beta_1 + \frac{cov(\tilde{X}_1, \tilde{\nu})}{var(\tilde{X}_1)}. \tag{B4}$$

Intuitively this makes sense – if the residual variation in X_1 is correlated with the residual prediction error, after controlling for X_2 in both cases, our estimate will be biased. Thus following the same logic as above, we can make the adversary a linear regression of $\tilde{\nu}$ on $\tilde{X}_1$.

A similar argument can be extended to the instrumental variables case with controls. Following the two-stage least squares estimation procedure, we first use covariates X_2 and instruments Z to predict X_1:

$$\widehat{X}_1^{IV} = C(C'C)^{-1}C'X_1 \tag{B5}$$

where $C = [1, X_2, Z]$. Then, we regress the outcome against the predicted values of X_1 and the covariates X_2 and take the estimated coefficient for $\widehat{X}_1^{IV}$ in this second stage regression as our estimate of the true β_1. By the Frisch-Waugh-Lovell theorem, we have

$$\begin{aligned} \widehat{\beta}_1^{2SLS} &= (\tilde{\widehat{X}}_1{}'\tilde{\widehat{X}}_1)^{-1}\tilde{\widehat{X}}_1{}'\tilde{\widehat{Y}} \\ &= (\tilde{\widehat{X}}_1{}'\tilde{\widehat{X}}_1)^{-1}\tilde{\widehat{X}}_1{}'\tilde{Y} + (\tilde{\widehat{X}}_1{}'\tilde{\widehat{X}}_1)^{-1}\tilde{\widehat{X}}_1{}'\tilde{\nu} \end{aligned} \tag{B6}$$

where $\widetilde{\widehat{X}_1}$ are the residuals from regressing $\widehat{X}_1^{IV}$ on X_2, $\tilde{Y}$ are the residuals from regressing Y on X_2, and $\tilde{\nu}$ are the residuals from regressing ν on X_2.

In the case of a single instrument Z, this estimator of β_1 can be rewritten more simply following the indirect least-squares procedure. For this approach, we perform linear regressions for the models

$$\widehat{Y}_i = \gamma_0 + \gamma_1 X_{1i} + X_{2i}\gamma_2 + w_i; \tag{B7}$$

$$\widehat{X}_{1i} = \alpha_0 + \alpha_1 Z_i + X_{2i}\alpha_2 + u_i \tag{B8}$$

This produces the following estimate of β_1, which coincides with $\widehat{\beta}_1^{2SLS}$ for this special case:

(B9)
$$\widehat{\beta}_1^{ILS} = \frac{\widehat{\gamma_1}}{\widehat{\alpha_1}} = \frac{(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\hat{\tilde{Y}}}{(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{X_1}} = \frac{cov(\tilde{Z}, \hat{\tilde{Y}})}{cov(\tilde{Z}, \tilde{X_1})}$$
$$= \frac{cov(\tilde{Z}, \tilde{Y})}{cov(\tilde{Z}, \tilde{X_1})} + \frac{cov(\tilde{Z}, \tilde{\nu})}{cov(\tilde{Z}, \tilde{X_1})}$$

In the above expressions for $\widehat{\beta}_1^{2SLS}$ and $\widehat{\beta}_1^{ILS}$, we see that the coefficient estimates are the sum of the coefficient estimate that we would obtain from performing these procedures given Y without measurement error - which is consistent for β_1 given IV assumptions - and an additional bias term involving the measurement error ν. To minimize this bias, we propose an adversary in the form of a regression of $\tilde{\nu}$ on $\tilde{Z}$ for the single instrument case, or $\tilde{\nu}$ on $\hat{\tilde{X}}_1$ more generally.