Debiasing Estimates of Global Forest Cover Loss

By MATTHEW GORDON, ELIANA STONE, MEGAN AYERS, AND LUKE SANFORD *

Expanded access to satellite data, reduced computational costs, and continued advances in machine learning have catalyzed a measurement revolution in the environmental and social sciences. These innovations allow for the analysis of previously difficult-to-measure variables at high spatial and temporal resolution, allowing for precise causal evaluation of policies. This research is crucial to the design and implementation of interventions that effectively combat climate change, alleviate poverty, and improve food security.

A common measurement strategy is to train a machine learning model to predict an outcome Y, using satellite derived features k, generating predictions $\widehat{Y}$. However, when these models are to minimize a standard loss function, the resulting predictions can bias analyses when used in regressions. 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 can occur even in cases when researchers have a valid instrument or an experimental research design.

New methods have been developed to correct these biases; however, most methods require some representative ground-truth labeled data. In this paper, we use active learning techniques to collect ground truth data on deforestation in Africa and check for biases in widely used satellite-derived data on forest cover change.

I. Prediction Errors and Causal Inference

To fix ideas, consider a regression with one independent variable X measured for observations indexed by i. We seek to estimate the effect of changes in X on an outcome Y according to:

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

In this case, β measures 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 from a machine learning model: $\widehat{Y}_i = Y_i + \nu_i$, where ν_i is the measurement error for a given observation.

When we estimate β using OLS with $\widehat{Y}_i$ as the dependent variable instead of Y_i, our estimate 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 on the right-hand-side is standard endogeneity bias. The third term is the result of using machine learned proxies for Y_i instead of the true values. In words, the measurement error from our machine learning model will bias our estimate if it is correlated with the treatment variable. This can happen even when X_i is assigned randomly. For example, if a treatment has no effect on outcomes of interest, but it induces non-random measurement error, researchers could measure a spurious effect where none existed (see panels b and d in Figure 1).

* Gordon: Paris School of Economics (email: matthew.gordon@psemail.eu); Stone: Yale University (email: eliana.stone@yale.edu; Ayers: Reed College (email: meganayers@reed.edu); Sanford: Yale University (email: luke.sanford@yale.edu). 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. We also thank Refine.ink for help with manuscript preparation.

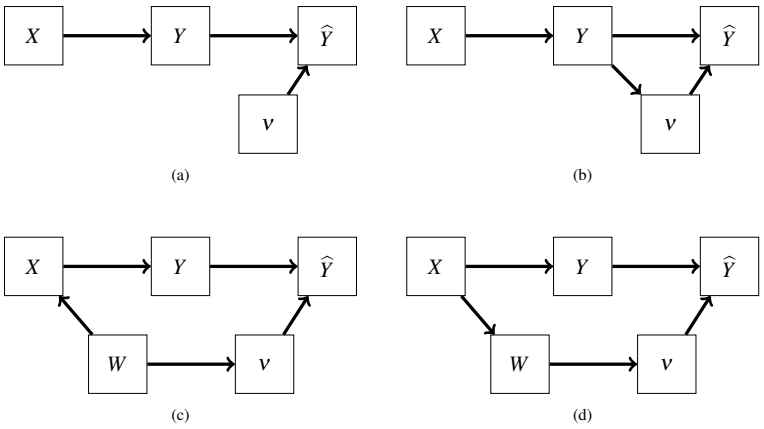

Figure 1. : Four Directed-Acyclic-Graphs

Note: Diagrams illustrate potential relationships between treatment (X), outcomes (Y), measurement error (v), machine learning model predictions ($\widehat{Y}$), and unobserved variables (W). (a) shows classical measurement error, (b) shows outcome-induced bias, (c) shows confounder-induced bias, and (d) shows treatment-induced bias.

A. Related Work

New work on this topic shows how to correct for this type of error using a subsample of points where the ground truth measures of Y are known. One approach, prediction-powered-inference, develops a loss function for the downstream estimator that corrects for biases (Angelopoulos et al., 2023; Carlson and Dell, 2025). A related method known as predict-then-debias, estimates the bias and subtracts it from the coefficient of interest (Kluger, 2025). A third set of papers use a variety of approaches to correct the predictions themselves (Ratledge, 2021; Proctor, Carleton and Sum, 2023; Battaglia et al., 2024; Rambachan, Singh and Viviano, 2025).

In Sanford et al. (2025), the authors investigate prediction models that avoid differential measurement error by construction. They analyze a class of models with parameters ω^* that satisfy:

$$\omega^* = \arg\min_{\omega} L_p(\widehat{Y}(\omega), Y, k) \quad \text{s.t.} \quad \text{Cov}(X, Y - \widehat{Y}(\omega)) = 0 \tag{3}$$

where L_p is a loss function, such as mean squared error. In that paper, the authors show that the adversarial debiasing framework proposed by Zhang, Lemoine and Mitchell (2018), which was originally used to debias machine learning model predictions with respect to race or gender, can be isomorphic to the constrained loss function in (3) under some conditions. Roughly speaking, in an adversarial debiasing setup, a primary model simultaneously tries to minimize its own loss function, while maximizing the loss function of an adversarial model.

If the adversary is a linear regression of the form:

$$v_i = \mu + \gamma X_i + \varepsilon_i, \tag{4}$$

then the primary model will try to make predictions such that the errors are uncorrelated with the treatment variable — exactly the condition needed to ensure unbiased estimates of β in equation 1.

For all of the methods described above, a major obstacle to implementation is the lack of high-quality, representative ground-truth data that is needed to debias predictions.

II. Case Study: Hansen Global Forest Change

The Global Forest Change (GFC) data estimates forest cover loss globally at high resolution using a machine learning model trained on satellite imagery (Hansen, 2013). These estimates are widely used

to study the drivers of deforestation. They have also been found to contain non-classical measurement error in certain settings. Tropek (2014) showed that the data often confuses tree plantations for forests, which may be relevant for research questions that study deforestation. Bastin (2017) showed underestimation of forest cover in a cross-section of ground truth points in global dryland regions.

For most applications in social sciences that attempt to answer causal questions about the drivers of deforestation, we require ground truth data on changes in deforestation rates over time to assess biases. One source of such data is Guo, Zhu and Gong (2022) (hereafter, Guo). The authors select a sample of pixels and visually interpret high-resolution historical imagery from various sources to manually label forest cover change from 2000-2020. In the left panel of Figure 2, we match these pixels to the corresponding GFC predictions[1] and calculate v_i, treating the Guo labels as the ground-truth. We count a prediction of deforestation as accurate if the pixel was deforested in any year, not necessarily the exact year that was predicted. We then aggregate the average prediction error to the country level, attempting to weight for their unequal probability sampling strategy. Positive values (countries in green) indicate that the predictions tend to overestimate deforestation relative to the Guo labels, while negative values (countries in red) indicate that the predictions underestimate deforestation.

While sample sizes for some countries are small, the map shows that the GFC predictions may be differentially biased across countries. There are several issues with the Guo labels that limit their usefulness, however. The authors do not record their probabilities of sampling within each strata, and furthermore, they drop some observations, making it difficult to determine representativeness. The sample is also small and mostly static. For example, there are only 149 observations in Africa where any forest cover loss is recorded, limiting statistical power to detect biases. For these reasons, we undertake our own data collection exercise in Africa to build a more complete and representative labeled data set on the continent with the highest rates of deforestation.

A. *Active Sampling and Approximately Optimal Verification of Forest Change in Africa*

To collect this data, we use approximately optimal active sampling methods (Zrnic and Candes, 2024; Gordon and Papp, Forthcoming). The basic idea is to first train a machine learning model on whatever ground-truth data is available. In our case, we use the Guo data to train a 3-layer LSTM neural network (Hochreiter and Schmidhuber, 1997).

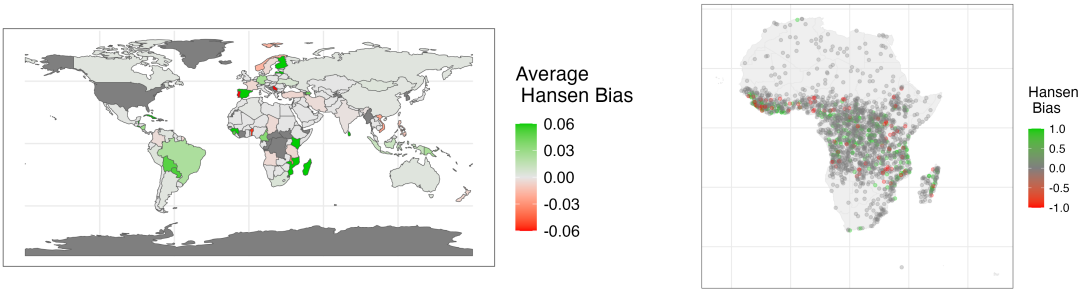

Figure 2. : Measurement Error in Global Tree Cover Loss Estimates

Note: Left panel shows average bias in GFC predictions calculated using Guo, Zhu and Gong (2022) ground truth data. Values beyond ± 0.06 saturated. Right panel shows bias for a sample of pixels in Africa selected using active learning methods (see text). Positive values indicate predicted deforestation where none occurred in any year (false positives), and negative values indicate that deforestation occurred where it was not predicted in any year (false negatives).

[1] GFC 2023 Version 1.11 is used throughout the analysis.

Next we use this model, which, given the limited training data, has fairly low out-of-sample-accuracy and may be biased, to make a preliminary set of predictions in a representative sample of unlabeled points. We then use these predictions in combination with Neyman (1934) optimal sampling methods to choose a representative set of these points to label. In particular we choose points with probability $\pi_i \propto \sqrt{p_i(1-p_i)}$, where p_i is our 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.

We select 2,000 pixels using this method. We then verify whether deforestation has occurred in these pixels using visual interpretation of historical high-resolution satellite imagery in Google Earth Pro. For each point, we have two annotators independently inspect all available imagery, and mark the first year between 2000-2017 in which deforestation is observed. If the annotators disagree, we have one of the authors break the tie. To mimic the structure of the GFC data, we define a forested pixel as having 30% tree cover (vegetation $>$ 5m tall), and deforestation as the loss of more than half of pre-existing tree-cover. As with the GFC data, we do not record multiple instances of deforestation in the same location or regrowth. This process is imperfect, especially in the earlier years of our study period when high-resolution data is less frequently available, yet we believe it offers a substantial improvement over current alternatives, and it should prove valuable to researchers in stimulating further investigation when bias is detected.

The right panel of Figure 2 shows a map of our labeled points. Note that the active sampling strategy leads us to over sample 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 GFC prediction errors. Green points are false positives, points predicted to have deforestation that were not deforested. These appear to be more concentrated in the most densely forested regions of central Africa. Red points show false negatives — points that saw deforestation but were missed by the GFC predictions. These appear to be more common in areas with lower baseline forest cover.

III. Results and Conclusion

To test for systematic biases in the GFC predictions, we estimate equation 4 using this labeled data, where v_i is the GFC prediction error, and X_i is a geographic characteristic (elevation, rainfall, slope, travel time to the nearest city, and deforestation in neighboring pixels). The results in Appendix Table A1 show systematic biases with respect to several variables. Rainfall is associated with false negatives (column 2), possibly because vegetation regrowth is relatively rapid. Distance to the nearest city has a positive and significant coefficient (column 4), indicating that deforestation in urban environments is less likely to be captured. Finally, we see large negative coefficients on deforestation in neighboring pixels, indicating that deforestation is underestimated in high-deforestation areas.

These results demonstrate that empirical work attempting to infer the causal effects of policies using machine-learned variables must carefully consider how differential prediction error will affect the analysis. Recent research in this area offers a variety of methods to detect and correct for potential biases; however, most of these methods rely on a representative sample of ground-truth data. In this paper, we collect ground-truth data on deforestation in Africa and assess the bias in commonly used satellite-derived measures of deforestation. There is a great need for future research in this area to determine which methods work best under which circumstances, optimize the collection of ground-truth data, and gather ground-truth data for a greater variety of outcomes.

REFERENCES

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

Bastin, Jean-Franois et al. 2017. "The extent of forest in dryland biomes." *Science*, 356(6338): 635–638.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher. 2024. "Inference for Regression with Variables Generated from Unstructured Data."

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

Farr, T. G. et al. 2007. "The Shuttle Radar Topography Mission." *Reviews of Geophysics*, 45(2): RG2004.

Funk, Chris et al. 2015. "The climate hazards infrared precipitation with stationsa new environmental record for monitoring extremes." *Scientific Data*, 2: 150066.

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. et al. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science*, 342(6160): 850–853.

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

Kluger, Dan M. et al. 2025. "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling."

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.

Ratledge, Nathan et al. 2021. "Using Satellite Imagery and Machine Learning to Estimate the Livelihood Impact of Electricity Access."

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

Tropek, Robert et al. 2014. "Comment on High-resolution global maps of 21st-century forest cover change." *Science*, 344(6187): 981–981.

Weiss, D. J. et al. 2018. "A global map of travel time to cities to assess inequalities in accessibility in 2015." *Nature*.

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

Zrnic, Tijana, and Emmanuel J. Candes. 2024. "Active Statistical Inference." arXiv:2403.03208.

APPENDIX

	v_i = *GFC Predicted Deforestation - Observed Deforestation*					
	(1)	(2)	(3)	(4)	(5)	(6)
(Intercept)	−0.022	0.005	−0.009+	−0.064***	−0.007+	−0.005
	(0.029)	(0.006)	(0.005)	(0.016)	(0.004)	(0.005)
Log Elevation (m)	0.001					
	(0.005)					
Rain (mm/day)		−0.009***				
		(0.002)				
Log Slope (degrees + 1)			−0.009			
			(0.008)			
Log travel time to city (minutes)				0.009**		
				(0.003)		
Neighbor mean (k=10)					−0.137***	
					(0.041)	
Neighbor mean (k=50)						−0.231**
						(0.073)
Num.Obs.	1997	1998	1998	1998	2000	2000
R2	0.000	0.009	0.001	0.005	0.006	0.005
F	0.095	17.723	1.218	10.504	11.117	10.026

Table A1—: Predicted Deforestation

Results from estimating equation 4. Dependent variable is GFC predicted deforestation (Hansen, 2013) minus deforestation observed in high resolution imagery. Elevation and slope are calculated using a digital elevation map in Farr (2007), rainfall is from the CHIRPS dataset (Funk, 2015) and travel time to city is from Weiss (2018). Neighbor mean deforestation calculated using k nearest points in the labeled data, weighted by inverse sample inclusion probabilities. All coefficients were estimated with OLS using inverse sample inclusion probability weights. Observations with missing data dropped.