Remote Control: Debiasing Remote Sensing Predictions for Causal Inference

Matthew Gordon
School of the Environment
Yale University

Megan Ayers*
Department of Statistics & Data Science
Yale University

Eliana Stone
School of the Environment
Yale University

Luke Sanford
School of the Environment
Yale University

Abstract

Understanding and properly estimating the impacts of environmental interventions is of critical importance as we work towards achieving global climate goals. Remote sensing has become an essential tool for evaluating when and where climate policies have positive impacts on factors like greenhouse gas emissions and carbon sequestration. However, when machine learning models trained to predict outcomes using remotely sensed data simply minimize a standard loss function, the predictions that they generate can produce biased estimates in downstream causal inference. If prediction 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 policies can be biased. In this paper, we demonstrate how this bias can arise, and we propose the use of an adversarial debiasing model (Zhang, Lemoine, and Mitchell 2018) in order to correct the issue when using satellite data to generate machine learning predictions for use in causal inference. We apply this method to a case study of the relationship between roads and tree cover in West Africa, where our results indicate that adversarial debiasing can recover a much more accurate estimate of the parameter of interest compared to when the standard approach is used.

1 Introduction

Advances in machine learning and the increasing availability of satellite imagery have catalyzed the use of remotely sensed measures of human activity or environmental outcomes in research seeking to infer environmental policy impacts. The effort required to directly observe and collect outcome data on-the-ground may be prohibitively expensive or time-consuming. Remote sensing, particularly via satellite imaging, offers a huge opportunity for measuring outcomes at a much lower cost and time investment to researchers. This is especially applicable to those developing and testing interventions to mitigate the causes and effects of climate change, which frequently need to be evaluated at large spatial scales across the globe and in remote locations that are not easily accessible.

To translate satellite imagery to manageable representations of outcomes, researchers typically take a small number of images that have been hand-labeled or cross-referenced with on-the-ground measurements, and use them to train various machine learning models to predict the desired variable (for example, proportion of tree cover). Once these models have been trained, they can be deployed on a much larger set of images to obtain outcome predictions at all points of interest, requiring only a fraction of the data used for impact evaluation to be manually labelled or collected.

*Corresponding author (m.ayers@yale.edu)

A common application of these machine learned predictions for impact evaluation is to use causal inference techniques to estimate the effect of an intervention, or "treatment," on the outcome of interest. One of the assumptions made within the standard causal inference framework is that there are no unobserved variables that "confound" the relationship between the treatment and outcome by depending on or influencing both the treatment and outcome variables. When using predictions from remote sensing methods as outcome values for standard causal inference techniques, the relevant assumption then applies to the outcome *predictions*, not the ground-truth outcomes that are generally unobservable (see examples of these relationships in Figure 1).

Even in a randomized controlled trial where researchers have full control over a randomized treatment assignment mechanism, it is possible for this assumption to be violated if the assignment of treatment influences the outcome prediction error. For example, imagine a randomized experiment where treatment (perhaps a carbon credit) inadvertently increases nearby manufacturing and thereby air pollution. If increased levels of air pollution result in systematic under-predictions of the outcome of interest, then there is a problematic dependency between the prediction error and the treatment. In non-experimentally randomized settings, opportunities for dependencies between the treatment mechanism and outcome prediction error proliferate, related to common concerns about confounding variables and ignorability in observational studies. Suppose that we want to understand the impact of building roads on deforestation. Roads tend not to be built on steep slopes, where remote sensing methods also tend to over-predict tree cover. The effect of slope on tree cover prediction error may be misattributed to being far from roads, in this case.

To remove potential dependencies of this nature from remotely sensed outcome predictions, we propose the use of an adversarial debiasing model, inspired by prior work in algorithmic fairness (Zhang, Lemoine, and Mitchell 2018, Celis and Keswani 2019). Adversarial debiasing has been used in this field to ensure that algorithms do not make systematically inaccurate predictions on the basis of race or gender. Our approach adapts this tactic to generate satellite-derived measures that are unbiased for true outcome values across the range of the treatment variable, improving the accuracy of downstream estimation procedures for impact evaluation. With the use of a hand-labeled data set of forest cover in West Africa (Bastin et al. 2017), we demonstrate the usefulness of this method as a means of obtaining accurate estimates of policy-relevant parameters in cases where standard methods fail.

2 Related work

Recently, other studies have pointed out cases of systematic prediction errors in a variety of remotely sensed variables. Balboni, Barbier, and Burgess 2022 review recent work focused on these issues in deforestation prediction models. One such model which has received almost 10,000 citations, Hansen et al. 2013, has received criticism for systematically under-predicting tree cover in tropical forests as well as dryland biomes (Balboni, Barbier, and Burgess 2022, Bastin et al. 2017). Fowlie, Rubin, and Walker 2019 show that two satellite-based air quality prediction methods are down-biased for high ground-truth air pollution values. This behavior could attenuate downstream estimates of true causal effects, since control regions that continue polluting at higher levels would be underestimated, while treatment regions which were caused to pollute at lower levels would be more accurately detected. In econometrics, where remotely sensed variables are also becoming more commonly used, work has shown the prediction error of night lights data to be greater in rural areas (Gibson et al. 2021).

There has also been recent work focused on techniques for handling this kind of prediction error to avoid downstream biases during estimation for impact evaluation (Alix-Garcia and Millimet 2020, Garcia and Heilmayr 2022, Ratledge et al. 2022, Proctor, Carleton, and Sum 2023). Our work is distinct from these as it focuses on addressing the issue in the prediction modeling stage, rather than the estimation stage, and by the approach of directly incentivizing the prediction model to avoid dependency between the prediction error and the treatment variable.

Ozery-Flato et al. 2020 also present an algorithm that applies adversarial models to a causal inference framework. Their approach involves re-weighting observations to improve covariate balance between treatment groups, without considering outcome data. Instead, they focus on addressing omitted variable biases that may result from comparing systematically different populations in observational studies. In comparison, our approach focuses on estimation bias that is propagated from certain outcome prediction error structures.

3 Methodology

Our methodology is motivated by a simple example of how impact evaluation may be biased in the presence of systematic prediction error in remotely sensed variables. Consider a scenario where we seek to estimate the effect of some treatment X (ex. carbon crediting or road construction) on an observed outcome Y (ex. tree cover). Imagine that the true model relating these is simply

$$Y = \gamma + \tau X + \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, γ represents the expected outcome without treatment, and $\tau \in \mathbb{R}$ represents the homogeneous treatment effect for all units. Suppose that it is too costly or time consuming for us to directly observe all of the outcome data $Y_1, \ldots, Y_n$ for n units in a study. Instead, we use remote sensing methods to obtain machine learned predictions of them, which we assume are the sum of the true outcome values Y and a potentially unit-dependent prediction error ν, whose marginal and any relevant joint distributions are unobserved to us:

$$\hat{Y}_i = Y_i + \nu_i = \gamma + \tau X_i + \epsilon_i + \nu_i. \tag{2}$$

If there is no prediction error, $\nu = 0$, then we have obtained perfect predictions of the observed outcome of unit i using these methods.

Suppose that given these values of $\hat{Y}_i$ for units $i = 1, \ldots, n$, we go to estimate the treatment effect τ using linear regression, as is common in this setting. This will produce an estimate $\hat{\beta}$ for τ that in expectation is equal to

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \frac{Cov(X, \hat{Y})}{Var(X)} \\
&= \frac{Cov(X, \gamma + \tau X + \epsilon + \nu)}{Var(X)} \\
&= \tau + \frac{Cov(X, \epsilon) + Cov(X, \nu)}{Var(X)}.
\end{aligned} \tag{3}$$

We see that for the regression coefficient $\hat{\beta}$ to be unbiased for the treatment effect, it is necessary that the second term be zero. It is typical to assume that $Cov(X, \epsilon) = 0$; in a randomized controlled trial because the treatment assignment X is randomized independently of ϵ, and in an observational study where the ignorability assumption is necessary. However, even if this is true, the prediction error ν may continue to bias the estimate of τ to the extent that ν is correlated with the treatment variable X.

Our approach for generating accurate predictions $\hat{Y}_i$ with minimal correlation between ν and X consists of combining a primary prediction model, $\mathcal{M}_p$, and an adversarial model $\mathcal{M}_a$. In the work described here, we have implemented $\mathcal{M}_p$ as a logistic regression classification model, and simple linear regression for $\mathcal{M}_a$, which makes predictions $\widehat{X}$ of the treatment variable X using the residuals from the primary model. To train the model we use a loss function combining the log-loss of $\mathcal{M}_p$ and the scaled MSE of $\mathcal{M}_a$:

$$\begin{aligned}\mathcal{L}_{total} &= \textit{log-loss}_p - \alpha \cdot MSE_a \\ &= -\sum_{i=0}^{n} \Big(Y_i \log(\sigma(\hat{Y}_i)) + (1 - Y_i) \log(1 - \sigma(\hat{Y}_i)) \Big) - \alpha \sum_{i=0}^{n} (\widehat{X}_i - X_i)^2. \end{aligned} \tag{4}$$

where α is a tuneable hyperparameter. In our evaluation of the model, we cross-fit predictions in three folds to optimize the use of our fully labelled data set, though the results are very similar when a standard train-test split is used. We evaluate the model for a range of values of α. At higher values of α, we expect to approach predictions whose errors have no correlation with the treatment variable X, but potentially at the cost of overall prediction accuracy. When $\alpha = 0$, the model reduces to a standard logistic regression.

4 Experiments

We apply the adversarial debiasing method to a case study on tree cover in West Africa. From the data set produced in Bastin et al. 2017, we use 16,908 hand-labeled points that are coded as forest or not forest as the "true" measure of forest cover in this region. We utilize time series of corresponding satellite records from the Landsat 7 ETM sensor, and calculate median and first and third quartile values over time for each point, following the approach described in Hansen et al. 2013. Our key independent variable for this case study is the distance from each point to the nearest road, which was obtained from the GLOBIO global roads database. Because this data is top-coded, we train and test our models on only the observations with distance to road below 32km. We also incorporate control variables in some experiments including slope, elevation, aspect, and aridity zone to compare model performances.

Initially, we compare the performances of baseline approaches, where $\alpha = 0$ and our prediction model is a standard logistic regression, when 1. only satellite data is used to predict forest cover and 2. when satellite data and the additional covariates are used. We compare the estimates produced by regressing these forest cover predictions on the variable of interest, distance to the nearest road. Because different classification thresholds led to very different results in our experiments, we use the predicted probabilities of forest cover, rather than binary forest cover class in our exploratory work for this case study. In our ongoing work we plan to address how choices of the classification threshold factor into our modeling procedure. In addition to these baseline models, we repeat the procedure for a sequence of nonzero α values to assess the adversarial debiasing model performance.

Note that while we developed this model with causal questions in mind, we are not claiming identification of a causal relationship between distance to roads and forest cover in this case study. Rather, we wish to see regardless of if someone has found a way to get $Cov(X, \epsilon) = 0$ whether the analysis would still be biased by the presence of prediction error. Here, setting aside the issue of likely omitted variable bias, we find that prediction error presents a *separate* source of bias.

5 Results

Our baseline experiments demonstrate that compared to the parameter estimate obtained using the Bastin et al. 2017 ground-truth data, both estimation attempts using baseline model predictions result in inaccurate values (Table 1). When looking at the map of prediction errors (Figure 2) from the standard logistic regression model without covariates, we observe that the errors have a spatial structure beyond the spatial presence of forest cover (including covariates does not meaningfully change this behavior). In this instance, we chose the classification threshold of predicted probabilities that optimized the overall mean squared error to visualize classification error.

However, we find that when we use predictions from the adversarial debiasing model, we can recover an estimate much closer to the true parameter at only a small cost to the overall

	ground truth	baseline	baseline + covars	adversarial model ($\alpha = 60$)
(Intercept)	0.278	0.563	0.572	0.511
	(0.006)	(0.004)	(0.004)	(0.003)
log_distance	−0.040	−0.079	−0.087	−0.042
	(0.003)	(0.002)	(0.002)	(0.002)
Num.Obs.	16 908	16 908	16 908	16 908
R2 Adj.	0.007	0.059	0.075	0.030
RMSE	0.41	0.27	0.27	0.21

Table 1: Linear regression summaries of ground truth forest cover, baseline model forest cover probability predictions, baseline model forest cover probability predictions when additional covariates are included in the model, and adversarial model forest cover probability predictions regressed against the log of distance to nearest road (in kilometers). Heteroskedasticity-robust standard errors (HC2) are reported. Adjusted R-squared values are low - in practice we may be concerned with capturing more variability in the outcome by including additional covariates, but here we focus specifically on removing coefficient estimation bias stemming from outcome prediction error.

accuracy of the prediction model. This depends on the choice of the α, but we find that the estimate quickly converges to the ground-truth value as we increase this parameter (Figure 3). Work is ongoing to experiment further and provide guidelines for tuning α in realistic settings where the true parameter of interest is unknown.

6 Conclusions

The use of remotely sensed outcomes in fields such as environmental policy impact research is rapidly expanding, particularly in the areas of deforestation, electrification (night lights), and air pollution. Outcomes related to many of these measures are key to achieving global climate goals, and are the focus of many climate-focused impact evaluations. However, researchers who use off-the-shelf satellite-derived measures of these key dependent variables are likely introducing bias to the results of their inferences. We formalize the problem, identify potential sources of bias, and provide researchers with an easy to implement method to reduce bias originating from dependencies between prediction error of remotely sensed variables and the key independent variable or treatment.

There are limitations to the current work, and efforts are ongoing to continue experimentally assessing the behavior of this method given different choices of data sets, primary and adversarial model combinations, and loss functions. A challenge in this setting is finding data that allows a "ground-truth" causal effect to be reasonably recovered. Another ongoing line of work is determining appropriate standard errors for inference in this setting. The typical OLS and heteroskedasticity-robust standard errors and do not capture uncertainty in the outcome prediction model - we may wonder how different our estimates would be if we resampled the training data set and repeated the modeling process. Bootstrap methods can capture this uncertainty, but are computationally intensive. Additionally, in a causal inference setting, dependencies introduced by this model between the treatment status of some samples and the outcome predictions of others may warrant additional statistical developments to ensure that variance is estimated appropriately for causal effect estimators.

References

Alix-Garcia, J. and D. L. Millimet (Aug. 2020). *Remotely Incorrect?* URL: https://drive.google.com/file/d/1YjhnqL68vIsZd0TXYVhtBzeCAbEmyaYF/view?usp=sharing (visited on 01/04/2021).

Balboni, C., E. B. Barbier, and J. C. Burgess (Sept. 2022). "The Economics of Tropical Deforestation". en. In: *Journal of Economic Surveys* 15.3, pp. 413–433. ISSN: 09500804, 14676419. DOI: 10.1111/1467-6419.00144. URL: https://onlinelibrary.wiley.com/doi/10.1111/1467-6419.00144 (visited on 09/12/2022).

Bastin, J.-F. et al. (May 2017). "The extent of forest in dryland biomes". In: *Science* 356.6338. Publisher: American Association for the Advancement of Science, pp. 635–638. DOI: 10.1126/science.aam6527. URL: https://www.science.org/doi/10.1126/science.aam6527 (visited on 09/06/2022).

Celis, L. E. and V. Keswani (Jan. 2019). *Improved Adversarial Learning for Fair Classification*. arXiv:1901.10443 [cs, stat]. URL: http://arxiv.org/abs/1901.10443 (visited on 09/12/2022).

Fowlie, M., E. Rubin, and R. Walker (May 2019). "Bringing Satellite-Based Air Quality Estimates Down to Earth". en. In: *AEA Papers and Proceedings* 109, pp. 283–288. ISSN: 2574-0768. DOI: 10.1257/pandp.20191064. URL: https://www.aeaweb.org/articles?id=10.1257/pandp.20191064 (visited on 09/12/2022).

Garcia, A. and R. Heilmayr (Aug. 2022). *Conservation Impact Evaluation Using Remotely Sensed Data*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4179782. URL: https://papers.ssrn.com/abstract=4179782 (visited on 09/29/2022).

Gibson, J. et al. (Mar. 2021). "Which night lights data should we use in economics, and where?" en. In: *Journal of Development Economics* 149, p. 102602. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2020.102602. URL: https://www.sciencedirect.com/science/article/pii/S0304387820301772 (visited on 08/03/2022).

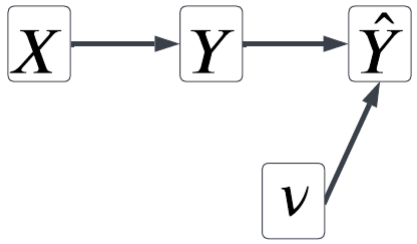
Hansen, M. C. et al. (Nov. 2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change". In: *Science* 342.6160. Publisher: American Association for the Advancement of Science, pp. 850–853. DOI: 10.1126/science.1244693. URL: https://www.science.org/doi/10.1126/science.1244693 (visited on 09/13/2022).

Ozery-Flato, M. et al. (Sept. 2020). *Adversarial Balancing for Causal Inference*. arXiv:1810.07406 [cs, stat]. URL: http://arxiv.org/abs/1810.07406 (visited on 03/03/2023).

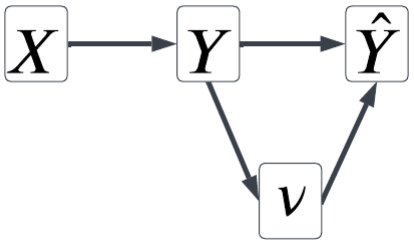
Proctor, J., T. Carleton, and S. Sum (Jan. 2023). *Parameter Recovery Using Remotely Sensed Variables*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4334940. URL: https://papers.ssrn.com/abstract=4334940 (visited on 01/26/2023).

Ratledge, N. et al. (Nov. 2022). "Using machine learning to assess the livelihood impact of electricity access". en. In: *Nature* 611.7936. Number: 7936 Publisher: Nature Publishing Group, pp. 491–495. ISSN: 1476-4687. DOI: 10.1038/s41586-022-05322-8. URL: https://www.nature.com/articles/s41586-022-05322-8 (visited on 11/17/2022).

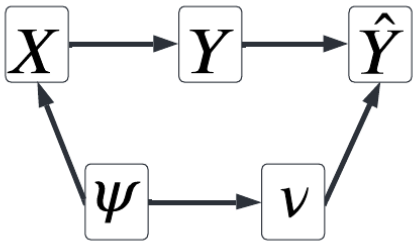
Zhang, B. H., B. Lemoine, and M. Mitchell (Jan. 2018). *Mitigating Unwanted Biases with Adversarial Learning*. arXiv:1801.07593 [cs]. URL: http://arxiv.org/abs/1801.07593 (visited on 07/20/2022).



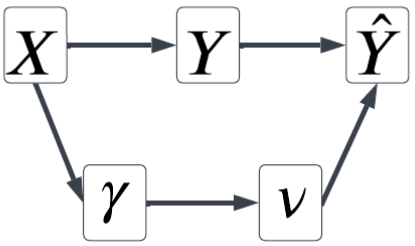
(a) Classical measurement error



(b) Attenuation

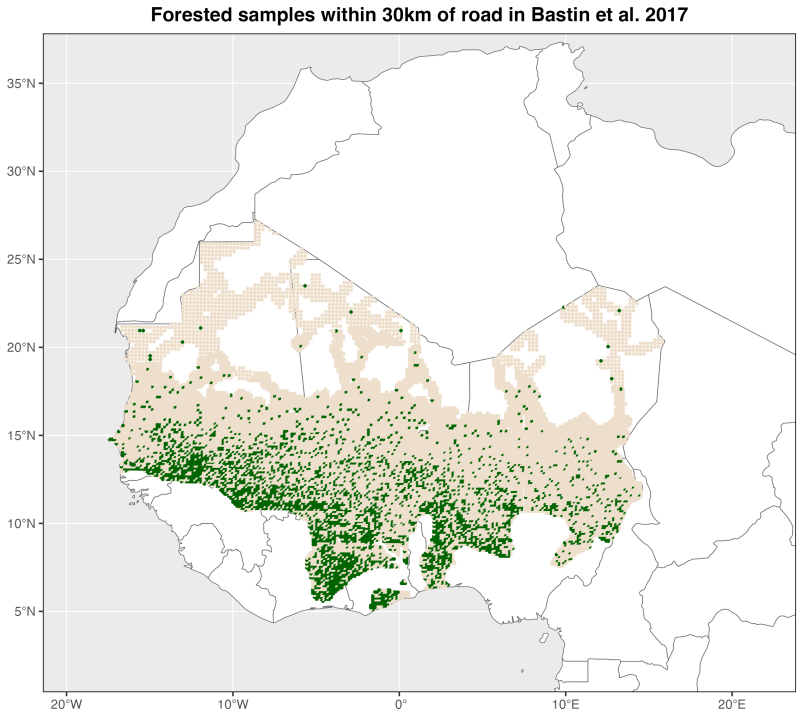


(c) Simple confounder

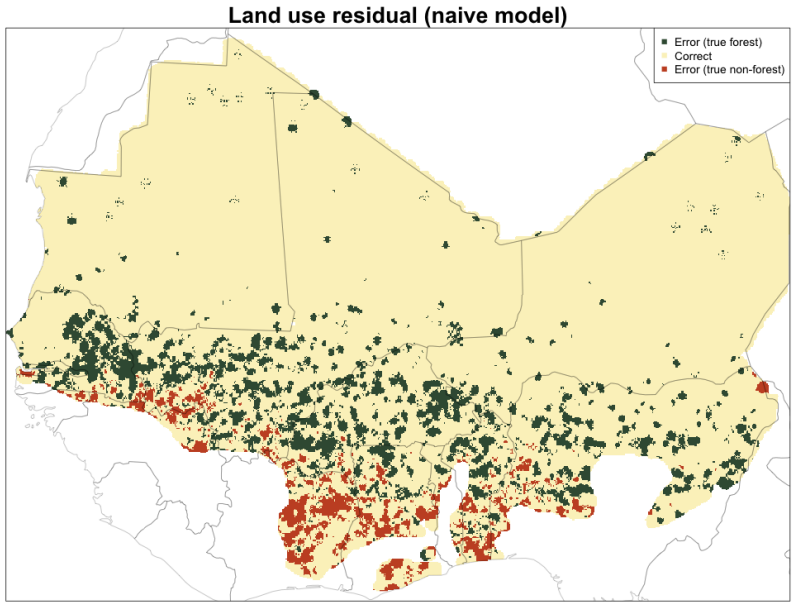


(d) Mis-identification

Figure 1: Causal diagrams where X represents an intervention or treatment, Y represents a ground-truth (but unobservable) outcome, $\hat{Y}$ represents the remotely sensed prediction of Y, ν represents measurement error of $\hat{Y}$, and γ and Ψ represent various confounding variables. (a) Classical measurement error: measurement error term ν of $\hat{Y}$ is independent of observable variables. This is the most commonly assumed structure, if measurement error is considered at all. (b) Attenuation: measurement error ν of $\hat{Y}$ depends on the value of the true outcome variable Y. (c) Simple confounder: measurement error ν depends on a variable Ψ, which also influences the treatment variable X. (d) Mis-identification: The treatment variable X influences a confounding variable γ, which contributes to measurement error ν. The contribution of γ to ν may be mistaken as a contribution of X to the outcome Y.

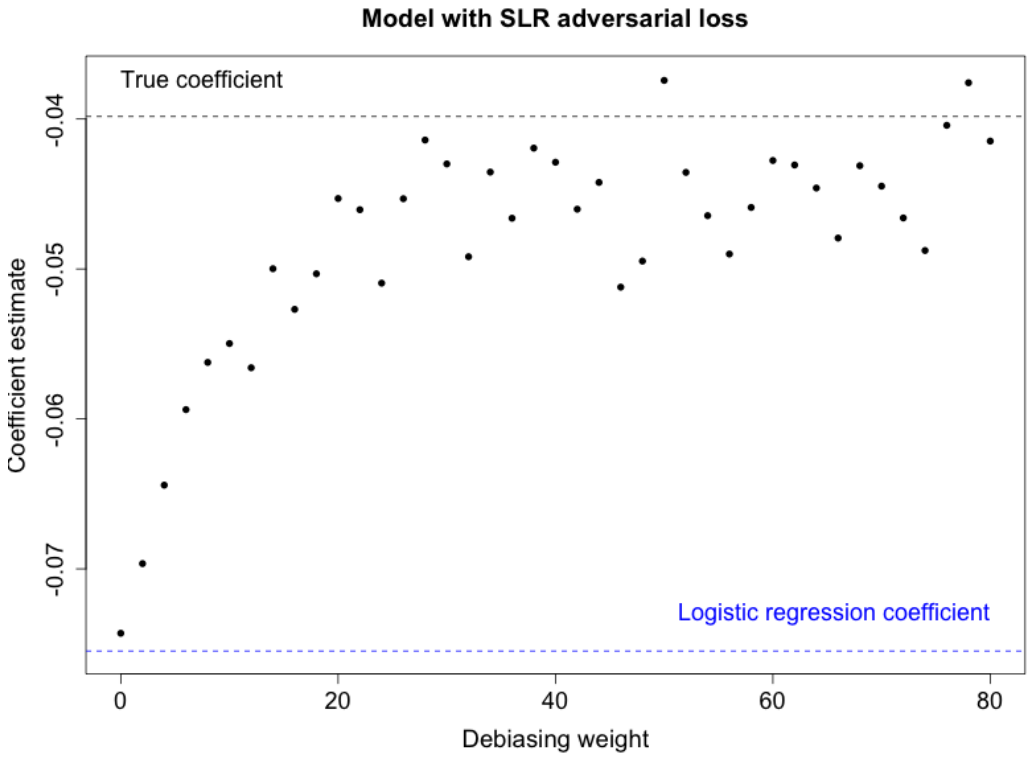


(a)

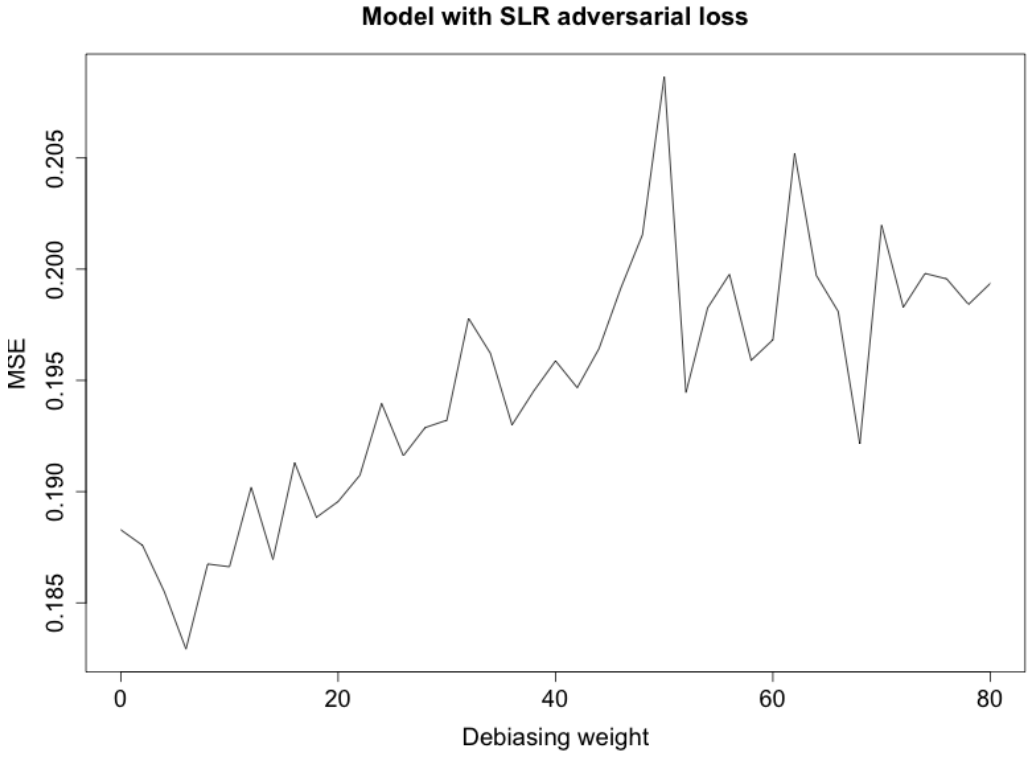


(b)

Figure 2: (a) Distribution of points labeled as forested in Bastin et al. 2017. Green points are forested, beige are not. Note that the sampling intensity of this data set varies by aridity zone (generally sparser in the north, denser in the south). (b) Residuals from a baseline logistic regression model trained on satellite imagery data to predict forest cover. These prediction errors have a clear spatial structure.



(a)



(b)

Figure 3: Results of tuning the hyperparameter α in the adversarial debiasing model for the Bastin et al. 2017 case study.