Satellite Data, Environmental Change, and Causal Inference

SUMMER SCHOOL

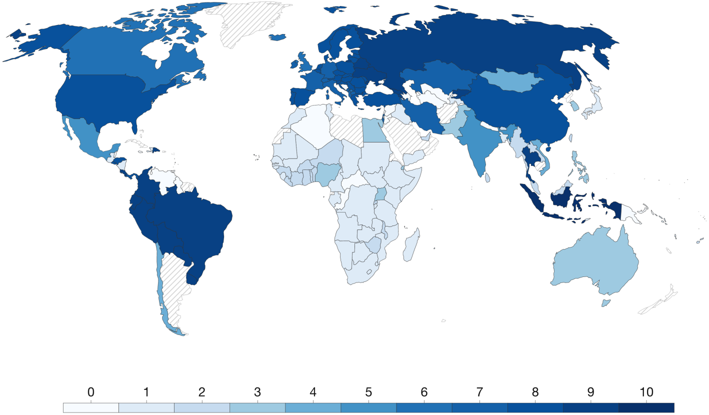
CLIMATE CHANGE

Matthew Gordon - Paris School of Economics

PSE Summer School: Climate Change 2025

How we evaluate policy

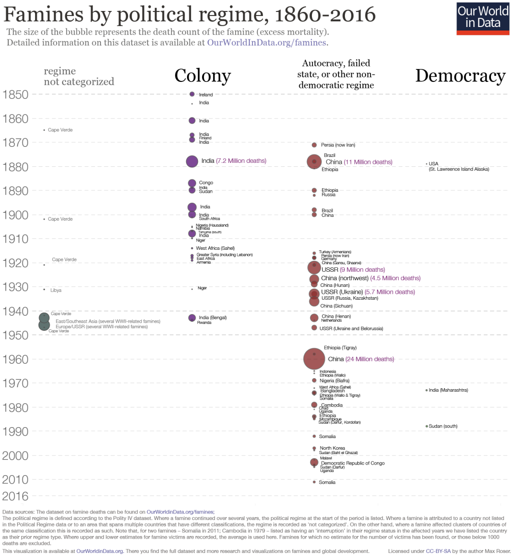
Number of income/consumption surveys in the past decade available via the World Bank, 2021



Source: World Bank Poverty and Inequality Platform
Note: Each decade comprises the current year and the nine years before.

This is an incredible accomplishment!

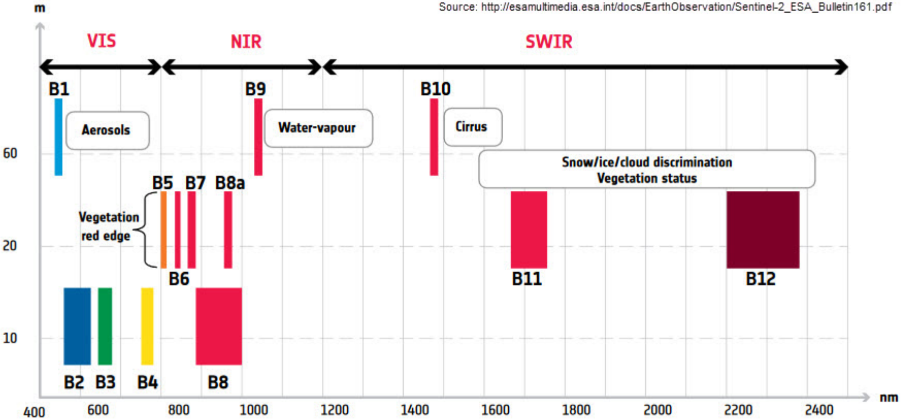
- 1950: PC Mahalanobis launches first large scale representative surveys in India
- late 1950s: China abandons comprehensive data collection



Is it possible to do better?

	Household Surveys	Satellite Data
Spatial Resolution	Selected villages or coarser geographies	Entire world at 10m or less
Temporal Resolution	Typically annual or greater	Every 2 weeks or less
Political vulnerability	Can be distorted or suppressed	Available publicly within 24 hours
Variables Measured	Direct measures of things we care about (income, consumption, etc...)	Light intensity at various wavelengths
Biases and uncertainty	Well understood, often quantifiable	???

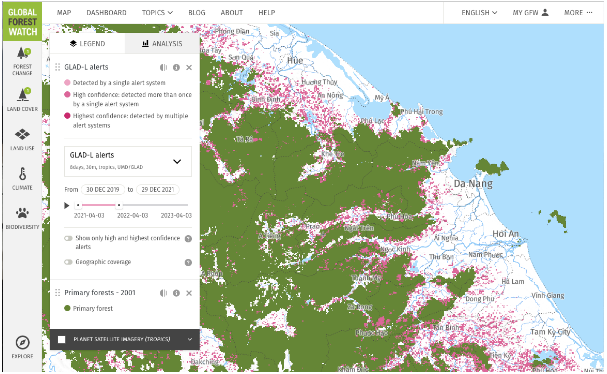
What is Satellite Data?



↑ Spatial resolution versus wavelength: Sentinel-2's span of 13 spectral bands, from the visible and the near-infrared to the shortwave infrared at different spatial resolutions ranging from 10 to 60 m on the ground, takes land monitoring to an unprecedented level

Remote Sensing and Policy Evaluation

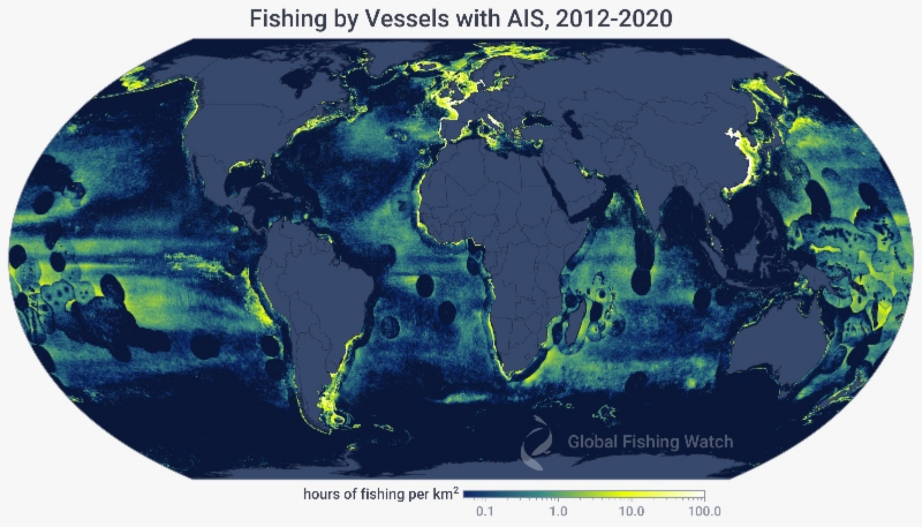
Typical approach is to train a machine learning model (or use off the shelf dataset)



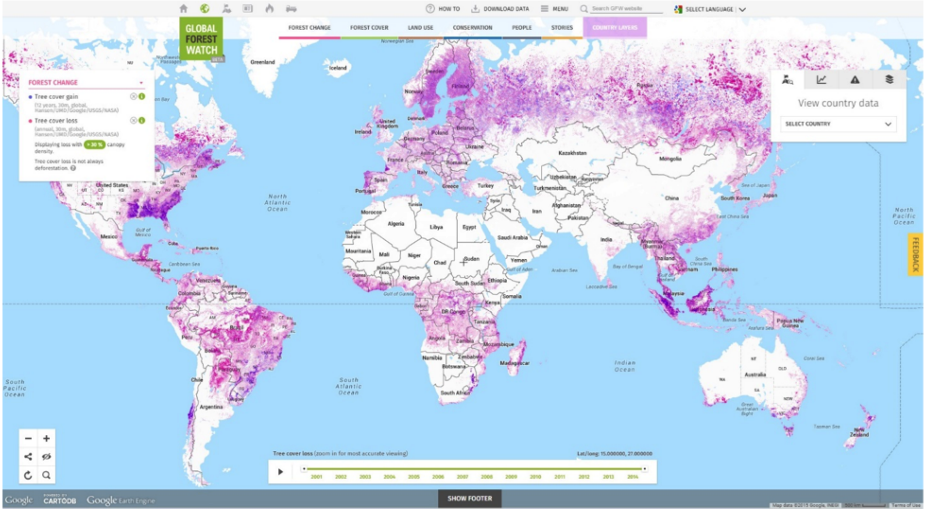
https://www.globalforestwatch.org/map/global/

- Hand labeling all observations is expensive and time-consuming
- Machine learning methods + satellite data $\rightarrow$ data sets of outcome variables with minimal labeling
- Global Forest Watch (Hansen et al., 2013). Global Fishing Watch. Air Pollution (van Donkelaar et al., 2019). Wealth Indices (Yeh et al., 2020).

Applications



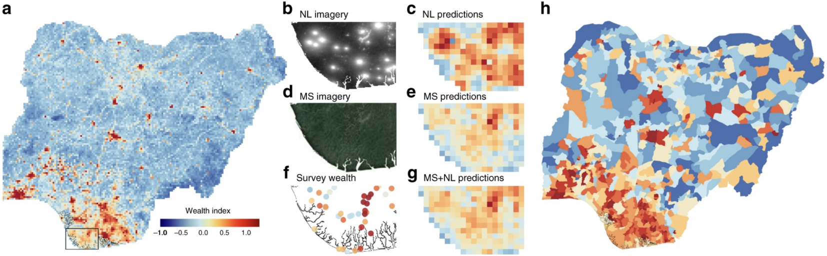
Applications



Applications

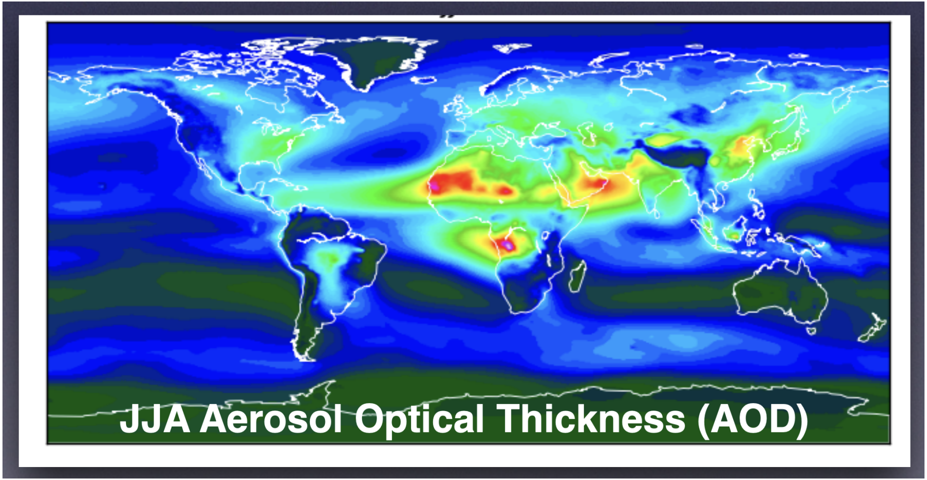
Fig. 6: Spatial extent of imagery allows wealth predictions at scale.

From: Using publicly available satellite imagery and deep learning to understand economic well-being in Africa

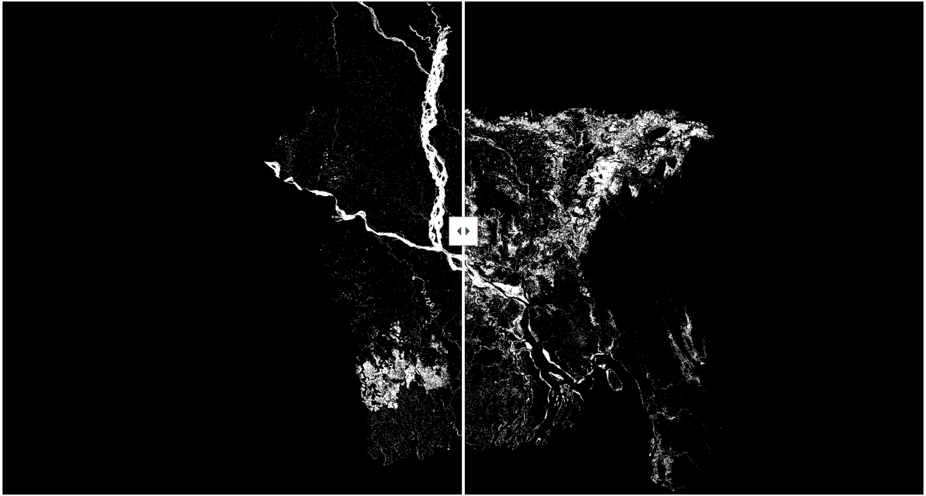


a Satellite-based wealth estimates across Nigeria at pixel level. **b**, **d** Imagery inputs to model over region in Southern Nigeria depicted in box in **a**. **f** Ground truth input to model over the same region. **c**, **e**, **g** Model predictions with just nightlights (NL) as input, just multispectral (MS) imagery as input, and the concatenated NL and MS features as input. In this region, the model appears to rely more heavily on MS than NL inputs, ignoring light blooms from gas flares visible in **b**. **h** Deciles of satellite-based wealth index across Nigeria, population weighted using Global Human Settlement Layer population raster, and aggregated to Local Government Area level from the Database of Global Administrative Areas.

Applications



Applications

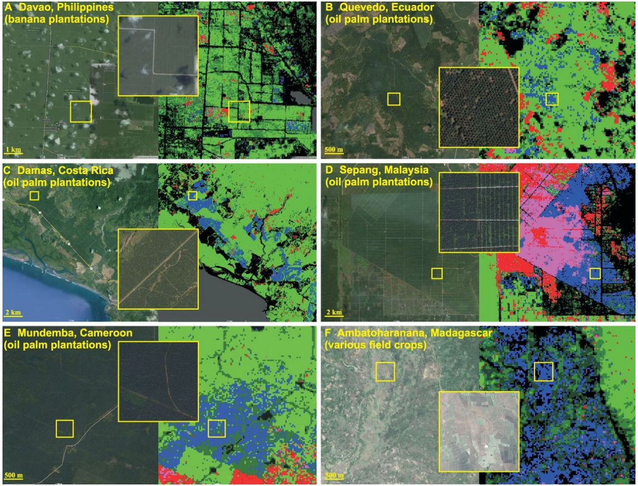


Scrolling between the left image (March) and the right image (September) shows how surface water changes over the calendar year.

Beware Prediction Errors

Well documented biases can pose problems for causal inference

- Air Pollution (Fowlie, Rubin and Walker, 2019)
- Forest Cover (Tropek et al., 2014)
- Wealth (Ratledge et al., 2021)



Remotely sensed variables have measurement error

- Remotely sensed data products will likely have minimum global prediction error, but may make systematic errors within subsets of the training data feature space.
- These might include:
  - Any geographic area
  - Treatment status
  - The range of the ground-truth measure
- Especially likely for 'underrepresented' areas of feature space
  - 'A Fairness Accuracy Frontier' (Liang, Lu and Mu, 2023)

Outline

Relevant Literature:

- Correcting for measurement error in remote sensing (Alix-García and Millimet, 2023; Torchiana et al., 2023; Proctor, Carleton and Sum, 2023; Angelopoulos et al., 2023; Kluger et al., 2025).
- Machine learning, adversarial debiasing, 'Algorithmic Justice', active learning (Zhang, Lemoine and Mitchell, 2018; Chernozhukov et al., 2020; Liang, Lu and Mu, 2023; Zrnic and Candès, 2024)

Agenda for Today:

- Preliminiaries: Problem setup, existing solutions
- Paper 1: Remote Control: Debiasing Remote Sensing Predictions for Causal Inference (with Luke Sanford, Megan Ayers, and Eliana Stone)
  - ML algorithm to 'debias' measurement error using labelled data
- Paper 2: Dumps (with Anna Papp)
  - Optimal selection of labelled points

A Simple Setup:

We want to estimate:

$$Y_i = \alpha + \tau X_i + e_i$$

A Simple Setup:

We really estimate:

$$\widehat{Y}_i = \alpha + \tau X_i + e_i$$

A Simple Setup:

We really estimate:

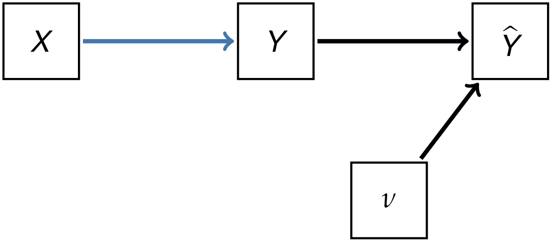
$$\widehat{Y}_i = \alpha + \tau X_i + e_i$$

$$\widehat{Y}_i = Y_i + \nu_i$$

$$\widehat{\tau} = \tau + \frac{cov(X, e)}{var(X)} + \frac{cov(X, \nu)}{var(X)}$$

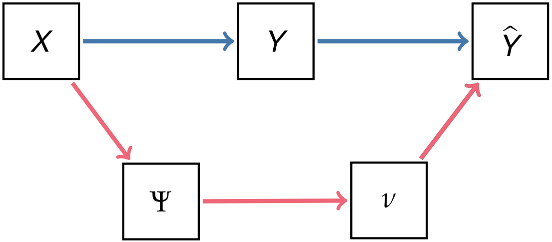
Prediction error can bias causal inference–RCT

- X: Unconditional Cash Transfer, Y: Forest cover
- $\hat{Y}$: RS measure of forest cover, ν: measurement error
- Want to estimate effect, τ, of X on Y. Are able to estimate effect of X on $\widehat{Y}$, $\tilde{\tau}$.



Prediction error can bias causal inference–RCT

- X: Unconditional Cash Transfer, Y: Forest cover
- $\hat{Y}$: RS measure of forest cover, ν: measurement error
- Want to estimate effect, τ, of X on Y. Are able to estimate effect of X on $\widehat{Y}$, $\tilde{\tau}$.
- Problem: ψ is irrigated cropland, more often misclassified as tree cover.



Idea for a simple solution:

$$\hat{\tau} = \tau + \frac{cov(X, e)}{var(X)} + \frac{cov(X, \nu)}{var(X)}$$

Make sure $cov(X, \nu) = 0$!

Prediction Powered Inference (Angelopoulos et al., 2023)

Consider a linear regression of treatment status on prediction errors:

- Note we estimate this regression in the labelled observations $j \in S$

$$\nu_j = \gamma X_j + \epsilon_j$$

Prediction Powered Inference (Angelopoulos et al., 2023)

Consider a linear regression of treatment status on prediction errors:

- Note we estimate this regression in the labelled observations $j \in S$

$$\nu_j = \gamma X_j + \epsilon_j$$
$$\hat{\gamma} = \frac{cov(X, \nu)}{var(X)}$$

This is the bias term in our estimate of τ!

A bias test

$$\text{Is } \gamma = \frac{cov(X, \nu)}{var(X)} = 0?$$

- Standard confidence intervals of $\hat{\gamma}$ tell us whether we can rule out large biases
- Using straightforward power analysis techniques, we can calculate the minimum detectable bias
    - How many points would you have to verify in order to be confident in your treatment effect estimates?
    - Can use high resolution imagery, hand labelling

Predict-then-Debias

$\hat{\tau}$ is estimated using predictions, $\hat{\gamma}$ is estimated in labelled set:

$$\hat{\tau}^{ppi} = \hat{\tau} - \hat{\gamma} \tag{1}$$

Consistent if $\hat{\gamma} \rightarrow_p E(\gamma)$.

- Satisfied if labelled set is representative
- Works for multivariate X
- Optimal weights for ground truth/predicted data (Kluger et al., 2025).
- Variance has 2 components, $\hat{\tau}$ and $\hat{\gamma}$
  - Both are constrained by accuracy of original predictions

Predict-then-Debias

$\hat{\tau}$ is estimated using predictions, $\hat{\gamma}$ is estimated in labelled set:

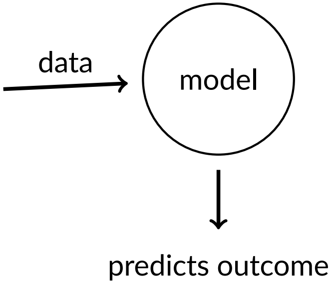
$$\hat{\tau}^{ppi} = \hat{\tau} - \hat{\gamma} \tag{1}$$

Consistent if $\hat{\gamma} \rightarrow_p E(\gamma)$.

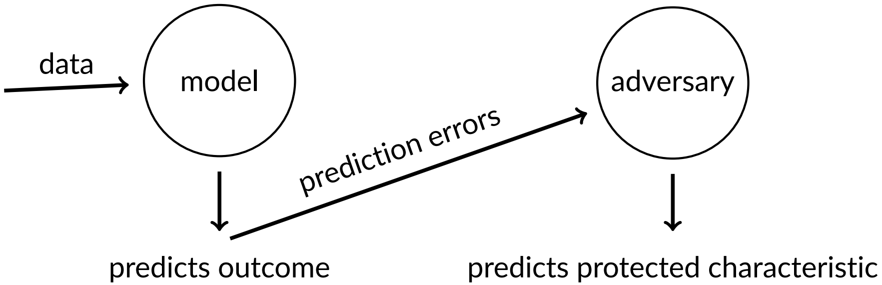
- Satisfied if labelled set is representative
- Works for multivariate X
- Optimal weights for ground truth/predicted data (Kluger et al., 2025).
- Variance has 2 components, $\hat{\tau}$ and $\hat{\gamma}$
  - Both are constrained by accuracy of original predictions

How can we create debiased predictions that are as efficient as possible?

Introduction to adversarial debiasing

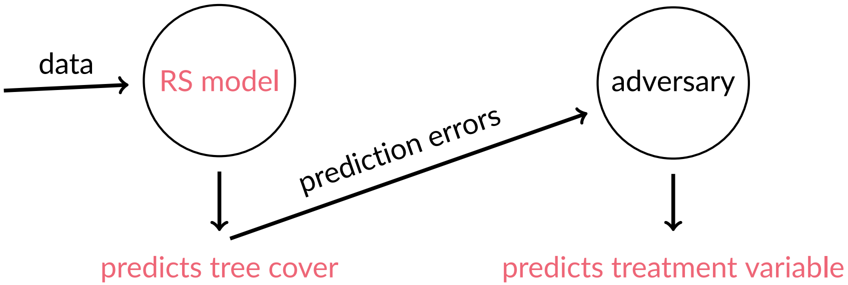


Introduction to adversarial debiasing



Used in computer science for making sure e.g. job/loan application ratings do not discriminate on the basis of race or gender (Zhang et al. 2018)

Introduction to adversarial debiasing



Used in computer science for making sure e.g. job/loan application ratings do not discriminate on the basis of race or gender (Zhang et al. 2018)

Adversarial Debiasing Formally

Loss function for a standard model: choose model weights (ω^*) such that

$$\omega^* = \arg\min_{\omega} L_p(\widehat{Y}(\omega), Y, k)$$

- L_p is a loss function (e.g. mean squared error)
- $\widehat{Y}$ are predictions of true outcomes Y, k are input features

Adversarial Debiasing Formally

Loss function for an adversarial model: choose model weights (ω^*) such that

$$\omega^* = \arg\min_{\omega} L_p(\widehat{Y}(\omega), Y, k) - \alpha L_a(\hat{X}(\mu), X, Y, \hat{Y}(\omega))$$
$$\text{subject to: } \mu \in \operatorname{argmin} L_a(\hat{X}(\mu), X, Y, \hat{Y}(\omega))$$

- L_p is a loss function (e.g. mean squared error)
- $\widehat{Y}$ are predictions of true outcomes Y, k are input features
- L_a is adversary's loss function
- $\widehat{X}$ are predictions of treatment variables X
- μ are adversary's model weights
- α is a tuning parameter: weight on the adversary's loss function

Unbiased Predictions

Consider an adversary that is our linear regression of treatment status on prediction errors:

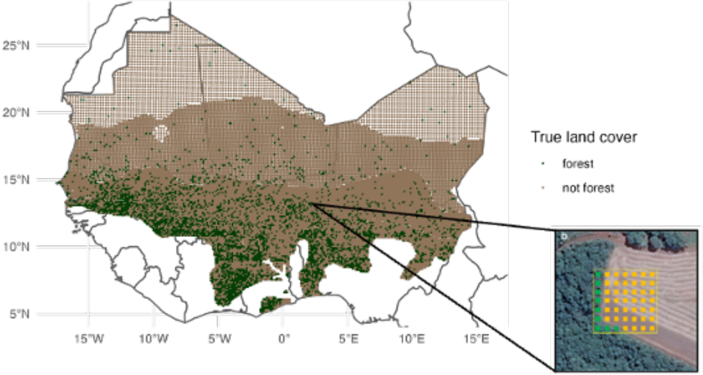
$$\nu_i = \gamma X_i + \epsilon_i$$
$$\widehat{\gamma} = \frac{cov(X, \nu)}{var(X)}$$

Intuitively, for a given accuracy, adversary MSE maximized when $\widehat{\gamma} = 0$

- Can also penalize bias directly (e.g. $|\gamma|$ or γ^2

Our application space

- Bastin et al. (2017) 23,000 hand-labeled points on dryland forest in W. Africa
  - motivated by Hansen et al. (2013) under-estimate of forest in dryland biomes
- Landsat 7 Surface Reflectance (7 bands)



Experiments: Can we recover τ using ML predictions of Y?

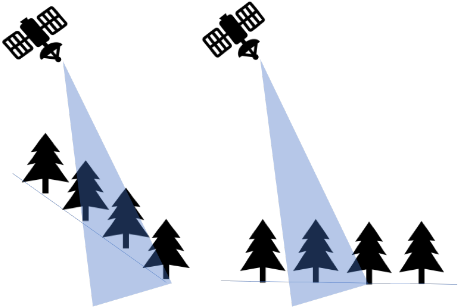
- Train baseline ML model (simple 3-layer neural net) using satellite data and 2/3 of labeled points
  - cross-fit the model to get OOS predictions for each labeled point
- Train adversarial model in the same way
- Use ground truth, baseline model measurements, and adversarial model measurements to estimate τ

Application 1: Simulated Data

- $\psi \sim$ Poisson (think e.g. slope)
- $p(X)$ decreasing in ψ (think e.g. infrastructure)
- Y: % Forest cover - randomly draw Bastin points and associated satellite data
    - Treatment effect $\tau = 0$ if using true labels

Application 1: Simulated Data

- $\psi \sim$ Poisson (think e.g. slope)
- $p(X)$ decreasing in ψ (think e.g. infrastructure)
- Y: % Forest cover - randomly draw Bastin points and associated satellite data
  - Treatment effect $\tau = 0$ if using true labels
  - One catch: if $\psi > 0$, we make the points look 'greener'



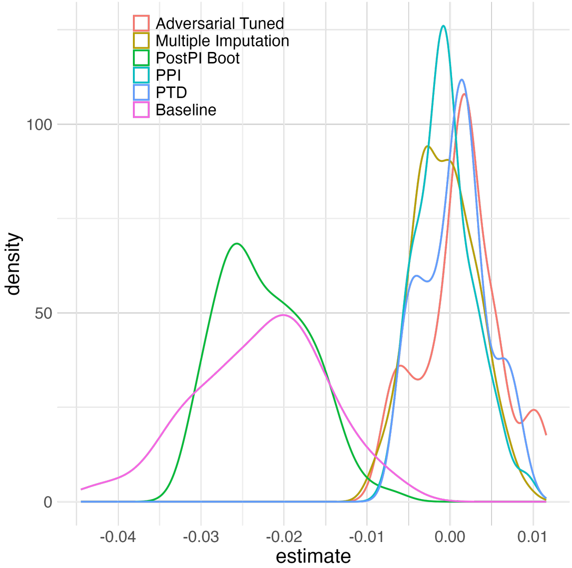
Application 1: Simulated Data

- $\psi \sim$ Poisson (think e.g. slope)
- $p(X)$ decreasing in ψ (think e.g. infrastructure)
- Y: % Forest cover - randomly draw Bastin points and associated satellite data
  - Treatment effect $\tau = 0$ if using true labels
  - One catch: if $\psi > 0$, we make the points look 'greener'
- Since high ψ points are less common, standard ML model learns that green usually means trees
- Debiased model notices these errors are correlated with X, does worse on low ψ points, better on high ψ points

Application 1: Simulated Data

- $\psi \sim$ Poisson (think e.g. slope)
- $p(X)$ decreasing in ψ (think e.g. infrastructure)
- Y: % Forest cover - randomly draw Bastin points and associated satellite data
  - Treatment effect $\tau = 0$ if using true labels
  - One catch: if $\psi > 0$, we make the points look 'greener'
- Since high ψ points are less common, standard ML model learns that green usually means trees
- Debiased model notices these errors are correlated with X, does worse on low ψ points, better on high ψ points
  - This is despite not knowing ψ!

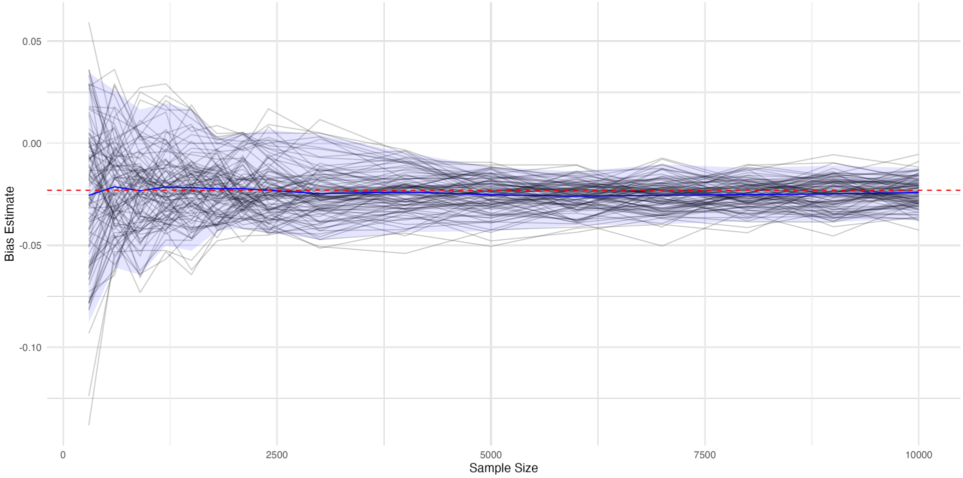
Application 1: Assessing bias



- Adversarial models correctly estimate $\tau = 0$
- Coefficient distributions from 100 bootstrapped iterations, models trained on 10,000 observations

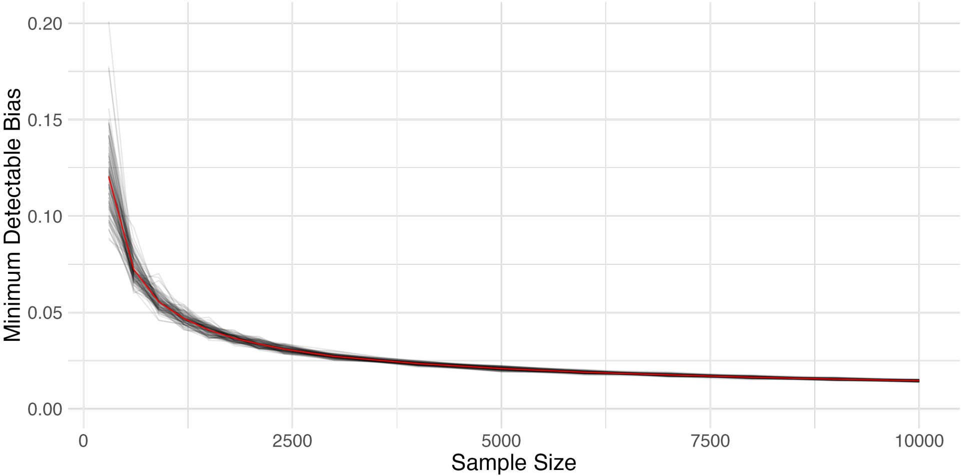
What if we progressively label more observations?

How well can we estimate bias if we have *n* labelled points:



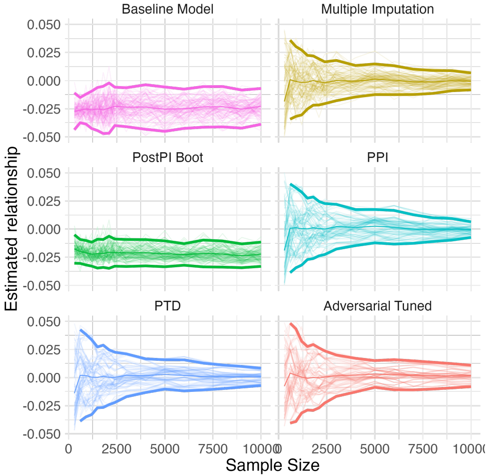
What if we progressively label more observations?

We can also bootstrap SEs for the bias statistic to estimate power:



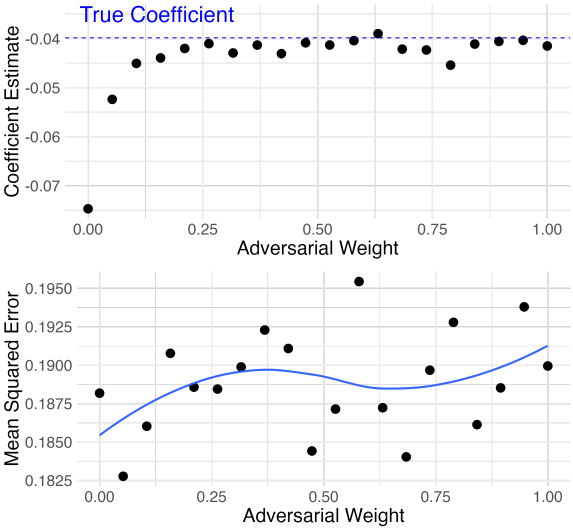
What if we progressively label more observations?

Adversarial models do well with sufficient training data:



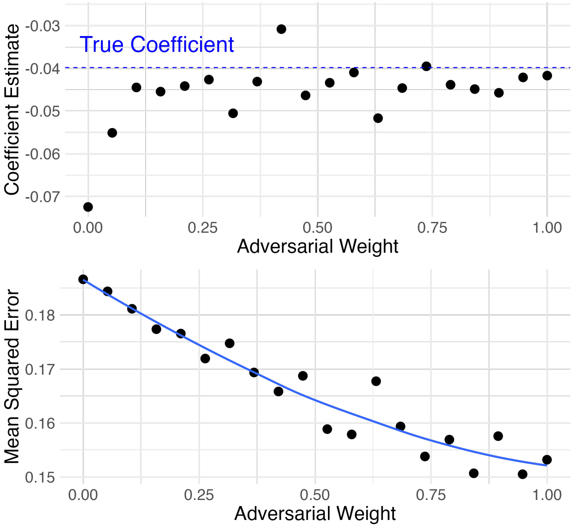
Tuning adversarial weight α: decreased bias, precision loss

Logistic Regression:



Tuning adversarial weight α: decreased bias, precision loss

Deep Neural Net:



Takeaways

- Caution when using proxies for outcome variables
  - Issue is not unique to remote sensing! Many uses of machine-learned outcomes as variables in other models (eg. text-based outcomes)
- If you are training a remote sensing model for a causal inference task you can:
  - Check to see if your estimates are biased by measurement error
  - See how many points you would need to label to detect bias of a certain size
  - Train a model that produces measurements that are suitable for your task
- Conduct measurement with the application in mind
  - A simple unbiased model can be better than a more powerful more accurate model

Takeaways

- Caution when using proxies for outcome variables
  - Issue is not unique to remote sensing! Many uses of machine-learned outcomes as variables in other models (eg. text-based outcomes)
- If you are training a remote sensing model for a causal inference task you can:
  - Check to see if your estimates are biased by measurement error
  - See how many points you would need to label to detect bias of a certain size
  - Train a model that produces measurements that are suitable for your task
- Conduct measurement with the application in mind
  - A simple unbiased model can be better than a more powerful more accurate model
- **Next:** An empirical research question where remote sensing data would be useful
  - Can we do better than randomly selecting which points to label?

Dumps

- Waste management often one of the largest line-items in low-income city municipal budgets: 20% of total expenditures Hoornweg and Bhada-Tata (2012)
- Estimated 90% of waste in low-income countries is disposed in unregulated landfills or burned
- Solid waste generation expected to increase 73% by 2050

Motivation: Trade

- Approximately 50% of plastic waste collected for recycling is traded internationally Kellenberg (2015)
  - The US exported $\sim$2 million tons of plastic scrap to 89 trade partners in 2016. About 8x US auto exports by weight!
- What are the welfare implications of the global trade in waste?

  "the economic logic behind dumping a load of toxic waste in the lowest wage country is impeccable" – Larry Summers, 1991

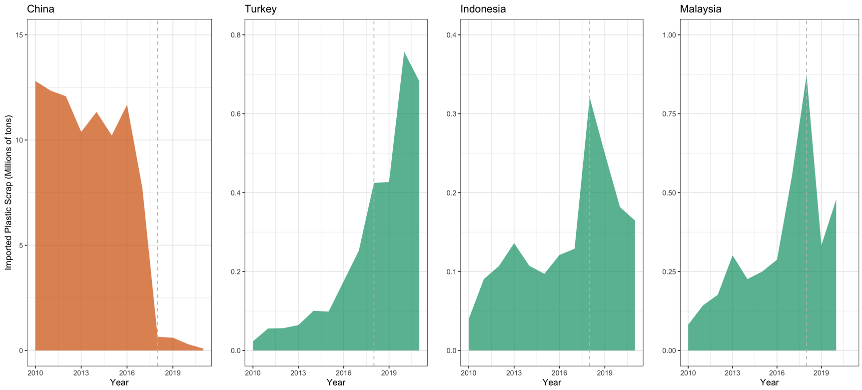
- Waste trade is heavily cross-subsidized
- Poor institutions as a source of comparative advantage? (Chichilnisky 1994)

The China Waste Import Ban

- China's National Sword Policy (also known as China's waste ban) was announced in mid-2017 and enacted in January 2018
- Prior to the ban, China handled around half of the world's traded recycling waste (around 70% of US exports and 95% of EU exports)
- After the ban, shippers diverted a significant fraction of this waste to countries across Southeast Asia and the rest of the world

Motivation

- This led to a huge increase in imported plastic waste in some countries



- Current discussions to limit trade in plastic waste under the Basel Convention

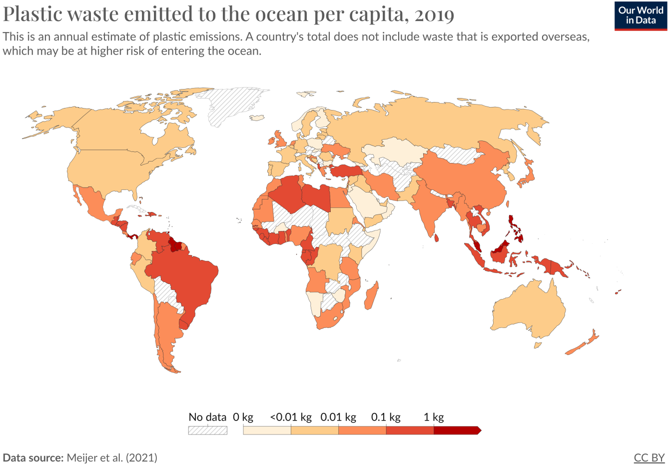
This Paper

- A method combining crowd-sourced data, machine learning, and econometric methods to create a globally-representative time-series of dumps for 2011-2023
  - Could be applied to other hard-to-study land uses

- Preliminary Findings:
  - Global ‘dump’ area increased 4x after the China Waste Ban
  - Increase is widespread - including in countries that saw imports fall
  - 1-2% of dump pixels are near a fire
    - Spikes around time of waste ban, but goes back down

- Future work to look at health/labor market outcomes

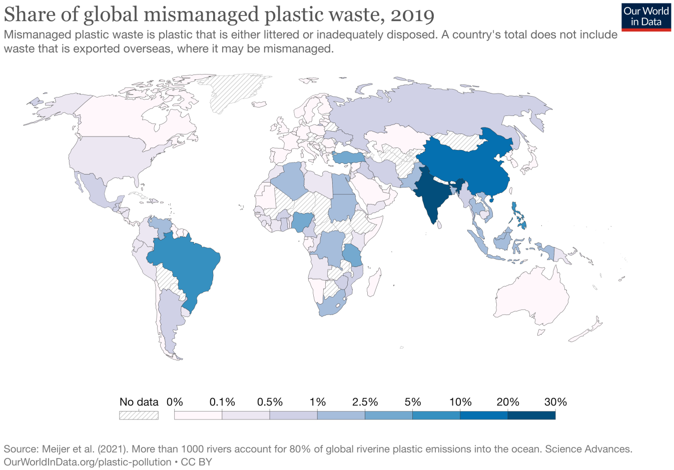
Existing Research on Dumps



Meijer et al. (2021). More than 1000 rivers account for 80% of global riverine plastic emissions into the Ocean. *Science Advances*.

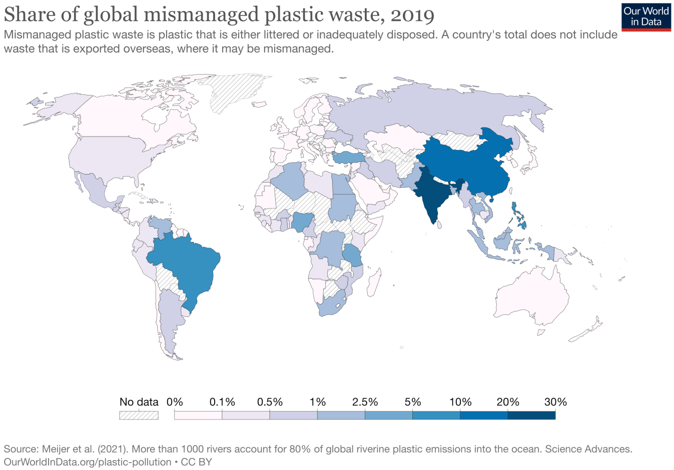
- Emissions = Population x MPW/capita x % Mismanaged

Mismanaged Waste



(Kaza et al., 2018). Solid waste data should be considered with a degree of caution due to...

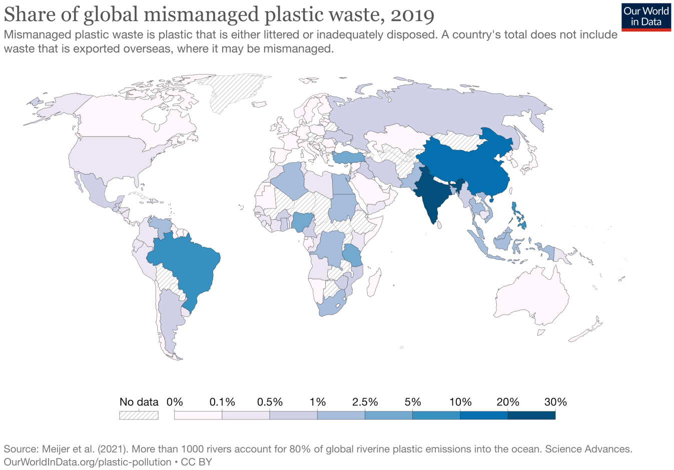
Mismanaged Waste



(Kaza et al., 2018). Solid waste data should be considered with a degree of caution due to...

- Undefined words or phrases

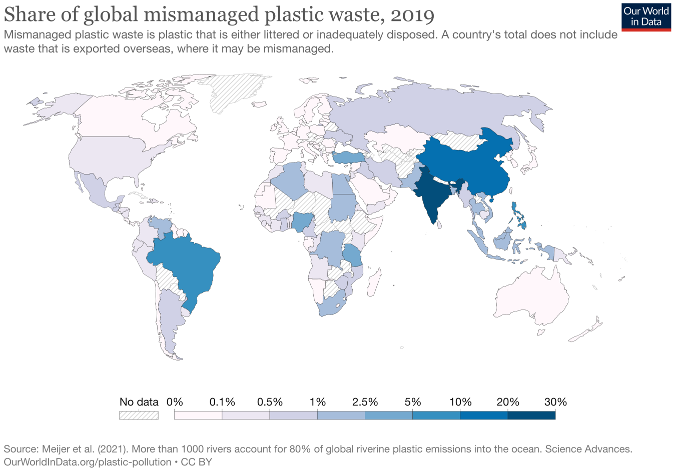
Mismanaged Waste



(Kaza et al., 2018). Solid waste data should be considered with a degree of caution due to...

- Undefined words or phrases
- Inconsistent or omitted units

Mismanaged Waste



(Kaza et al., 2018). Solid waste data should be considered with a degree of caution due to...

- Undefined words or phrases
- Inconsistent or omitted units
- Estimates made without basis

The Role of Trade

Law et al (2020). The United States' contribution of plastic waste to land and ocean. Science Advances

- To our knowledge, no quantitative estimates exist of the proportion of material exported for recycling that is ultimately discarded as waste or of the methods of disposal

The Role of Trade

Law et al (2020). The United States' contribution of plastic waste to land and ocean. Science Advances

- To our knowledge, no quantitative estimates exist of the proportion of material exported for recycling that is ultimately discarded as waste or of the methods of disposal
- we applied a credible range estimate of between 25 and 75% of plastic waste discarded during the processing of plastic and paper scrap that was inadequately managed in receiving countries that have greater than 20% inadequately managed waste.

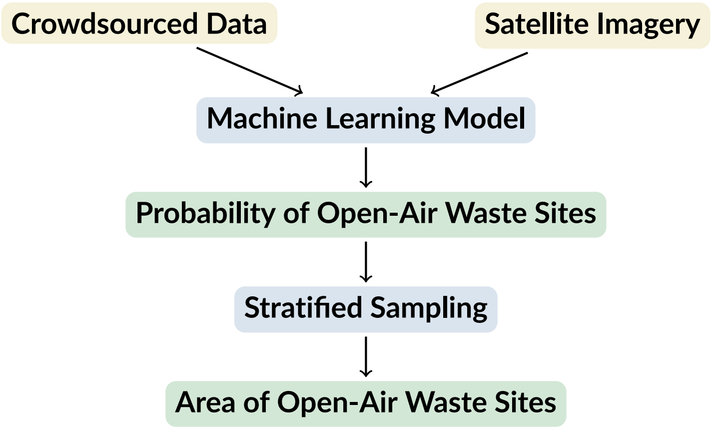
The Role of Trade

Law et al (2020). The United States' contribution of plastic waste to land and ocean. Science Advances

- To our knowledge, no quantitative estimates exist of the proportion of material exported for recycling that is ultimately discarded as waste or of the methods of disposal
- we applied a credible range estimate of between 25 and 75% of plastic waste discarded during the processing of plastic and paper scrap that was inadequately managed in receiving countries that have greater than 20% inadequately managed waste.
- By our upper-bound estimate, in 2016, the United States was the third largest contributor of mismanaged plastic waste to the coastal environment globally

Model-Assisted Stratified Sampling

- Our goal is to create a time series of open-air waste sites by country



Training Data Collection

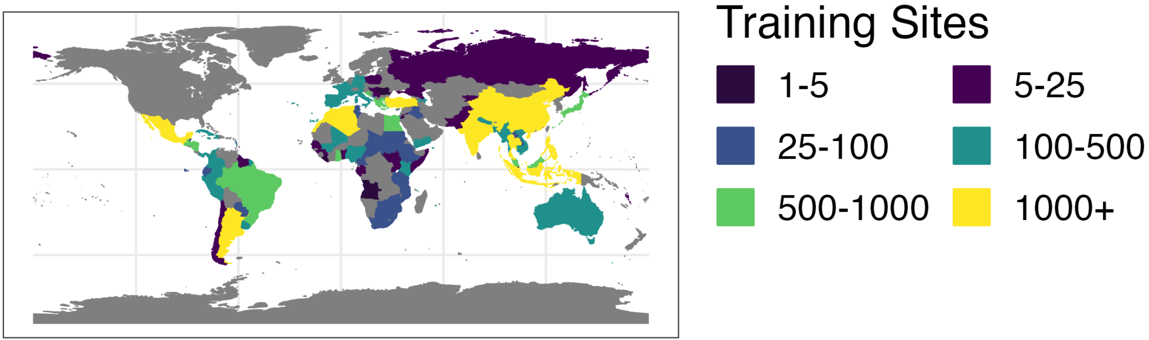
Crowdsourced training site collection:

- Partnership with NGOs, researchers, and activists; set up Atlas of Plastic Waste portal www.ban.org
- 270 unique open-air waste sites across 24 countries

Other Data Sources: D-waste, Earthrise, Greenpeace:

- $\approx$ 2,000 additional sites from 80+ countries
- 1,300 more sites from initial experiments with model

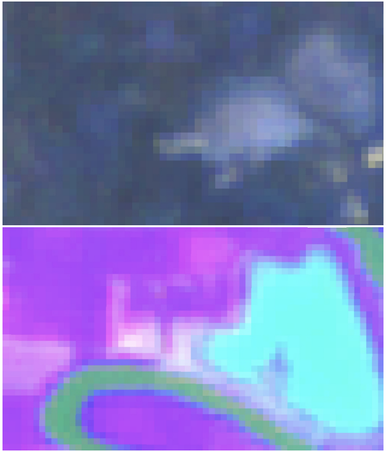
Full Training Data



Satellite Data

- Optical (Sentinel-2) and radar (Sentinel-1)
- 10m resolution, 11 bands + derived indices, 2 week overhead time

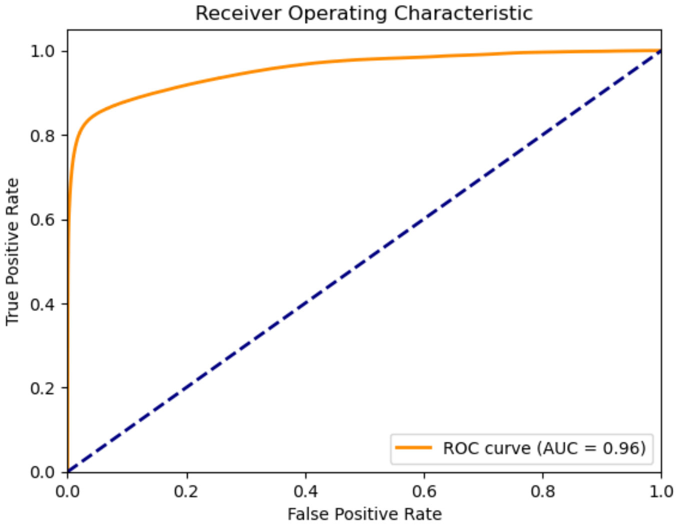
High Resolution Imagery



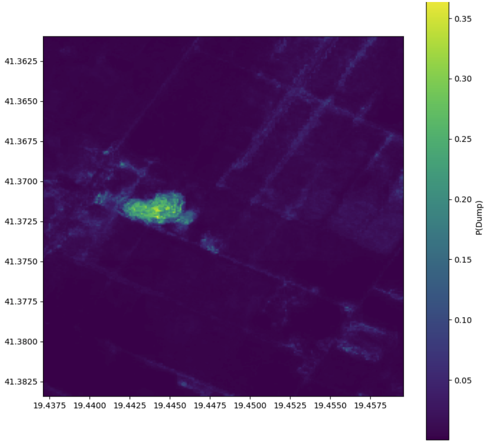
Sentinel 2 False Color

Machine Learning Model

- We train an XGBoost model on 80% of the clusters, evaluate accuracy on remaining 20%

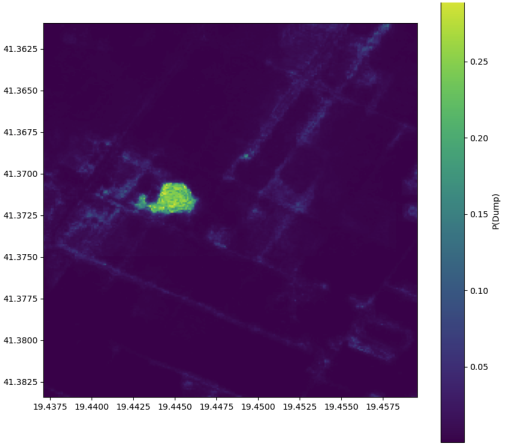


Machine Learning Model: Accuracy



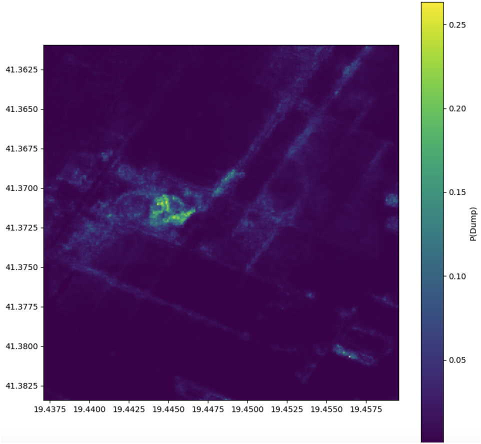
Durres, Albania 2018

Machine Learning Model: Accuracy



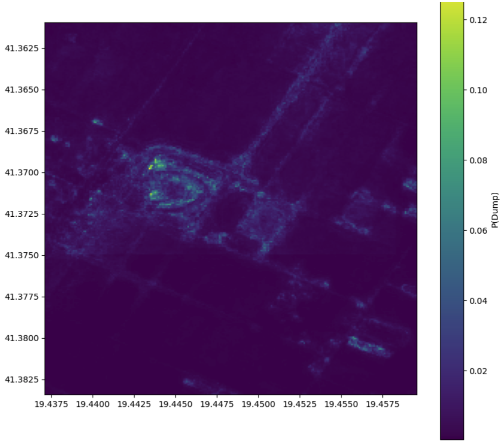
Durres, Albania 2019

Machine Learning Model: Accuracy



Durres, Albania 2020

Machine Learning Model: Accuracy



Durres, Albania 2021

Machine Learning Model: Bias

- Accuracy Measures: Hold-out clusters
  - Balanced Accuracy: 90%
  - Precision: 92%
  - Recall: 85%
- Hopelessly biased
  - Training data not representative
  - Image quality changes over time and between areas
  - Massively unbalanced classes exacerbate issues
    - If dumps make up 0.1% of landcover, and my accuracy varies by 0.05% between periods/countries, I could find a 50% increase in dumps when no change has occurred.

Machine Learning Model: Bias

- Accuracy Measures: Hold-out clusters
  - Balanced Accuracy: 90%
  - Precision: 92%
  - Recall: 85%
- Hopelessly biased
  - Training data not representative
  - Image quality changes over time and between areas
  - Massively unbalanced classes exacerbate issues
    - If dumps make up 0.1% of landcover, and my accuracy varies by 0.05% between periods/countries, I could find a 50% increase in dumps when no change has occurred.
- Yet still useful...

Model Assisted Active Sampling

Standard Horvitz-Thompson unequal probability estimator of bias, with weights $w_i = 1/\pi_i$, where π_i is inclusion probability:

$$\widehat{\mu}^{maas} = \frac{1}{N}\sum_{i \in N} p_i + \frac{1}{N}\sum_{i \in N} w_i s_i \nu_i \tag{2}$$

How can we choose which points to sample? Select inclusion probabilites π_i to solve:

$$\min_{\pi_i} Var(\widehat{\mu}^{maas}) \tag{3}$$

$$\text{such that } \sum_i \pi_i = S$$

Model Assisted Active Sampling

First term is constant since we observe whole population. With iid assumption, variance of the second term is:

$$Var(\widehat{\mu}^{maas}) = \frac{1}{N^2} \sum_{i \in N} Var(w_i s_i (p_i - D_i)) \tag{4}$$
$$= \frac{1}{N^2} \sum_{i \in N} \frac{p_i(1 - p_i)}{\pi_i}$$

Taking FOC and solving for π gives closed for solution for optimal sampling inclusion probabilities:

$$\pi_i = S \frac{\sqrt{p_i(1 - p_i)}}{\sum_j \sqrt{p_i(1 - p_i)}} \tag{5}$$

Model Assisted Active Sampling

$$\pi_i = S \frac{\sqrt{p_i(1-p_i)}}{\sum_j \sqrt{p_i(1-p_i)}}$$

This expression is not new (Neyman (1934))

- What is new is using ML model to get predictions of p_i
- If model is biased, it doesn't affect consistency, just efficiency

Model Assisted Active Sampling

Plug inclusion probabilities back into formula for the variance:

$$Var(\widehat{\mu}^{maas}) = \frac{1}{N^2 S}\left[\sum_{i \in N} \sqrt{p_i(1-p_i)}\right]^2 \tag{6}$$

Compare to uniform sampling:

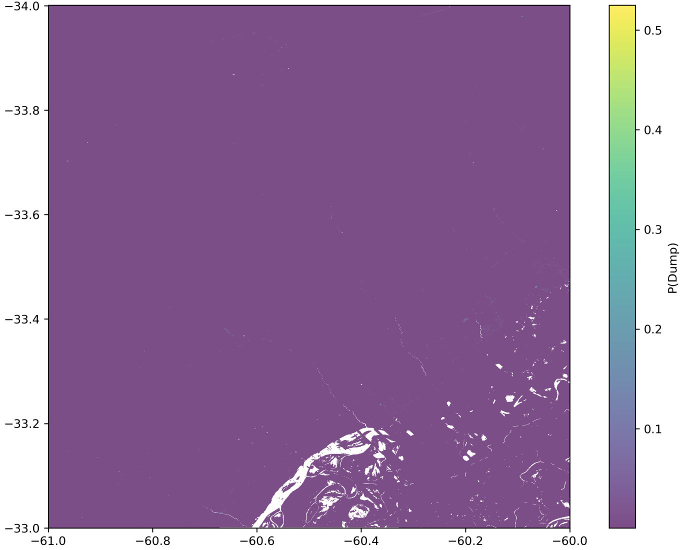
$$Var(\widehat{\mu}^{ppi}) = \frac{1}{NS}\sum_{i \in N} p_i(1-p_i)$$

$$\frac{Var(\widehat{\mu}^{maas})}{Var(\widehat{\mu}^{ppi})} = \frac{1}{1 + CV_\sigma^2} \leqslant 1$$

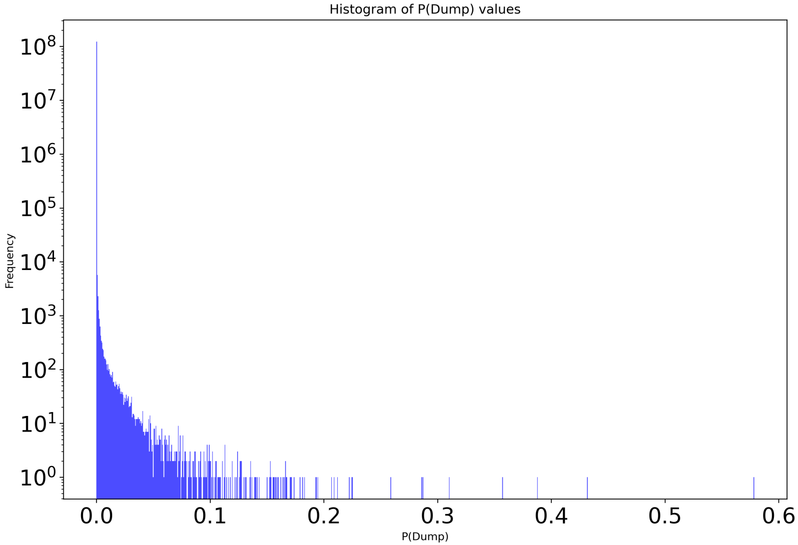
Optimal Sampling: Example Tile

Optimal Sampling: Example Tile

Optimal Sampling: Example Tile



Optimal Sampling: Example Tile

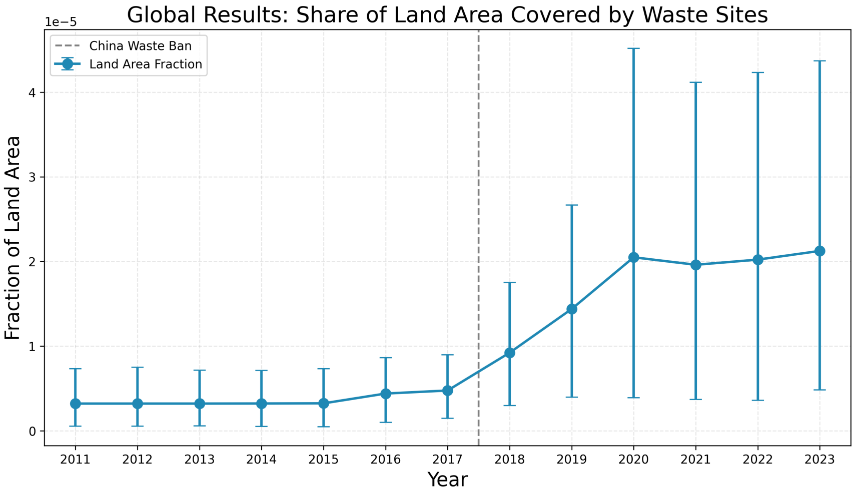


Optimal Sampling: Example Tile

Theoretical efficiency gains:

- For the same sample size, variance decreases by 99.88% relative to uniform sampling
- Might not be realistic:
  - Some scenes are more homogeneous
  - p_i probably (definitely) not perfectly calibrated
- still, potential gains are large
- Apply a similar approach to estimate changes. Need to estimate p_{chg} based on vector of probabilities over time.

Time Trend in Dumps



Thank you!

matthew.gordon@psemail.eu

https://sites.google.com/view/mdgordon/

Thanks to Anna Papp, Luke Sanford, Megan Ayers, Eliana Stone, Marion Chadal, and STEG, the IGC, the Minderoo Foundation, Earthrise, and the Basel Action Network for supporting the work.