Scalable Vision-Guided Crop Yield Estimation

Harrison H Li[1*], Medhanie Irgau[2*], Nabil Janmohamed[3], Karen Solveig Rieckmann[3], David B. Lobell[4]

[1]Department of Mathematics, Harvey Mudd College, Claremont, CA USA
[2]Department of Computer Science, Stanford University, Stanford, CA USA
[3]Pula Advisors AG, Mollis, Switzerland
[4]Department of Earth System Science and Center on Food Security and the Environment
Stanford University, Stanford, CA USA
harrli@hmc.edu, mirgau@stanford.edu, nabil@pula.io, krieckmann@pula.io, dlobell@stanford.edu

Abstract

Precise estimation and uncertainty quantification for average crop yields are critical for agricultural monitoring and decision making. Existing data collection methods, such as crop cuts in randomly sampled fields at harvest time, are relatively time-consuming. Thus, we propose an approach based on prediction-powered inference (PPI) to supplement these crop cuts with less time-consuming field photos. After training a computer vision model to predict the ground truth crop cut yields from the photos, we learn a "control function" that recalibrates these predictions with the spatial coordinates of each field. This enables fields with photos but not crop cuts to be leveraged to improve the precision of zone-wide average yield estimates. Our control function is learned by training on a dataset of nearly 20,000 real crop cuts and photos of rice and maize fields in sub-Saharan Africa. To improve precision, we pool training observations across different zones within the same first-level subdivision of each country. Our final PPI-based point estimates of the average yield are provably asymptotically unbiased and cannot increase the asymptotic variance beyond that of the natural baseline estimator — the sample average of the crop cuts — as the number of fields grows. We also propose a novel bias-corrected and accelerated (BCa) bootstrap to construct accompanying confidence intervals. Even in zones with as few as 20 fields, the point estimates show significant empirical improvement over the baseline, increasing the effective sample size by as much as 73% for rice and by 12-23% for maize. The confidence intervals are accordingly shorter at minimal cost to empirical finite-sample coverage. This demonstrates the potential for relatively low-cost images to make area-based crop insurance more affordable and thus spur investment into sustainable agricultural practices.

Code — https://github.com/medhanieirgau/scalable-vision-guided-crop-yield-estimation

Datasets — https://doi.org/10.5281/zenodo.17626117

Extended version — https://www.arxiv.org/abs/2511.12999

[*]These authors contributed equally.

1 Introduction

Improved agricultural productivity is critical to achieve greater food security and economic growth worldwide. A significant impediment to progress in many countries has been the lack of reliable data on cropping outcomes such as yield, which can impede many forms of investments. A prime example is that farmers often lack access to affordable and effective crop insurance, which would enable higher levels of investment and risk-taking needed to raise productivity. Effective insurance, in turn, relies on the ability to accurately assess cropping conditions and yields across large and heterogeneous cropping regions. Historically, this information has proven to be difficult and costly to obtain.

Recent progress has been made by conducting systematic large-scale crop cutting measurements during the harvest season. These crop cuts are averaged over pre-defined regions called zones to provide an objective measure of average yields underlying area-based insurance products.

The motivation for the present work comes from the promise of photographs or aerial imagery as more affordable and scalable approaches to yield estimation (Khaki et al. 2021; Li et al. 2022). While they have commonly been promoted as potential alternatives to ground-based measures like crop cuts, they often explain only half or less of true yield variability in complex smallholder farming settings (Darra et al. 2023) and can be biased (Lobell et al. 2020). This makes using such yield proxies in the place of ground measurements inadequate for insurance and re-insurance providers. We thus seek to instead supplement ground measurements with images — specifically, photos of fields taken at harvest time — to increase the effective sample size without introducing bias.

To do so, we propose leveraging recent statistical methods to efficiently combine small amounts of expensive, error-free data with larger amounts of cheaper but less accurate measures. After training a computer vision model to predict yields from photos, we post-process these predictions by adopting ideas from the framework of prediction-powered inference (PPI). Importantly, we are able to guarantee that in large samples, our zone-level average yield estimates remain unbiased and cannot have larger variance than using the crop cuts alone, regardless of the performance of the computer vision model. However, a better computer vision

Crop	Country	Harvest year	Zones	Fields
Rice	Nigeria	2022	29	826
Maize	Zambia	2023	126	3759
Maize	Zambia	2024	342	10727
Maize	Zimbabwe	2024	87	4173

Table 1: The number of eligible zones (i.e., zones with at least 20 fields) and the total number of fields across the eligible zones in the dataset by country and harvest year

model enables a lower asymptotic variance for the final estimator, to a degree that we can quantify. We empirically validate these theoretical guarantees under relatively small sample sizes using an extensive dataset from Pula, a provider of area-based crop insurance products in sub-Saharan Africa.

The remainder of the paper is organized as follows. Section 2 describes the dataset and preprocessing steps. Section 3 reviews the prediction-powered inference framework and introduces our estimation approach. Section 4 details the photo-based yield prediction models underlying our methodology, and Section 5 presents some innovations that leverage the nature of area-based insurance data to improve learning of the control function that underpins PPI. Section 6 reports empirical findings and Section 7 concludes.

2 Data

Our dataset contains field-level observations from four (country, harvest year) pairs: rice fields from (Nigeria, 2022) and maize (corn) fields from (Zambia, 2023), (Zambia, 2024), and (Zimbabwe, 2024). For each field, two random crop cuts were taken at harvest time. Their average is treated as the ground truth yield, in units of $\texttt{mt ha}^{-1}$. Photographs were also manually captured for each field. Some examples are provided in Fig. 1. The data from Zambia are commercially sensitive; however, we provide the complete data from Nigeria and Zimbabwe in the link following the abstract of this paper. Currently, we do not have access to photos from additional fields; in our evaluations we use bootstrapping to simulate the desired future setting where there are additional "unlabeled" fields with photos but not crop cuts (Section 6).

Each field is assigned to a unique zone for area-based insurance purposes. Since we are interested in improving estimates of zone-level mean yields, to reduce noise in our performance metrics we filter our dataset to only zones with at least 20 observations. Table 1 provides summary statistics for the final filtered dataset.

Our released dataset excludes approximately 80 out of ∼7,000 field photos in Nigeria and Zimbabwe with identifiable individuals or other sensitive content.

3 Prediction-powered Inference

We now provide some background on prediction-powered inference (PPI) and show the ideas can be leveraged in our context to effectively combine field photos with crop cuts for mean yield estimation.

For clarity, we begin by considering the data in a single zone, although we will end up pooling data from multiple

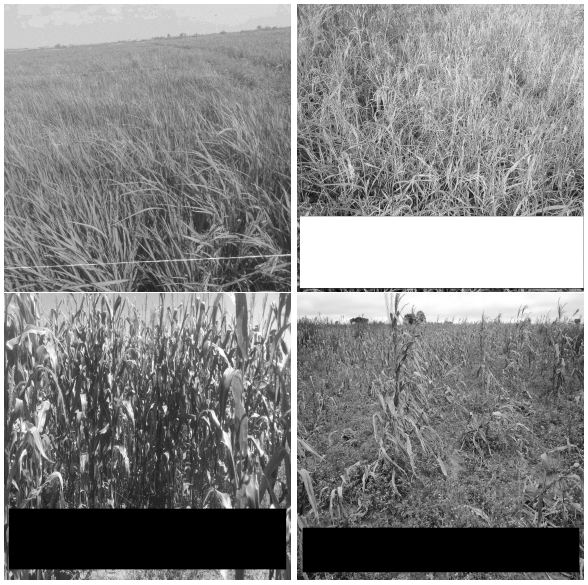

Figure 1: (Top row) Sample rice field photos from Nigeria in harvest year 2022. The fields have ground truth (crop cut) yields of $4.7\ \texttt{mt ha}^{-1}$ (left) and $0.0088\ \texttt{mt ha}^{-1}$ (right). (Bottom row) Sample maize field photos from Zimbabwe in harvest year 2024. The fields have ground truth (crop cut) yields of $9.1\ \texttt{mt ha}^{-1}$ (left) and $0.0061\ \texttt{mt ha}^{-1}$ (right). Some identifying information is blacked out.

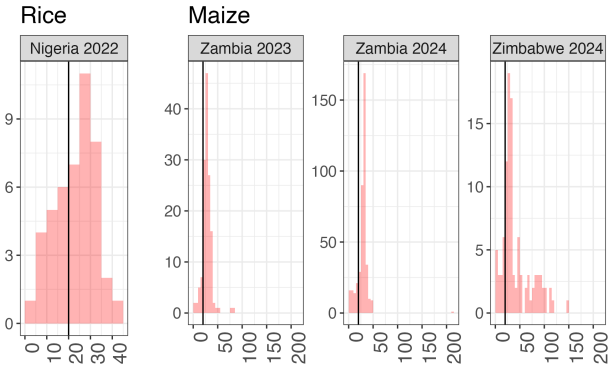

Figure 2: Histograms of the number of fields within each zone in our dataset, separated by country and harvest year

zones to improve the finite sample performance of our final zone-level average yield estimates (Section 5). The data in a zone consists of "labeled" fields $i = 1, \ldots, n$ where we observe the ground truth yields ("labels") Y_i, photos V_i, and covariates X_i, along with "unlabeled" fields $i = n+1, \ldots, n+N$ for which only the photos and covariates are available. As noted in Section 2, there are no unlabeled samples in our dataset, but we present our methodology in anticipation of a future application with additional unlabeled samples that can be cheaply collected. For our evaluations, we only use latitude and longitude — readily available in our dataset — as our two covariates X_i, though future im-

Algorithm 1: PPI++

Require: Labeled data $\{(Y_i, W_i)\}_{i=1}^{n}$, unlabeled data $\{W_i\}_{i=n+1}^{n+N}$, control function $f(\cdot)$

$\hat{\theta}_{\text{lbl}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} Y_i$

$\bar{f}_n \leftarrow \frac{1}{n}\sum_{i=1}^{n} f(W_i)$

$\bar{f}_N \leftarrow \frac{1}{N}\sum_{i=n+1}^{n+N} f(W_i)$

$\bar{f} \leftarrow \frac{1}{n+N}\sum_{i=1}^{n+N} f(W_i)$

$\widehat{\text{cov}}(Y, f(W)) \leftarrow \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{\theta}_{\text{lbl}})(f(W_i) - \bar{f}_n)$, $\widehat{\text{Var}}(f(W)) \leftarrow \frac{1}{n+N-1}\sum_{i=1}^{n+N}(f(W_i) - \bar{f})^2$

$\hat{\lambda} \leftarrow \frac{N}{n+N}\frac{\widehat{\text{cov}}(Y, f(W))}{\widehat{\text{Var}}(f(W))}$

return $\hat{\theta}_{\text{lbl}} - \hat{\lambda}\left(\bar{f}_n - \bar{f}_N\right)$

plementations could incorporate more covariates. The estimand of interest is $\theta = \mathbb{E}[Y_i]$. We access the information in the photos V_i through predictions $\hat{Y}_i = g(V_i)$ from a computer vision model g. We defer the details of how to train $g(\cdot)$ to Section 4; for now, we treat it as a black box.

Algorithm 1 details how to compute the proposed PPI++ estimator $\hat{\theta}_{\text{PPI++}}$ for θ given these data. Following Angelopoulos, Duchi, and Zrnic (2024), the PPI++ estimator takes the form

$$\hat{\theta}_{\text{PPI++}} = \hat{\theta}_{\text{lbl}} - \hat{\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \frac{1}{N}\sum_{i=n+1}^{n+N} f(W_i)\right) \quad (1)$$

where $\hat{\theta}_{\text{lbl}} = n^{-1}\sum_{i=1}^{n} Y_i$ is simply the sample average of the yields of the labeled fields. The additional components of the PPI++ estimator (1) are the control function f — designed to predict the yields Y from the combination $W = (\hat{Y}, X)$ of photo model predictions and covariates — and the coefficient $\hat{\lambda}$, which can depend on f. If the observations in the labeled and unlabeled fields are independent and drawn from the same distribution, then provided $\hat{\lambda}$ converges to a non-random limit, the PPI++ estimator $\hat{\theta}_{\text{PPI++}}$ is asymptotically unbiased for θ for any control function f. While this assumption can be restrictive in some applications, an area-based crop insurer can ensure it holds by design by choosing which fields receive crop cuts uniformly at random. We thus assume it holds throughout this paper.

Angelopoulos, Duchi, and Zrnic (2024) showed that given any square integrable control function f, the coefficient

$$\hat{\lambda} = \frac{N}{n+N}\frac{\widehat{\text{cov}}(Y_i, f(W_i))}{\widehat{\text{Var}}(f(W_i))} \quad (2)$$

where $\widehat{\text{cov}}$ and $\widehat{\text{Var}}$ denote the usual sample-based estimates of covariance and variance, respectively, minimizes the asymptotic variance of $\hat{\theta}_{\text{PPI++}}$ as the sample sizes n and N grow. The choice of $\hat{\lambda}$ in (2) ensures the estimator $\hat{\theta}_{\text{PPI++}}$ asymptotically dominates (i.e., cannot have higher asymptotic variance than) $\hat{\theta}_{\text{lbl}}$ which ignores the photos and corresponds to setting $\hat{\lambda} = 0$ in (1), along with the original PPI estimator of Angelopoulos et al. (2023) which corresponds

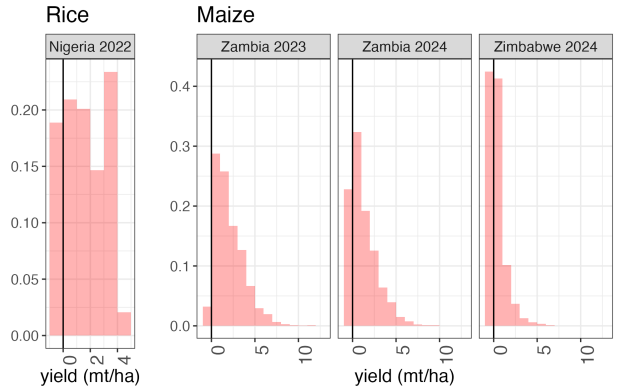

Figure 3: Histograms of field-level yields, separated by country and harvest year. All observations to the left of the solid vertical lines at 0 are exact zeros.

to setting $\hat{\lambda} = 1$ in (1). While recent work by Ji, Lei, and Zrnic (2025) has found that a "recalibrated PPI" (RePPI) estimator can further asymptotically dominate PPI++, we show in Appendix A that RePPI is in fact equivalent to our PPI++ estimator in our setting. All appendices are available in the extended version of our paper at the link indicated after the abstract.

It remains to make a prudent choice of the control function f. Let $\mu(w) = \mathbb{E}[Y \mid W = w]$ be the conditional mean outcome given the photo predictions and covariates (i.e., W). When $f = \mu$ we have $\hat{\lambda} = N/(n+N)$ from (2). Plugging these choices of $(f, \hat{\lambda})$ into (1) yields the augmented inverse propensity weighted (AIPW) estimator $\hat{\theta}_{\text{AIPW}}$, which is known to be semiparametrically efficient for this problem (Robins, Rotnitzky, and Zhao 1994; Angelopoulos, Duchi, and Zrnic 2024). This means that picking $(f, \hat{\lambda}) = (\mu, N/(n+N))$ minimizes the asymptotic variance not only among all estimators of the form (1), but also among all regular, asymptotically linear estimators for θ.

However, in reality μ is not known so we cannot simply "pick" $f = \mu$. Naturally, we can take f to be a data-driven estimate $\hat{\mu}$ of μ, learned via some supervised learning algorithm to predict Y from W. Semiparametric efficiency, however, would require this estimated $\hat{\mu}$ to be asymptotically equivalent to (i.e., consistent in mean square for) μ. With n generally on the order of dozens (Fig. 2), it is not prudent to assume that μ can be well approximated nonparametrically. This is an important argument in favor of using the PPI++ estimator in lieu of the AIPW estimator. The PPI++ and AIPW estimators differ in their choice of $\hat{\lambda}$. Whereas AIPW sets $\hat{\lambda} = N/(n+N)$ based on the assumption that $f \approx \mu$, the PPI++ estimator adaptively chooses $\hat{\lambda}$ via (2) to approximate the optimal choice for the f actually learned. The efficiency of AIPW, however, still incentivizes learning a control function f that is as close to μ as possible.

We also seek the quantify the uncertainty in our point estimate $\hat{\theta}_{\text{PPI++}}$. While $\hat{\theta}_{\text{PPI++}}$ satisfies a central limit theorem (e.g., Theorem 1 below), enabling asymptotically valid confidence intervals based on the normal distribution as

Algorithm 2: `PPBootBCa`

Require: Labeled data $\{(Y_i, W_i)\}_{i=1}^{n}$, unlabeled data $\{W_i\}_{i=n+1}^{n+N}$, control function $f(\cdot)$, significance level α, number of bootstrap samples B

Compute point estimate $\hat{\theta}_{\text{PPI++}} \leftarrow \texttt{PPI++}\left(\{(Y_i, W_i)\}_{i=1}^{n}, \{W_i\}_{i=n+1}^{n+N}, f(\cdot)\right)$
for $b = 1, \ldots, B$ **do**
  Get bootstrap samples $\left\{(Y_i^{(b)}, W_i^{(b)})\right\}_{i=1}^{n}$, $\left\{W_i^{(b)}\right\}_{i=n+1}^{n+N}$ by sampling with replacement
  Compute bootstrap point estimate $\hat{\theta}_{\text{PPI++}}^{(b)} \leftarrow \texttt{PPI++}\left(\left\{(Y_i^{(b)}, W_i^{(b)})\right\}_{i=1}^{n}, \left\{W_i^{(b)}\right\}_{i=n+1}^{n+N}, f(\cdot)\right)$
end for
$z_0 \leftarrow \Phi^{-1}\left(B^{-1}\sum_{b=1}^{B}\mathbf{1}\left[\hat{\theta}_{\text{PPI++}}^{(b)} \leqslant \hat{\theta}_{\text{PPI++}}\right]\right)$ // *Bias-correction parameter;* Φ *is* $\mathcal{N}(0,1)$ *CDF*
$z_L \leftarrow z_0 + \Phi^{-1}(\alpha/2), \quad z_U \leftarrow z_0 + \Phi^{-1}(1-\alpha/2)$
for $m = 1, \ldots, n+N$ **do**
  $\hat{\theta}^{(-m)} \leftarrow \texttt{PPI++}\left(\{(Y_i, W_i)\}_{i\in\{1,\ldots,n\}\setminus\{m\}}, \{W_i\}_{i\in\{n+1,\ldots,n+N\}\setminus\{m\}}, f(\cdot)\right)$
end for
$\bar{\theta} \leftarrow (n+N)^{-1}\sum_{m=1}^{n+N}\hat{\theta}^{(-m)}$
$u \leftarrow [\bar{\theta} - \hat{\theta}^{(-1)}, \ldots, \bar{\theta} - \hat{\theta}^{(-[n+N])}]$
$\gamma \leftarrow \frac{\sum_{m=1}^{n+N} u_m^3}{6(\sum_{m=1}^{n+N} u_m^2)^{3/2}}$ // *Acceleration parameter*
return $\left[\texttt{quantile}\left(\{\hat{\theta}_{\text{PPI++}}^{(b)}\}_{b=1}^{B}, \Phi\left(z_0 + \frac{z_L}{1-\gamma z_L}\right)\right), \texttt{quantile}\left(\{\hat{\theta}_{\text{PPI++}}^{(b)}\}_{b=1}^{B}, \Phi\left(z_0 + \frac{z_U}{1-\gamma z_U}\right)\right)\right]$

$n, N \to \infty$, ground truth yields tend to be substantially skewed and zero-inflated, particularly for maize (Fig. 3), so such intervals do not attain the desired coverage levels for small n. To ameliorate these issues, in Algorithm 2 we propose a nonparametric bias-corrected and accelerated (BCa) bootstrap (Efron 1987) to generate confidence intervals for θ. Confidence intervals from the BCa bootstrap satisfy desirable second-order asymptotic coverage properties and are designed to correct for issues like skewness in the data distribution (Efron and Hastie 2021). Algorithm 2 can be seen as a variant of the prediction-powered bootstrap `PPBoot` (Zrnic 2025). We find empirically that our BCa bootstrap exhibits superior performance over an alternative based on the bootstrap-t method as well as over both standard asymptotic confidence intervals based on the central limit theorem (CLT) and the percentile bootstrap intervals suggested by `PPBoot` (Appendix B.1).

4 Computer Vision Models

Our vision models are convolutional neural networks based on the ResNet-50 architecture (He et al. 2015), pretrained on ImageNet (Russakovsky et al. 2015) and fine-tuned to predict the ground truth maize or rice yield from crop cuts. To adapt each model for regression, we replace the final classification layer with a fully connected layer that outputs a single scalar representing the predicted yield. Training minimizes mean squared error (MSE) using the Adam optimizer.

We aim to learn a separate model for each (country, harvest year) pair. To ensure robust evaluation and avoid overfitting, we adopt a parallelized 5-fold cross-fitting approach. That is, for each (country, harvest year) pair, labeled data within each zone is split randomly into five folds and five models are trained in parallel. Each of these models is

Algorithm 3: Full procedure

Require: Ground truth yields $\{Y_{ji}\}_{i=1}^{n_j}$, photos $\{V_{ji}\}_{i=1}^{n_j+N_j}$, and covariates $\{X_{ji}\}_{i=1}^{n_j+N_j}$ for zones $j = 1, \ldots, J$; number of cross-fitting folds K

for zone $j = 1, \ldots, J$ **do**
  Partition indices $(j,1), \ldots, (j, n_j)$ randomly and as evenly as possible into folds $I_{j1}, \ldots, I_{jK}$
  $\mathcal{I}_k \leftarrow \mathcal{I}_k \cup I_{jk}, \quad k = 1, \ldots, K$
end for
$\mathcal{I} \leftarrow \cup_{k=1}^{K}\mathcal{I}_k$
for fold $k = 1, \ldots, K$ **do**
  Train photo model $\hat{g}^{(-k)}(\cdot)$ to predict Y_{ji} from V_{ji} using all observations with $(j,i) \in \mathcal{I} \setminus \mathcal{I}_k$
  Compute photo predictions on labeled observations: $\hat{Y}_{ji} \leftarrow \hat{g}^{(-k)}(V_{ji}), \quad (j,i) \in \mathcal{I}_k$
end for
Compute photo predictions on unlabeled observations by averaging all K models: $\hat{Y}_{ji} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\hat{g}^{(-k)}(V_{ji}), \quad j = 1, \ldots, J, \quad i = n_j+1, \ldots, n_j+N_j$
$W_{ji} \leftarrow (\hat{Y}_{ji}, X_{ji}), \quad \psi(W_{ji}) \leftarrow (1, \hat{Y}_{ji}, X_{ji})', \quad j = 1, \ldots, J, \quad i = 1, \ldots, n_j+N_j$
Partition zones $j = 1, \ldots, J$ into R sets $\mathcal{S}_1, \ldots, \mathcal{S}_R$ by study region
for study region $r = 1, \ldots, R$ **do**
  Learn coefficients $\hat{\beta}_r$ via 5-fold cross-validated LASSO to predict Y from $\psi(W)$ using all labeled observations in $\mathcal{S}_r$
  $f_r(\cdot) \leftarrow \hat{\beta}_r^{\top}\psi(\cdot)$
end for
`estimates` $\leftarrow []$, `CIs` $\leftarrow []$
for $r = 1, \ldots, R$ **do**
  for $j \in \mathcal{S}_r$ **do**
    $\hat{\theta}_{\text{PPI++},j} \leftarrow \texttt{PPI++}\left(\{(Y_{ji}, W_{ji})\}_{i=1}^{n_j}, \{W_{ji}\}_{i=n_j+1}^{n_j+N_j}, f_r(\cdot)\right)$
    `estimates.append` $\left(\hat{\theta}_{\text{PPI++},j}\right)$
    `CIs.append(` `PPBootBCa`$\left(\{(Y_{ji}, W_{ji})\}_{i=1}^{n_j}, \{W_{ji}\}_{i=n_j+1}^{n_j+N_j}, f_r(\cdot)\right)$
  end for
end for
return `estimates, CIs`

R^2	(NG, 2022)	(ZM, 2023)	(ZM, 2024)	(ZW, 2024)
Within-zone	0.198	0.145	0.143	0.261
Cross-zone	0.666	0.201	0.404	0.448

Table 2: R^2 scores of our photo model predictions per (country, harvest year) pair. The within-zone numbers are simple averages across all zones.

trained on the observations across all zones from four of the folds and then used to predict the crop cut yields from the photos in the held-out fold. Then each field has a predicted yield from exactly one of the five models, which was trained on the folds not containing that field.

At the end of an epoch, we aggregate such predictions to form a full set of validation predictions and compute an evaluation metric: the average within-zone R^2 score, computed from held-out predictions. Specifically, we compute the R^2 between predicted and true yields separately within each zone, and then take a simple average across all zones in the (country, harvest year) pair. This metric is sensitive to how well the model captures yield variation within individual zones, and directly governs the amount variance reduction we can expect from our methods (Section 5). Through training, we track the five model checkpoints with the highest validation average within-zone R^2 (one for each fold) and save these to make our final predictions. The within-zone R^2 estimates for these models are given for each (country, year) pair in Table 2. For reference, we also provide the substantially higher "cross-zone" R^2 estimates based on predictions across all zones.

We trained all models using the Adam optimizer for 10 epochs with an initial learning rate selected from a sweep over 1e-4, 2e-4, 3e-4, 5e-4, 1e-3, choosing the best based on validation average within-zone R^2 for each (country, harvest year) pair. A learning rate of 2e-4 was selected for (ZM, 2023) and (ZW, 2024), while a learning rate of 3e-4 was selected for (ZM, 2024) and (NG, 2022). All experiments were run on a cluster with 20 CPU cores, 20 GB RAM, and 5 NVIDIA A4000 GPUs, running Ubuntu 20.04.6 LTS. We used a batch size of 128 during training.

5 Learning the Control Function

The theoretical discussion in Section 3 suggests learning the control function f in each zone by using the field observations within the zone to predict the mean ground truth yield Y from the covariates X and the computer vision model predictions $\hat{Y}$. However, we can exploit the fact that we have observations across many zones to improve the stability of the learned f. While heterogeneity across zones means that pooling observations to learn f likely introduces asymptotic bias in learning the true optimal control function within each zone, it reduces the finite sample variance in the learned f, translating to improved finite sample precision for the downstream PPI estimator $\hat{\theta}_{\text{PPI++}}$. We emphasize that this bias does not translate to asymptotic bias in $\hat{\theta}_{\text{PPI++}}$, but can cause it to have a higher asymptotic variance.

We find the best empirical performance when learning a single control function f for each first-level administrative division in each country (states in Nigeria, provinces in Zambia and Zimbabwe). We assign each zone to a single such administrative division (hereafter "study region") based on plurality membership of the fields in that zone, with ties broken at random. Appendix B.2 provides additional results for when a single control function f is learned for all the zones in each (country, harvest year), or when a different control function is learned for each zone.

Another important decision point for learning f is the choice of learning algorithm. We found that learning f using regularized linear regression via the LASSO (Tibshirani 1996) leads to better results on our data than more complex learning algorithms such as random forests (Breiman 2001) (Appendix B.3). We choose the LASSO penalty factor via 5-fold cross validation as implemented by the `glmnet` package in R (Friedman, Hastie, and Tibshirani 2010). We include the photo model predictions $\hat{Y}_{ji}$ linearly, and the covariates X_{ji} are expanded to include interactions up to second order in the LASSO regression.

Algorithm 3 summarizes our proposed methodology from end to end. In Theorem 1 below, we formally justify that the final PPI++ estimator from Algorithm 3 attains the same asymptotic variance as an oracle PPI++ estimator that doesn't need to estimate the optimal coefficient λ from (2). The need for this result comes from our specific choices for learning the control function f, so that f can be viewed neither as fully nonrandom (Angelopoulos et al. 2023; Angelopoulos, Duchi, and Zrnic 2024) nor as fully compliant with the cross-PPI paradigm (Zrnic and Candès 2024; Ji, Lei, and Zrnic 2025). We provide a full proof of this result and a brief technical discussion in Appendix C.

Theorem 1. *In the setting of Algorithm 3, suppose that for each zone $j = 1, \dots, J$, the sample sizes n_j and N_j satisfy $n_j/N_j \to \rho_j \in (0, \infty)$ as $N_j \to \infty$, the labeled observations $\{(Y_{ji}, V_{ji}, X_{ji})\}_{i=1}^{n_j}$ are independent and identically distributed (i.i.d.) and independent of the unlabeled observations $\{(V_{ji}, X_{ji})\}_{i=n_j+1}^{n_j+N_j}$ (which are also i.i.d.), and the photos and covariates have the same distribution between the labeled and unlabeled datasets (i.e., $\{(V_{ji}, X_{ji})\}_{i=1}^{n_j+N_j}$ are i.i.d.). Further assume that the photo models $\hat{g}^{(-k)}(\cdot), k = 1, \dots, K$ converge in mean square to a common square integrable function $g(\cdot)$, $\mathbb{E}[\|X_{ji}\|_2^2] < \infty$, $\mathbb{E}[Y_{ji}^2] < \infty$, and $\hat{\beta}_{r(j)} \xrightarrow{p} \beta_{r(j)}$ as $N_j \to \infty$, where $r(j) \in \{1, \dots, R\}$ is the study region for zone j. Then for $\theta_j = \mathbb{E}[Y_{ji}]$ we have*

$$\sqrt{n_j}(\hat{\theta}_{\text{PPI++},j} - \theta_j) \xrightarrow{d} \mathcal{N}(0, V_j),$$

$$V_j = \rho_j \lambda_j^2 \text{Var}(f^*_{r(j)}(W_{ji})) + \text{Var}\left(Y_{ji} - \lambda_j f^*_{r(j)}(W_{ji})\right)$$

where $f^*_{r(j)}(W_{ji}) = (1, g(V_{ji}), X_{ji})\beta_{r(j)}$ and $\hat{\lambda}_j \xrightarrow{p} \lambda_j = \frac{\text{cov}\left(Y_{ji}, f^*_{r(j)}(W_{ji})\right)}{(1+\rho_j)\text{Var}(f^*_{r(j)}(W_{ji}))}$ if $\text{Var}(f^*(W_{ji})) > 0$.

Supposing that the limiting control function $f^*_{r(j)}$ defined in the statement of Theorem 1 is in fact the conditional

mean function $\mu_j(\cdot) = \mathbb{E}_j(Y \mid W = \cdot)$ (here the subscript j denotes the expectation is taken over the distribution of the fields in zone j) we can take the ratio of the variance $\mathrm{Var}(Y_{ji})/n_j$ V_j of $\hat{\theta}_{\text{lbl}}$, the sample average of the labeled field yields, to V_j/n_j, the asymptotic variance of our PPI++ estimator from Theorem 1, to compute an asymptotic relative efficiency in terms of

$$R_j^2 = \frac{\mathrm{cov}(\mu_j(W_{ji}), Y_{ji})}{\mathrm{Var}(Y_{ji})} = \frac{\mathrm{Var}(\mu_j(W_{ji}))}{\mathrm{Var}(Y_{ji})} = 1 - \frac{\mathrm{Var}(Y_{ji} - \mu_j(W_{ji}))}{\mathrm{Var}(Y_{ji})},$$

the R^2 of the oracle prediction function μ_j for predicting the ground truth yields Y:

$$\begin{aligned}
\text{RE} &= \frac{\mathrm{Var}(Y_{ji})/n_j}{V_j/n_j} \\
&= \left(\rho_j \lambda_j^2 \frac{\mathrm{Var}(\mu_j(W_{ji}))}{\mathrm{Var}(Y_{ji})} + \frac{\mathrm{Var}(Y_{ji} - \lambda_j \mu_j(W_{ji}))}{\mathrm{Var}(Y_{ji})}\right)^{-1} \\
&= \left(1 + R_j^2 \left[\rho_j \lambda_j^2 - 1 + (1-\lambda_j)^2\right]\right)^{-1} \\
&\approx \left(1 - R_j^2 \frac{N_j}{N_j + n_j}\right)^{-1} \qquad (3)
\end{aligned}$$

where the last step follows from $\lambda_j = (1 + \rho_j)^{-1}$ and $\rho_j \approx n_j/N_j$. Equation (3) shows how higher signal in the photos (i.e., greater attainable zone-level R^2) directly improves the performance of our downstream PPI++ estimator, justifying our use of within-zone R^2 as our evaluation metric in checkpointing our computer vision models, as discussed in Section 4. We also observe from (3) that full potential of the photos is harnessed when the number of unlabeled fields is much larger than the number of labeled fields ($N_j \gg n_j$).

6 Results

We evaluate Algorithm 3 using our dataset by sampling with replacement to generate synthetic datasets respecting the observed distribution of ground truth yields and photos. The code repository linked after the abstract reproduces these experiments. For each zone, we draw a random sample of n fields and photos with replacement (where n is the number of fields in the zone) to form synthetic labeled data. By the bootstrap principle, for evaluation purposes the (sample) average ground truth yield of the original fields can be viewed as the unknown estimand of interest θ. We then take an independent sample of $N = 4n$ photos with replacement to form the synthetic unlabeled data. The larger amount of unlabeled data reflects the desired future setting with abundant field photos and limited crop cuts. We repeat all of these sampling steps 10 times per zone to get more precise performance estimates. For computational efficiency, we do not retrain the vision models across resamples, but we do refit the control functions via the cross-validated LASSO.

We compare alternative estimators including AIPW (`aipw`), the original PPI estimator that takes $\hat{\lambda} = 1$ in (1) (`ppi`), and a variant of PPI++ that ignores the photos (`nophoto`) — instead only adjusting for latitude and longitude — to our proposed method based on the PPI++ estimator (`ppipp`). Confidence intervals for the average yield based on each of these estimators are computed following Algorithm 2 with $B = 1000$, replacing the calls to the function `PPI++` of Algorithm 1 with calls to functions computing the relevant estimator as needed. Our evaluation metrics for each estimator are relative efficiencies comparing with $\hat{\theta}_{\text{lbl}}$, the sample average of the labeled fields. We consider both an "MSE-based relative efficiency," defined as the ratio of the mean squared error (MSE) of $\hat{\theta}_{\text{lbl}}$ across all zones in a given (country, year) to the MSE of the relevant estimator, and a "CI-based relative efficiency," defined as the squared ratio of the average widths of the respective confidence intervals. Both of these relative efficiencies can be viewed as effective sample size multipliers. By first-order asymptotics, they should be equivalent and independent of the zone size n for sufficiently large n. For the PPI++ estimator, they can be approximated (optimistically, assuming $\hat{f}_j \approx \mu_j$ for all j) by averaging the right-hand side of (3) across all zones j.

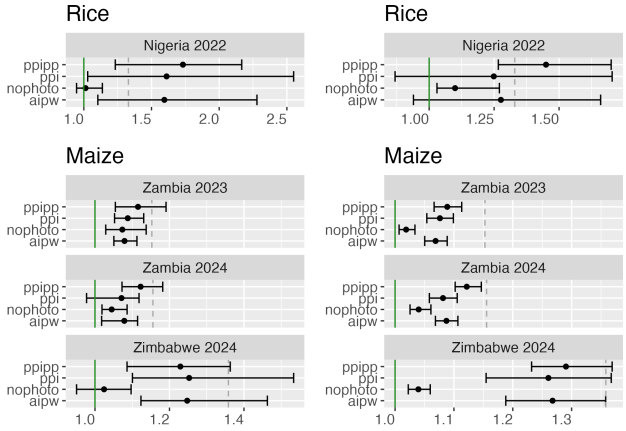

Figure 4: The estimated "MSE-based" (left) and "CI-based" (right) relative efficiencies of the proposed method (`ppipp`) and other alternatives described in the text for each (country, harvest year) pair. The dots are point estimates for each relative efficiency, and computed as averages of squared error ratios (left) and squared CI width ratios (right) across all synthetic bootstrap datasets for all zones. The error bars are 95% BCa bootstrap confidence intervals for the true relative efficiencies, where the uncertainty stems from having only sampled a finite number of zones in each (country, harvest year) pair. The dashed vertical grey lines indicate the theoretical asymptotic relative efficiencies of the PPI++ estimator based on (3) with $N_j/n_j = 4$ and R_j^2 estimated in each zone by the squared Pearson correlations between the computer vision model predictions and ground truth yields.

In Fig. 4 we see that our proposed PPI++ estimator has the lowest average mean squared error (highest MSE-based relative efficiency) across all (country, harvest year) pairs except for (Zimbabwe, 2024), where the AIPW and PPI estimators edge out PPI++ very slightly. The relative efficiencies are

significantly larger than 1, indicating strong statistical evidence for improvement over the baseline. In particular, the MSE-based relative efficiencies imply a 73% increase in the effective sample size from using PPI++ to estimate average rice yields in (Nigeria, 2022), along with 12%, 12%, and 23% increases in the effective sample sizes for estimating average maize yields in (Zambia, 2023), (Zambia, 2024), and (Zimbabwe, 2024), respectively. The markedly superior empirical performance on rice compared to maize is not consistent with the theoretical estimates based on (3), though these estimates all lie within the error bars for the empirical MSE-based relative efficiency estimates for the PPI++ estimator for all four (country, harvest year) pairs. One factor that would favor empirical performance on rice is that our rice yields are noticeably less skewed than our maize yields (Fig. 3). It is clear that adjusting for only latitude and longitude and ignoring the photos (the `nophoto` estimator) performs worse than the proposed PPI++ estimator, demonstrating the value of the photos and computer vision model, although we see that $\hat{\theta}_{\text{nophoto}}$ provides a statistically significant improvement over $\hat{\theta}_{\text{lbl}}$ at the unadjusted 95% confidence level in both Zambia years, suggesting spatial coordinates have some predictive power independent of photos.

The PPI++ estimator also has the narrowest CI's among all methods considered for all (country, harvest year) pairs, supporting the conclusions from the MSE-based relative efficiencies. We see in Table 3 that this narrowness does come at some slight cost to coverage, though the empirical coverage differential between our intervals based on PPI++ and any of the other estimators is no more than 1% in all of the (country, harvest year) pairs. For rice yields in (Nigeria, 2022), the empirical coverage for all estimators exceeds the nominal 95% level, while we observe slight undercoverage for all estimators for the three maize (country, harvest year) pairs, again consistent with the skewness in the maize yields 3. We show in Appendix B.1 that our use of BCa bootstrap confidence intervals improves the empirical coverage over both standard CLT-based intervals and the percentile bootstrap intervals suggested by Zrnic (2025).

Continuing the interpretation of our relative efficiency metrics as sample size multipliers, Fig. 5 suggests that the PPI++ estimator improves the effective sample size — according to both MSE and CI width — for all zone sizes and all (country, harvest year) pairs. Notably, the improvement expected in large samples from Theorem 1 is observed empirically in zones as small as our *a priori* cutoff of 20 fields.

7 Summary and Discussion

We have shown that PPI provides a mathematically principled and empirically performant way to use a computer vision model to leverage easily obtained field photos to supplement limited ground-truth field observations for precisely estimating average yields. The method provably does not introduce any asymptotic bias in the point estimates, nor can it decrease the asymptotic precision relative to taking the simple average of ground truth observations. Key implementation decisions such as the supervised learning algorithm for fitting the control function, the scale at which observations

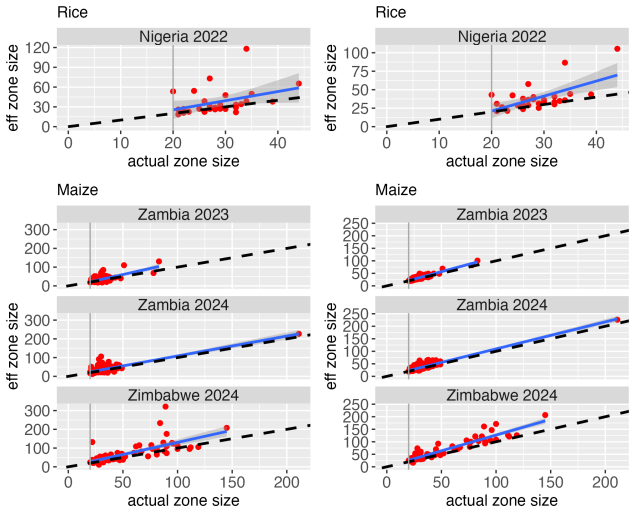

Figure 5: The estimated effective sample size is plotted on each vertical axis against the actual zone size on the horizontal axis for each (country, harvest year) pair. Effective sample size is defined for each zone as the ratio of squared error (left) or squared ratio of CI width (right) from $\hat{\theta}_{\text{lbl}}$ to that from $\hat{\theta}_{\text{PPI++}}$, multiplied by the zone size. The blue lines are ordinary linear regression fits to the scatterplots and the dashed lines pass through the origin with slope 1.

Estimator	(NG, 2022)	(ZM, 2023)	(ZM, 2024)	(ZW, 2024)
$\hat{\theta}_{\text{lbl}}$	0.962	0.939	0.929	0.923
$\hat{\theta}_{\text{AIPW}}$	0.959	0.936	0.923	0.926
$\hat{\theta}_{\text{PPI}}$	0.955	0.940	0.928	0.922
$\hat{\theta}_{\text{PPI++}}$	0.952	0.931	0.927	0.916
$\hat{\theta}_{\text{nophoto}}$	0.952	0.939	0.931	0.921

Table 3: The empirical coverage probabilities of the various nominal 95% confidence intervals across zones in each (country, harvest year).

are pooled to train the control function, and the method used to compute confidence intervals can have practically relevant impacts on performance. Thus, we recommend that prior to deployment, a user of our method validates their choices using simulations that reflect their data-generating process, as we have in Section 6.

Standardized photo-taking guidelines were absent when our photos were taken. Hence, our photos vary significantly in terms of the portion of photo covered by the field and the zoom level of the field (see Fig. 1). A more standardized photo-taking procedure would likely enhance performance and perhaps inform different implementation choices.

The PPI framework extends to a wide range of estimands beyond means, so our methods could be extended to estimate other quantities of interest like specific yield quantiles. When available, one could also integrate additional sources of low-cost data such as satellite imagery.

References

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

Angelopoulos, A. N.; Duchi, J. C.; and Zrnic, T. 2024. PPI++: Efficient Prediction-Powered Inference. arXiv:2311.01453.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45: 5–32.

Darra, N.; Anastasiou, E.; Kriezi, O.; Lazarou, E.; Kalivas, D.; and Fountas, S. 2023. Can Yield Prediction Be Fully Digitilized? A Systematic Review. *Agronomy*, 13(9): 2441.

Efron, B. 1987. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397): 171–185.

Efron, B.; and Hastie, T. 2021. *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press.

Friedman, J. H.; Hastie, T.; and Tibshirani, R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33: 1–22.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Ji, W.; Lei, L.; and Zrnic, T. 2025. Predictions As Surrogates: Revisiting Surrogate Outcomes in the Age of AI. arXiv:2501.09731.

Khaki, S.; Pham, H.; Han, Y.; Kuhl, A.; Kent, W.; and Wang, L. 2021. Deepcorn: A Semi-Supervised Deep Learning Method for High-Throughput Image-Based Corn Kernel Counting and Yield Estimation. *Knowledge-Based Systems*, 218: 106874.

Li, C.; Chimimba, E. G.; Kambombe, O.; Brown, L. A.; Chibarabada, T. P.; Lu, Y.; Anghileri, D.; Ngongondo, C.; Sheffield, J.; and Dash, J. 2022. Maize Yield Estimation in Intercropped Smallholder Fields using Satellite Data in Southern Malawi. *Remote Sensing*, 14(10): 2458.

Lobell, D. B.; Azzari, G.; Burke, M.; Gourlay, S.; Jin, Z.; Kilic, T.; and Murray, S. 2020. Eyes in the Sky, Boots on the Ground: Assessing Satellite–and Ground-Based Approaches to Crop Yield Measurement and Analysis. *American Journal of Agricultural Economics*, 102(1): 202–219.

Robins, J. M.; Rotnitzky, A.; and Zhao, L. P. 1994. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427): 846–866.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 267–288.

Zrnic, T. 2025. A Note on the Prediction-Powered Bootstrap. arXiv:2405.18379.

Zrnic, T.; and Candès, E. J. 2024. Cross-Prediction-Powered Inference. *Proceedings of the National Academy of Sciences*, 121(15): e2322083121.