Addressing Bias from Misclassification in Predicted Data: Theory and Application to Satellite Crop Data*

Arthur R. Wardle†

August 2026

Abstract

Predicted data, generated by machine learning or other algorithms rather than direct measurement, are increasingly popular in empirical research but prone to measurement error. I introduce a flexible, nonparametric multiple imputation approach that uses a small validation sample with known true values to correct for misclassification in predicted categorical data, which would otherwise cause bias in model estimation. The method draws multiple plausible realizations of the unobserved true categories and estimates the model of interest on each completed dataset. I show that the simple average of these estimates is consistent for the true parameter of interest and how to calculate valid confidence intervals for inference. Simulations confirm the procedure effectively removes bias under known data-generating processes. I then apply the procedure to a multinomial crop choice model for Kern County, California. Whereas naive use of satellite-based predicted data results in substantial bias, the multiple imputation method recovers estimates statistically indistinguishable from those based on full ground truth observations.

*I thank Maximilian Auffhammer, Ellen Bruno, Tamma Carleton, Joel Ferguson, Simon Greenhill, Nick Hagerty, James Sallee, Aaron Smith, and participants at the 2024 AAEA, 2025 TWEEDS and WEA, and 2026 IHS Junior Fellowship conferences for helpful feedback. I am grateful to Luna Yue Huang for helpful code. This research is supported by the Foundation for Food & Agriculture Research (FFAR) under Grant ID NIA21-0000000036 and from the Institute for Humane Studies under Grant ID IHS020193. Previous versions of this work circulated under the title "Model Estimation using Categorical Satellite Data with Misclassification.''

†Department of Agricultural Economics & Economics, Montana State University, arthur.wardle@berkeley.edu

Introduction

Many new data sources for economically important variables now consist of *predicted data*: the output of statistical or machine learning models rather than direct measurements. Machine classification enables low-cost and scalable labeling of otherwise prohibitively large raw datasets. For example, machine classification of natural language is increasingly used for measuring features like topic or sentiment in speeches or online posts in political science and economics (de Slegte et al., 2024). A particularly widespread form of predicted data is *satellite data*, which extracts variables from satellite imagery using models trained on a small set of pre-labeled observations. Across agricultural, environmental, and development economics, satellite-based categorical data on forest cover, crop choice, burned areas, agricultural practice adoption, and urban land use are already playing a major role in applied research (Donaldson and Storeygard, 2016; Wuepper et al., 2025). As deep learning methods continue to improve, the availability and use of predicted data are likely to expand further (Dell, 2025).

Although these emerging data sources open new avenues for research, they require careful treatment because of potential measurement error. In particular, predicted data often involve categorical variables—such as land cover or crop type—that are subject to misclassification, meaning than an observation is assigned to the wrong category. Predicted data are typically more prone to misclassification than data obtained through traditional measurement. Such misclassification may be systematically related to other variables in a model, introducing a clear source of bias. However, even when misclassification is completely random, it remains an inherently non-classical form of measurement error that distorts parameter estimates.

Existing econometric methods for addressing measurement error are primarily designed for continuous variables, and those that are not generally require assumptions ill-suited to the predicted data context. These include parametric restrictions on the error process, valid instruments, or multiple proxies, sometimes in combination with validation data. The nature of misclassification in predicted data makes such assumptions difficult to justify. This paper

introduces a method that instead relies solely on validation data and minimal regularity assumptions.

The method I propose is a general purpose tool for end users of predicted data products, designed to remove bias from model estimates without reinventing the classification algorithm or requiring modification of standard statistical software. The approach starts by using a relatively small validation sample to learn the structure of misclassification in the predicted data—for example, that corn is more likely to be misclassified as wheat than apples, or that misclassification rates differ between irrigated and non-irrigated fields. This step is implemented nonparametrically to avoid misspecification bias.

Next, adapting techniques from the literature on *multiple imputation* for missing data, I generate many "corrected" datasets that represent plausible realizations of the unobserved true data, given what I learned about the structure of misclassification in the first step. The model of interest is then estimated separately on each "corrected" dataset, and the resulting estimates are combined using standard multiple imputation rules. The final estimator, the simple average of the estimates from each "corrected" dataset, is consistent for the parameter that would have been obtained using correctly classified data, with valid confidence intervals for inference.

More concretely, suppose the goal is to estimate a parameter β, using an estimator $\hat{\beta}(X, Z)$ that would satisfy $\mathbb{E}(\hat{\beta}(X, Z)) \approx \beta$ if the true covariates X and Z were observed. Instead, Z is unobserved for most units, and only a predicted proxy $\tilde{Z}$ is available. A small validation subsample contains both $\tilde{Z}$ and the true Z. Using the the law of iterated expectations, $\mathbb{E}(\hat{\beta}(X, Z)) = \mathbb{E}\left(\mathbb{E}\left(\hat{\beta}(X, Z)|X, \tilde{Z}\right)\right)$, multiple imputation estimates $\hat{\beta}(X, Z)$ by averaging over the estimates obtained from many simulated versions of Z. Each simulated Z is drawn from an *imputation model*, a model for $P(Z \mid \tilde{Z}, X)$ estimated using the validation data.

I evaluate the performance of the proposed method in a series of simulation exercises in which misclassification is introduced under a known data generating process. I then apply

the approach to estimating a multinomial crop choice model for Kern County, California, a major agricultural region at the southern end of California's Central Valley. Using both a comprehensive ground truth dataset and a satellite-derived predicted dataset for this area, I show that the method substantially mitigates the misclassification-induced biases observed in estimators that naively use the satellite data.

The remainder of the paper proceeds by discussing the sources of misclassification in predicted data, its consequences for estimation, and existing approaches for addressing it. I then introduce the general framework for multiple imputation and present a flexible, nonparametric implementation based on a hierarchical Dirichlet process mixture model. I then report results from simulation studies and apply the method to the Kern County crop choice data.

The Sources of Misclassification in Predicted Data

Predicted data, also referred to as *modeled* or *algorithmic* data, is data whose values are inferred by an algorithm rather than obtained through direct measurement. Continuing with the example of satellite crop data, a machine learning classifier (such as a random forest or decision tree) predicts which crop is grown in each pixel of a raster image or in each field of a shapefile dataset using satellite imagery. Such prediction algorithms vary widely in form and complexity, but they generally offer no guarantees that errors in the predicted data are independent of the true variable, other model variables, or of unobserved factors.

Figure 1 illustrates the provenance of predicted data and how it enters model estimation. Misclassification arises through several distinct channels.

First, errors in the raw input data propagate to predictions. For satellite data, these errors can come from imperfect sensors, distortions caused by cloud cover, smoke, or topography, errors in pre-processing that make imperfect adjustments for these issues, and errors introduced through interpolation to make up for incomplete satellite coverage.[1]

[1] Jain (2020) and Alix-Garcia and Millimet (2023) offer more thorough descriptions of how errors in the raw satellite data can impact downstream economics applications.

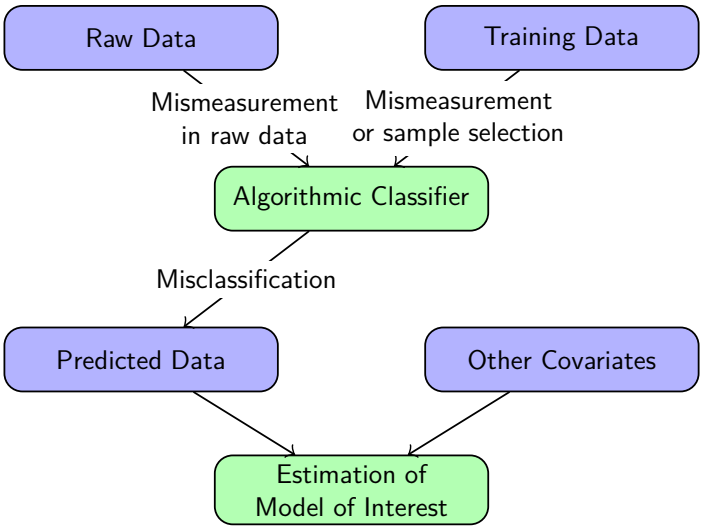


Figure 1: Origin and End Use of Predicted Data

Second, mismeasurement or sample selection in the training data can bias predictions. For example, the Cropland Data Layer, a satellite crop dataset used in the empirical application of this paper, relies on administrative data from federal farm support programs, oversampling crops that receive government support. Training on non-random samples introduces systematic bias into the predicted data.

Third, the algorithm itself produces prediction errors. Machine learning models are often "black boxes," capable of achieving high degrees of predictive accuracy but without guarantees that remaining misclassification is independent of other variables.[2] Biases are often baked in during model development to seek greater predictive accuracy in the face of a bias-variance tradeoff (Dell, 2025). Only some predicted datasets offer any uncertaintly quantification, and those that do typically offer it in the form of a confusion matrix that summarizes misclassification rates across classes.[3] Short of perfect prediction accuracy, some degree

[2]Sanford et al. (2025) offer a technique for applying "adversarial debiasing" at the classifier training stage to ensure that errors in modeled outputs are uncorrelated with all other model covariates. This technique works when the satellite data are being used as a dependent variable and the classifier uses stochastic gradient descent for optimization. This requires the classification model to be tailored to a specific research design.

[3]For a classifier with J classes, a confusion matrix is a $J \times J$ matrix giving counts among all observations in some ground truth sample (sometimes the same one used for training, sometimes a holdout sample) with the true class along the x-axis and the algorithm-assigned class on the y-axis. A perfect classifier would produce a diagonal matrix.

of random error is inevitable; at best, this error is completely independent other model covariates.

Even widely used satellite crop datasets with reported overall accuracies in the neighborhood of 90 percent still contain enough misclassification to matter for econometric analysis. Despite this, predicted datasets are increasingly employed in applied economics and other fields without systematic assessment of misclassification or strategies to mitigate its impact on estimation.

Estimation Problems Under Misclassification

In linear regression with classical measurement error, which is mean zero and independent of the true variable and other covariates, the impact of measurement error depends on which variable is mismeasured. When the dependent variable is mismeasured, coefficient estimates remain unbiased, though variance increases. If an independent variable is mismeasured, its coefficient is attenuated and inconsistent, and coefficients on correlated variables will also be biased. These results, which are widely known and frequently form the basis for economists' intuitions about the impact of measurement error, do not apply to non-classical measurement error, nor do they apply to non-linear models.

Misclassification refers specifically to measurement error in categorical variables. This form of error is inherently non-classical. Unlike classical measurement error, misclassification in categorical data necessarily involves correlation between the measurement error and the true value, even if the error is completely random. This property means standard intuitions from classical measurement error do not apply to misclassification.

To see why, consider a binary variable $x \in \{0, 1\}$ and a randomly misclassified version $\tilde{x}$. The misclassification error $\eta = \tilde{x} - x$ can only take values in $\{-1, 0, 1\}$; in particular, when $x = 1$, $\eta \in \{-1, 0\}$ and when $x = 0$, $\eta \in \{0, 1\}$. This induces a negative correlation between x and η, no matter how misclassification arises. More generally, any mismeasured discrete variable

exhibits this non-classical property.

Even small degress of misclassification can have large consequences. An early contribution by Aigner (1973) demonstrated that in simple linear regression with a single binary independent variable, completely random misclassification induces attenuation proportional to the misclassification rate. For non-linear models or dependent misclassification, it is difficult or impossible to derive formulas for how exactly misclassification will affect estimates. Simulation evidence confirms that even modest misclassification can produce extreme errors in parameter estimates (Millimet, 2010). Empirical evidence reinforces these findings: correcting for misclassification in employment status reversed a key result on fertility and female labor supply, despite an overall misclassification rate below 2 percent (Keane and Sauer, 2009).

In short, misclassification can substantially bias parameter estimates, particularly in non-linear models. This is true even when misclassification is completely random; potential dependence between misclassification in predicted data and other model variables only accentuates the problem. This motivates the need for a flexible correction method that can produce consistent estimates from predicted data subject to misclassification.

Existing Approaches

Identification under misclassification must rely on either additional structural assumptions or auxiliary data such as instruments, repeated measurements, or validation data.

One structural route to identification imposes restrictions on how misclassification arises. For instance, Hausman et al. (1998) show that for a categorical outcome model estimated via maximum likelihood, assuming that misclassification probabilities are independent of all other covariates ($\mathbb{P}(y = i|\tilde{y} = j, x) = \phi_{ij} \ \forall \ i, j$) enables consistent estimation through a simple adjustment to the likelihood function. However, this assumption is restrictive: even small violations can yield parameter estimates that differ significantly from those obtained using correctly classified data (Kreider, 2010).

Alix-Garcia and Millimet (2023) extend the Hausman et al. (1998) approach by allowing misclassification probabilities to depend on covariates, $\mathbb{P}(y = i|\tilde{y} = j, x, z) = G_{ij}(z'\beta)$, but this comes at the cost of requiring specific link functions G_{ij}. While this relaxes independence assumptions and is estimable without validation data, the true functional form of misclassification probabilities is generally unknown, even when researchers can identify which variables ought to be included in these functions.[4] It is not the case that *any* choice of function will alleviate bias either; Meyer and Mittag (2017) shows that misspecified misclassification functions can aggravate bias relative to a naive, uncorrected model. Because the algorithms underlying predicted data are often "black boxes" with unknown error-generating processes, strong functional form assumptions are generally implausible.[5]

An alternate route uses auxiliary data, such as instruments, repeated proxies, or validation samples. Methods using instruments often assume misclassification is independent of covariates (e.g. Frazis and Loewenstein, 2003), but Mahajan (2006) and Hu (2008) allow (in the binary and general discrete case respectively) for point identification even when misclassification probabilities are functions of observed data in the presence of an "instrument-like variable" that is correlated with the true underlying variable of interest but not the misclassification probability.[6] Predicted data will generally violate these requirements, since predicted data are a post-outcome noisy measurement of the true class, and therefore anything affecting the true class can also impact misclassification.[7] This literature continues to be active, offering

[4] See Mikhaeil et al. (forthcoming) for some specific cases where it is feasible to make reasonable assumptions about the functional forms of misclassification probabilities and how to estimate them in hierarchical Bayesian models.

[5] In cases where *all* variables in a model are categorical, Pepe and Fleming (1991) show how to estimate a nonparametric misclassification function to correct a maximum likelihood estimator making use of the empirical distribution of a validation sample, but this method quickly succumbs to the curse of dimensionality.

[6] Though DiTraglia and García-Jimeno (2019) show the Mahajan (2006) result to work only with exogeneity in the true value of the misclassified variable, with multiple "instrument-like" variables, or with multiple misclassified proxies.

[7] Rambachan et al. (2025) make a similar point to argue that surrogacy methods are inappropriate for satellite mismeasurement problems. However, these techniques can still be useful for statistics where correlation with other variables is not a concern. Torchiana et al. (2025) offer a specific implementation of the method from Hu (2008) to back out corrected transition probabilites using a Hidden Markov Model applied to a panel of satellite land classification data.

new approaches for point identification based on different moment restrictions, combinations of instruments, or combinations of misclassified proxies (Card, 1996; Lewbel, 2007; Battistin et al., 2014; DiTraglia and García-Jimeno, 2019; Nguimkeu et al., 2019; Yanagi, 2019; Calvi et al., 2022; Kasahara and Shimotsu, 2022). These approaches require either instruments, multiple misclassified proxies, or both for every observation.

Other work avoids the need for additional data by imposing restrictions on the relationship between misclassification and model error (Chen et al., 2008a,b, 2009). Chen et al. (2009) achieve nonparametric identification of a regression model with a misclassified covariate, relying on monotonicity of the regression function and conditional independence of both the true and misclassified covariate with the model error. Estimation, however, requires solving a complex system of equations, and the method does not apply to misclassified outcomes.

The method I propose relies only on a small, appropriately sampled validation dataset. It applies whether the misclassified variable is the outcome or an independent variable in the model of interest, and works for any asymptotically linear estimator, a wide class of estimators including OLS, MLE, GMM, and more. Moreover, the method does not require modification of an objective function, or anything else that would frustrate a practitioner's ability to use standard statistical estimation packages.

Proctor et al. (2023) also use multiple imputation, showing that it is highly effective for handling measurement error in continuous, remotely sensed data. They use a Bayesian linear regression for the imputation model, which does not accommodate imputation of categorical variables. My approach extends this idea to categorical data, allowing for nonparametric modeling of misclassification without relying on linearity in the imputation model.[8]

While others in the literature have sometimes considered the possibility of using multiple imputation to resolve misclassification in predicted data, they have generally written off the

[8]I will focus exclusively on the categorical case in this paper given its novelty, but the nonparametric imputation model does additionally allow for multiple imputation of continuous variables without the imposed linearity of the imputation model adopted by Proctor et al. (2023). In practice, imputation of continuous variables tends not to be overly sensitive to imputation model structure.

possibility due to potential bias from misspecification in the imputation model (e.g. Fong and Tyler, 2021, Kluger et al. (2025)). Most implementations of multiple imputation in other contexts have historically relied on simple, parametric imputation models, since they were designed for problems where minor misspecification would have only minor consequences for downstream analysis. This paper tackles this problem directly by using a Bayesian nonparametric imputation model.

Two other emerging techniques, prediction-powered inference (Angelopoulos et al., 2023) and predict-then-debias estimators (Chen and Chen, 2000, Egami et al. (2024), Kluger et al. (2025), Lu et al. (2025)), also leverage validation data to correct for misclassification.[9] The former requires adjustment to a loss function, complicating estimation and limiting its applicability to M-estimation.[10] The latter is similar, but adjusts parameter estimates directly rather than loss functions, but outside of generalized linear models, requires either custom calculations or bootstrapping to compute confidence intervals. In contrast, multiple imputation provides a framework that is readily implementable using standard statistical software.

Multiple Imputation

Problem Setup and Assumptions

I aim to estimate and conduct inference on a (potentially multidimensional) parameter β describing a probability distribution over random vectors (Z, X, Y) where Z and X are categorical and Y is continuous. This general setup allows Z to be either an outcome or independent variable. I observe a sample $\mathcal{S}$ of size N containing $(\tilde{Z}, X, Y)$, where $\tilde{Z}$ is a potentially misclassified predicted proxy for Z, e.g. a satellite-based prediction. For a subset

[9] Egami et al. (2024) refer to their method as design-based supervised learning, but the method's formulation is equivalent to the predict-then-debias approach.

[10] This can also result in a non-convex loss function, even when the loss function without the rectifier is convex (Ji et al., 2025).

$\mathcal{S}^* \subset \mathcal{S}$ of size N^*, I also observe the true value of Z; this subset forms the validation data. I denote by Z^+ the observed true value of Z for $i \in \mathcal{S}^*$ and missing otherwise, and by R an indicator equal to one if $i \in \mathcal{S}^*$. Let Z^- denote the unobserved true values of Z for $i \notin \mathcal{S}^*$. The observed dataset is therefore $(Z^+, \tilde{Z}, X, Y, R)$.

Analyzing misclassified data with accompanying validation data is analogous to analyzing data with missing values. Despite the analogy, two key differences distinguish the predicted data misclassification problem from a typical missing data problem. First, the extent of "missingness" is much higher: validation samples may cover only a small fraction of observations ($N^* << N$), whereas methods for missing data are generally designed assuming most values are observed. Second, I observe the proxy $\tilde{Z}$, which provides valuable information about the missing Z and simplifies imputation in practice. My strategy adapts multiple imputation methods from the missing data literature to exploit information in the misclassified proxy $\tilde{Z}$. Before continuing, I introduce the basic assumptions necessary to proceed with a multiple imputation based analysis. The first two assumptions concern sampling: the first deals with how the overall sample $\mathcal{S}$ is drawn from the population $\mathcal{P}$, while the second concerns how the validation sample $\mathcal{S}^*$ is drawn from $\mathcal{S}$.

Assumption 1 (Ignorable Sampling) *Consider the full population $\mathcal{P}$ of (Z, X, Y) of which our data are only a subset $\mathcal{S} \subset \mathcal{P}$. Let I be an indicator variable for whether an observation in $\mathcal{P}$ is also in $\mathcal{S}$. My analysis assumes that the sampling procedure generating $\mathcal{S}$ is ignorable (Rubin, 1987, p. 50), meaning*

$$P(I \mid Z, \tilde{Z}, X, Y, R) = P(I \mid \tilde{Z}, X, Y, R) \; \forall \; Z$$

This is weaker than simple random sampling and allows for stratifying on observable information. Since sampling is assumed ignorable, I will be suppressed throughout the rest of my analysis.

Assumption 2 (Ignorable Validation Sampling) *My analysis assumes that the sampling procedure generating the validation data $\mathcal{S}^* \subset \mathcal{S}$ is ignorable, which, conditional on the previous assumption, is equivalent to (Rubin, 1987, p. 51):*

$$P(R \mid Z, \tilde{Z}, X, Y) = P(R \mid \tilde{Z}, X, Y) \; \forall \; Z$$

This assumption is again weaker than simple random sampling but requires that the probability of inclusion in the validation data does not depend on the unobserved true value of Z.[11] In particular, this *does* allow for stratified validation sampling based on observed values of $\tilde{Z}$, a likely feature of validation datasets created for the purpose of assessing the quality of a predicted data product.

Unfortunately, the ground truth data used to train the machine classifier that produced $\tilde{Z}$ violates this assumption. Because in-sample prediction accuracy is almost always higher than out-of-sample accuracy, observations where $Z = \tilde{Z}$ are more likely to be included in the training set, violating ignorability. Using such data in imputation would understate variance, since the training data over represents cases the classifier predicts well. In contrast, holdout samples used to evaluate classifer accuracy after training avoid this bias and offer a promising source of validation data.

Following the introduction of multiple imputation and my imputation model, I will discuss how violations of these assumptions affect estimation. I particularly focus on departures from ignorability in validation data sampling, given the potential availability of non-ignorable validation samples of convenience.

The important consequence of Assumptions 1 and 2 is that $P(Z^- \mid Z^+, \tilde{Z}, X, Y, R, I) = P(Z^- \mid Z^+, \tilde{Z}, X, Y)$, i.e. when the validation sample is ignorable, there will not be systematic

[11] An alternate statement of this assumption is $R \perp\!\!\!\perp Z \mid \tilde{Z}, X, Y$. Ignorable sampling implies the main assumption in Chen et al. (2005), who use a very similar approach for using a small validation dataset in concert with a larger, mismeasured dataset to estimate a GMM model. Their focus is on continuous data and their method is exclusive to GMM estimators.

differences in the conditional probability of Z between observations inside and outside of the validation sample.[12] While consistent multiple imputation is still possible under certain violations of this assumption, implementation requires an additional auxiliary model relating observed variables to the validation sampling mechanism (Rubin, 1987, p. 212), which is beyond the scope of this paper.

Assumption 3 (Positivity in Validation Sampling) *The conditional probability of any given observation in the dataset $\mathcal{S}$ also appearing in the validation dataset $\mathcal{S}^*$ is positive, i.e.*

$$P(R \mid Z, \tilde{Z}, X, Y) > 0 \; \forall \; Z, \tilde{Z}, X, Y$$

This is a basic identifiability assumption; if certain values of $(Z, \tilde{Z}, X, Y)$ are possible but cannot appear in the validation data, the imputation model would require out-of-sample extrapolation, even in the limit as $N^* \to \infty$.

Assumption 4 (Asymptotic Linearity of the Estimator) *The estimator of interest $\hat{\beta}$ (what would be directly estimated if complete and correctly classified data were available), is asymptotically linear. Equivalently, the estimator $\hat{\beta}$ can be expanded*

$$\hat{\beta} = \beta + \frac{1}{N} \sum_{i=1}^{N} T(z_i, x_i, y_i) + o_p\left(\frac{1}{\sqrt{N}}\right)$$

Additionally, $Var(T(z_i, x_i, y_i))$ exists and is finite.

The parameter and its estimate β and $\hat{\beta}$ can be either scalars or vectors; if each element in a parameter vector is asymptotically linear, then so is the entire vector. Almost all estimators used by applied econometricians are asymptotically linear; Newey and McFadden (1994) demonstrate this property for maximum likelihood, generalized method of moments, and ordinary/non-linear least squares estimators. Even more generally, the class of asymptotically linear estimators includes standard M-estimators, L-estimators, R-estimators (Fernholz,

[12] This is analagous to the "missing at random" assumption in the missing data literature.

1983), and certain regularized M-estimators (Werner, 2021). Combined with the existence of $\text{Var}(T(z_i, x_i, y_i))$, asymptotic linearity implies asymptotic normality of the test statistic. Appendix 3 shows that asymptotic linearity of the estimator of interest is sufficient for the multiple imputation estimator to be asymptotically unbiased.

Multiple Imputation Basics

Although multiple imputation has strong frequentist justification, it is most easily understood in a Bayesian framework. I therefore describe the approach in terms of estimating a posterior for β.

With fully observed data, the posterior distribution of β with complete data is $\text{P}(\beta \mid Z, X, Y)$, which by assumption is equivalent to $\text{P}(\beta \mid Z, \tilde{Z}, X, Y)$ since the misclassified proxy $\tilde{Z}$ adds no information once Z is known. Instead, Z is only partially observed, so the actual posterior distribution is $\text{P}(\beta \mid Z^+, \tilde{Z}, X, Y)$. These two posteriors are related by

$$\text{P}(\beta \mid Z^+, \tilde{Z}, X, Y) = \int \text{P}(\beta \mid Z, X, Y) \times \text{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)\, dZ^-$$

Following Rubin (1987, p. 82), this formulation implies that draws from $P(Z^- \mid Z^+, \tilde{Z}, X, Y)$ can be used to simulate the actual posterior of β, since Z^+ and Z^- together determine Z. The following section will detail the precise imputation model I use to estimate this conditional distribution. Generating imputations of Z^- from this conditional distribution fills in the unobserved cases Z^-, after which the complete data estimator $\hat{\beta}$ can be applied without modification. Repeating this process M times and recording the estimate $\hat{\beta^m}$ and variance estimate U^m each time yields $\{\hat{\beta^m}, U^m\}_{m=1}^M$.

Again following Rubin (1987, p. 84), let $\hat{\beta}$ be the mean of the posterior distribution of β with complete data and U its associated variance-covariance matrix. Then by the law of iterated

expectations,

$$\begin{aligned}\mathbb{E}(\beta \mid Z^+, \tilde{Z}, X, Y) &= \mathbb{E}(\mathbb{E}(\beta \mid Z, X, Y) \mid Z^+, \tilde{Z}, X, Y) \\ &= \mathbb{E}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)\end{aligned}$$

and by the law of total conditional variance,

$$\begin{aligned}\mathrm{Var}(\beta \mid Z^+, \tilde{Z}, X, Y) &= \mathbb{E}(\mathrm{Var}(\beta \mid Z, X, Y) \mid Z^+, \tilde{Z}, X, Y) \\ &\quad + \mathrm{Var}(\mathbb{E}(\beta \mid Z, X, Y) \mid Z^+, \tilde{Z}, X, Y) \\ &= \mathbb{E}(U \mid Z^+, \tilde{Z}, X, Y) + \mathrm{Var}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)\end{aligned}$$

For large M, this suggests the following estimators (Rubin, 1987, p. 85):

$$\hat{\beta}_\infty \equiv \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} \hat{\beta^m} = \mathbb{E}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)$$

$$U_\infty \equiv \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} U^m = \mathbb{E}(U \mid Z^+, \tilde{Z}, X, Y)$$

$$B_\infty \equiv \lim_{M\to\infty} \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta^m} - \hat{\beta}_\infty\right)\left(\hat{\beta^m} - \hat{\beta}_\infty\right)' = \mathrm{Var}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)$$

These satisfy $\mathbb{E}(\beta \mid Z^+, \tilde{Z}, X, Y) = \hat{\beta}_\infty$ and $\mathrm{Var}(\beta \mid Z^+, \tilde{Z}, X, Y) = U_\infty + B_\infty$, implying $(\hat{\beta}_\infty - \beta \mid Z^+, \tilde{Z}, X, Y) \sim \mathcal{N}(0, U_\infty + B_\infty)$.

In practice, M is finite and may even be quite small depending on the computational burden of parameter estimation. Rubin (1987) assesses the small-M case and suggests using $T_M = U_M + (1 + \frac{1}{M})B_M$ with U_M and B_M to estimate the total variance.

For inference, multiple imputation estimators use t-distributions, but require calculating an adjusted degrees of freedom to account for the imputation. For general use, where v_{com} is what the degrees of freedom would be with complete truth data, Barnard and Rubin (1999)

and Reiter and Raghunathan (2007) recommend the following formula for its satisfactory frequentist coverage:

$$v_m = \left[\left((M-1) \left(1 + \frac{U_M}{\left(1 + \frac{1}{M}\right) B_M} \right)^2 \right)^{-1} + \left(v_{com} \frac{U_M}{T_M} \frac{v_{com}+1}{v_{com}+3} \right)^{-1} \right]^{-1}$$

With this foundation in place, the next task is to specify an imputation model for $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$. Because the dependence of misclassification on Z, X, and Y is uncertain and imputation model misspecification could cause bias, I turn next to a flexible nonparametric approach for generating draws of the unobserved Z values.

A Bayesian Nonparametric Imputation Model

The key step in implementing multiple imputation is obtaining a sensible estimate of the conditional distribution $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$, which governs the unobserved truth values for the misclassified variable. Estimating this probability mass function allows me to draw multiple sets of imputed Z values and proceed with inference as described in the previous section.

Since Z is categorical, this conditional distribution is itself categorical, with probabilities for each category dependent on $Z^+, \tilde{Z}, X$, and Y. However, imputing from a single best-fit distribution would treat the estimated probabilities as known with certainy, ignoring uncertainty about the distribution itself.[13] Instead, I seek a *posterior* distribution over $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$, informed by a minimally informative prior and the validation data. This posterior allows me to first draw a probability vector for each unvalidated observation and then draw one realization of z_i from that vector for each imputation m. As Rubin (1987) emphasizes, this approach constitutes a *proper* imputation model—one that propagates uncertainty in the imputation process and yields valid confidence intervals.

[13] Sturcutral approaches instead incorporate uncertainty over misclassification probabilities by embedding the parameters governing misclassification into the likelihood function and estimating them directly.

Many choices of imputation model are possible. When misclassification is known to take a specific form or to be independent of some covariates, bespoke models can deliver tighter inference. Here, my goal is to construct a general purpose imputation model that accomodates mixed continuous and categorical data, avoids parametric misspecification, and is parsimonious enough to be feasible to estimate.[14] A promising approach satisfying these criteria is the Hierarchically Coupled Mixture Model with Local Dependence (HCMM-LD) first introduced by Murray and Reiter (2016). This Bayesian nonparametric model flexibility represents the joint distribution of mixed continuous and discrete variables, growing more complex as data demand while remaning parsimonious.

Model Structure

Building on the success of mixture models methods for purely categorical or continuous datasets (Murray, 2018), the HCMM-LD couples two lower-level mixture models, one for categorical variables and one for continous variables conditional on the categorical variables, through a third, top-level mixture. This hierarchical coupling maintains flexibility while preventing overparameterization by allowing conditional independence between the lower-level components while preserving dependence through the higher-level one. The model is estimated using the fully observed validation data.[15]

Expositionally, it helps to split the data into categorical variables and continuous variables, lumping what I have been calling Z and $\tilde{Z}$ in with what I have been calling X. Let $X_i = (Z_i, \tilde{Z}_i, X_{i1}, \ldots, X_{ip-2})'$ be a vector of p categorical variables, with the first and second

[14] Even under these criteria, there are promising alternatives to the model I use. As one example, sequential regression trees (Burgette and Reiter, 2010) have shown promise in empirical applications (Akande et al., 2017).

[15] This differs from Murray and Reiter (2016), who estimate the HCMM-LD using all data and impute missing values within the Gibbs Sampler. This is due to their focus on situations in which many variables have a small degree of missing information, in which case using all data can be valuable, since an observation with missing data for one variable still contains information about the covariance of other variables that may be missing in other observations. This is not a concern for my intended application. The non-validated data do not contribute much information about the relationship of Z to other variables, and excluding them saves significant computational resources in estimating via the Gibbs Sampler.

elements equal to the true and potentially misclassified versions of the variable we are imputing for, respectively. Let $Y_i = (Y_{i1}, \ldots, Y_{iq})'$ be a vector of continuous variables in $\mathbb{R}^q$.

Each individual is assigned to a categorical component $s_i \in \{1, \ldots, k^{(\mathcal{X})}\}$ and a continuous component $r_i \in \{1, \ldots, k^{(\mathcal{Y})}\}$. Conditional on memberships in these components, densities for X_i and Y_i are given by

$$\mathbb{P}(X_i = x_i | S_i = s) = \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)}$$

$$(Y_i = y_i | R_i = r, X_i = x_i) \sim \mathcal{N}(D(x_i)B_r, \Sigma_r)$$

A higher-level component $z_i \in \{1, \ldots, k^{(\mathcal{Z})}\}$[16] links the lower-level components,

$$\mathbb{P}(S_i = s, R_i = r | Z_i = z) = \phi_{zs}^{(\mathcal{X})} \phi_{zr}^{(\mathcal{Y})}$$

with membership in the high-level component drawn from a categorical distribution with probabilities

$$\mathbb{P}(Z_i = z) = \lambda_z$$

Though categorical variables are mutually independent within a component, dependence among categorical variables arises through shared component membership; dependence between categorical and continuous variables is captured directly through the regression structure $D(x)B_r$ (with $D(x)$ being the design matrix of x and B_r being parameter estimates), with additional higher-order interactions captured via the high-level component.[17] Dependence between continuous variables is also modeled directly through Σ_r. As B_r and Σ_r are specific to each continuous component, these parameters capture *local* rather than global dependence.

This structure enables a great deal of flexibility from a parsimonious setup. The marginal

[16] This Z is not to be confused with my earlier notation for the true categorical variable, which is now a component of X. I use Z to refer to the latent top-level mixture component to maintain notation consistency with Murray and Reiter (2016).

[17] $D(x)B_r$ simply alleviates the proliferation of components.

distribution of X takes a latent class form,

$$\mathbb{P}(X_i = x_i) = \sum_{s=1}^{k^{(\mathcal{X})}} \left(\sum_{z=1}^{k^{(\mathcal{Z})}} \lambda_z \phi_{zs}^{(\mathcal{X})} \right) \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)}$$

which reduces to the model of Si and Reiter (2013) or (with an unrestricted number of components) Dunson and Xing (2009), which is consistent for *any* multivariate categorical distribution, a consequence of the fact that any such distribution can be expressed as a finite mixture of product multinomial distributions.[18] The marginal distribution of Y is a mixture of multivariate normals, which is itself consistent for any twice continuously differentiable smooth density (Ghosal and van der Vaart, 2007) or even for non-smooth densities satisfying mild regularity conditions (Wu and Ghosal, 2010).

Marginalizing over all the components yields the joint density:

$$f(X_i, Y_i) = \sum_{z=1}^{k^{(\mathcal{Z})}} \lambda_z \left(\sum_{r=1}^{k^{(\mathcal{Y})}} \phi_{zr}^{(\mathcal{Y})} \mathcal{N}(Y_i; D(x_i)B_r, \Sigma_r) \right) \left(\sum_{s=1}^{k^{(\mathcal{X})}} \phi_{zs}^{(\mathcal{X})} \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)} \right)$$

This structure straightforwardly enables drawing conditional probabilities for any categorical variable:

$$\mathbb{P}(X_{ij} = c_j | X_{i/j} = x_{i/j}, Y_i) \propto \sum_{z=1}^{k^{(\mathcal{Z})}} \lambda_z \left(\sum_{r=1}^{k^{(\mathcal{Y})}} (\phi_{zr}^{(\mathcal{Y})} \mathcal{N}(Y_i; D(\tilde{x}_i)B_r, \Sigma_r)) \right) \left(\sum_{s=1}^{k^{(\mathcal{X})}} \phi_{zs}^{(\mathcal{X})} \prod_{l=1}^{p} \psi_{s\tilde{x}_{il}}^{l} \right)$$

where $X_{i/j}$ is the X_i-vector with the jth element removed and $\tilde{x}_i$ is the vector with entries equal to x_i except that the jth element is set to c_j. Appendix 2 proves that both the joint and conditional distributions estimated in this model converge in L_1 to their true counterparts for any marginal distribution of the categorical variables and subject to only mild regularity conditions on the continuous variables.

[18] Of course, this applies only to multivariate categorical distributions with finite variables and finite categories. A "product multinomial" distribution is a joint distribution of independent categorical variables. See Dunson and Xing (2009) for proofs.

Estimation via Gibbs Sampling

The imputation model is estimated using a Gibbs Sampler. As an MCMC method, the Gibbs Sampler takes the posterior of every estimation iteration as the prior for the following iteration. After convergence, each imputation can come from a single iteration of the Gibbs Sampler, with successive imputations calculated from distinct iterations.[19] This meets the criteria for a proper imputation model, given that uncertainty in the categorical probabilities is captured by using separate draws for each imputation. Full algorithmic details and priors are described in detail in Appendix 1.

Given the central importance of the hierarchical component structure, it is worth paying some attention to the priors for component membership and how they are updated. The priors for the λs and ϕs that determine component membership are distributions defined using the stick-breaking representation of the Dirichlet process (Sethuraman, 1994):

$$\phi_{zs}^{(\mathcal{X})} = \xi_{zs}^{(\mathcal{X})} \prod_{l<s}(1-\xi_{zl}^{(\mathcal{X})}), \quad \phi_{zr}^{(\mathcal{Y})} = \xi_{zr}^{(\mathcal{Y})} \prod_{l<r}(1-\xi_{zl}^{(\mathcal{Y})}), \quad \lambda_z = \xi_z^{(\mathcal{Z})} \prod_{l<z}(1-\xi_l^{(\mathcal{Z})})$$

where

$$\xi_{zs}^{(\mathcal{X})} \sim \text{Beta}(1, \beta^{(\mathcal{X})}), \quad \xi_{zr}^{(\mathcal{Y})} \sim \text{Beta}(1, \beta^{(\mathcal{Y})}), \quad \xi_z^{(\mathcal{Z})} \sim \text{Beta}(1, \alpha)$$

except for $\xi_{zk^{(\mathcal{X})}}^{(\mathcal{X})}$, $\xi_{zk^{(\mathcal{Y})}}^{(\mathcal{Y})}$, and $\xi_{k^{(\mathcal{Z})}}^{(\mathcal{Z})}$, which are set equal to one, truncating the stick-breaking process. Truncation substantially eases computation with negligible loss in flexibility (Ishwaran and James, 2001), but requires checking that, at convergence, some components remain empty or near empty. If all components (in any dimension) are occupied by more than a trivial number of observations, that is evidence of excessive truncation and estimation should begin again with $k^{(\cdot)}$ set higher in the relevant dimension.

This construction permits data-driven complexity: the concentration parameters $\beta^{(\mathcal{X})}$, $\beta^{(\mathcal{Y})}$,

[19] In simulation and applications, I allow multiple iterations between imputations to account for the dependence between neighboring draws in the Markov chain, though this should not matter if the model has appropriately converged.

and α evolve with the number of occupied components. In each iteration, each observation is assigned to one of each of the three components, with components that accurately fit that observation receiving higher probability weight. When none of the existing components fit well, that leaves greater probability mass in empty components, which can lead that observation to initializing a new component to be considered by all observations in future runs. Abstracting from the model's probabilistic nature, the estimation procedure keeps the model's complexity static when existing observations are well described by some combination of existing components, and automatically increases the complexity otherwise.

Computational Feasibility

The proposed implementation of the HCMM-LD is computationally feasible even for extremely large predicted datasets. The computationally intensive step, which is estimating the imputation model, occurs only on the comparatively small validation dataset. The complexity of this computation scales roughly linearly with sample size, though complexity can rise quickly with the number of continuous variables due to the covariance matrices (Murray and Reiter, 2016). Crucially, estimation of the imputation model occurs only once and only until convergence. Once fitted, the model can generate imputations for the large predicted dataset with minimal additional computation, and this step can be parallelized to handle truly massive datasets efficiently.

Incomplete Covariates in the Validation Data

As discussed above, implementing multiple imputation requires that all covariates in the analysis model (that is, all covariates you would use to estimate $\hat{\beta}$ with complete data) be observed within the validation data, so their joint dependence with the true distribution of Z can be modeled through $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$. In practice, however, a validation

sample $\mathcal{S}^*$ containing Z^+ and $\tilde{Z}$ may lack some or all of the other model covariates.[20] In such settings, multiple imputation estimation remains feasible if the missing covariates are conditionally independent of Z given the observed information. Let X_a and Y_a be the observed covariates and X_b and Y_b the unobserved covariates in the validation sample. If $Z^- \perp\!\!\!\perp X_b, Y_b \mid Z^+, \tilde{Z}, X_a, Y_a$, then $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y) = \mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X_a, Y_a)$, which can be estimated using the imputation model.

This issue is particularly likely to arise in randomized program evaluations. Researchers may have comprehensive predicted data and a sample of validation data drawn from a disjoint or minimally overlapping portion of the same population over which treatment assignment is randomized. When Z is an independent variable in the model of interest $Z^- \perp\!\!\!\perp X_b \mid Z^+, \tilde{Z}, X_a, Y$ is clearly a safe assumption when X_b is randomized treatment. However, if Z is the outcome, this assumption no longer holds: when treatment potentially affects Z, it is generally implausible to assume that misclassification error in the predicted data is independent of treatment, as the predictions themselves are post-treatment quantities. This concern does not arise if treatment data are observed in the validation sample, since dependence can then be explicitly modeled. When this is not available, Rambachan et al. (2025) offer an approach for estimation tailored to this specific setting.

Non-Ignorable Validation Data

A key assumption underpinning the multiple imputation procedure is that the validation sampling mechanism is ignorable, that is, $R \perp\!\!\!\perp Z \mid \tilde{Z}, X, Y$, where R is an indicator for an observation having validation data available. Ignorability ensures that the imputation model estimated on the validation sample can be validly extrapolated to the full dataset. When this assumption fails, the validation data become a selected sample, and the imputation model inherits that selection bias.

[20] In this case, $\mathcal{S}^* \not\subset \mathcal{S}$; rather S and S^* are ignorable samples from the same population $\mathcal{P}$. This is an alternative scenario to that assumed throughout the rest of the paper, in which $\mathcal{S}^* \subset \mathcal{S}$.

In practice, however, validation data that does not arise from formal randomization are often the only validation data available. A key question is how sensitive multiple imputation is to a departure from this assumption, and whether the formally inconsistent estimates that arise from applying multiple imputation with a non-ignorable validation sample should be expected to mitigate bias relative to naively using predicted data.

The impact of non-ignorability depends on the relationship between the selection mechanism and the true underlying variable and the sensitivity of the model being estimated. Suppose W is an unobserved variable omitted from the imputation model and that $P(R) = f(W)$. If W is strongly correlated with Z (conditional on $\tilde{Z}$, X, and Y), bias can be substantial. Under weaker correlations, however, multiple imputation often remains beneficial: it reduces bias relative to naive estimation even when ignorability is moderately violated. Simulation and empirical results shown in the following sections will illustrate this robustness property, mirroring similar findings by Proctor et al. (2023) and others in the multiple imputation literature.

When the source of non-ignorability is known, it can often be addressed directly by including relevant covariates in the imputation model so long as the separate positivity condition still holds. For example, it may be significantly easier to collect validation data for a satellite dataset by sampling locations adjacent to accessible roads. Because road proximity might correlate with whatever land characteristic is measured in the satellite data even after conditioning on other covariates, this constitutes a non-ignorable sampling design. However, including road proximity as an auxiliary variable in the imputation model would restore ignorability, so long as all levels of road proximity have nonzero sampling probability. This does not require any adjustment to the ultimate model of interest—it pertains only to the imputation step.

Misclassification in the Validation Sample

Even "validation" data are rarely perfect. Ground truth measurements may contain their own modest degree of misclassification, reflecting the inherent limitation of any data collection process. In this setting, multiple imputation recovers the parameter estimates that would be obtained if *all* observations were measured with the same accuracy as the validation data, not the estimates that would arise under truly error-free measurement.

Error in the multiple imputation estimator can be decomposed as $\text{Error}(\hat{\beta}^{MI}) = (\hat{\beta}^{MI} - \hat{\beta}) + (\hat{\beta} - \beta)$, where $\hat{\beta}$ is the infeasible value of the estimator that would be recovered if data of the same quality as the validation data were available for the entire dataset and used in estimation. The first of these terms vanishes asymptotically: the multiple imputation estimator converges to the same limit as the complete data estimator.[21] Therefore, multiple imputation neither resolves nor amplifies bias present in the validation data; it merely extends the level of accuracy in that data to the full dataset.

Special Considerations for Panel Data

Econometric applications of panel data often include large sets of fixed effects, estimated as nuisance parameters. This poses a distinct challenge for imputation: fixed effects must be represented within the imputation model, or the dependence between observations sharing the same fixed effect will be lost in the imputed data. This requirement is often described as *congeniality* between the imputation model and the model of interest, i.e., the imputation model must allow for at least as rich a dependence structure as the model of interest (Meng, 1994).

The severity of this problem depends on the granularity of the fixed effects. In settings where the fixed effects correspond to coarser grouping, such as states or years in individual-level panels, incorporating them into the imputation model is straightforward. Each grouping

[21] Derivation of this fact is given in Appendix 3.

variable can be supplied as a categorical variable in the imputation model (Rubin, 1996). The design matrices for such variables are relatively compact, so computation remains tractable. This will generally suffice for models with misclassified outcome data, since the incidental parameters problem typically precludes the use of fine-grained fixed effects in nonlinear models anyway.

When fixed effects are high-dimensional, however, the curse of dimensionality becomes more acute. While there is nothing *mathematically* preventing the imputation model from accommodating thousands of fixed effects, estimation becomes unstable: sparse indicators contribute little information, runtime grows rapidly, and the Gibbs Sampler may fail to converge. In such cases, it is preferable to conduct imputation on a *wide* version of the dataset, where each time period's observation forms a separate variable with an individual's record, rather than the *long* version commonly used for estimation. If the validation data do not include complete time series for each individual, the Gibbs Sampler can be modified to draw the missing elements of the validation data in each iteration. This strategy would not alter the theoretical properties of the multiple imputation estimator, nor would it affect how downstream estimation occurs, since the imputed datasets can easily be converted back to long form.

Simulations

To evaluate the performance of multiple imputation using the HCMM-LD model, I conduct a series of simulations comparing four estimation strategies:

1. Complete Data Estimate: These estimates use the true value of the misclassified variable, Z. Because most of Z would be unobserved in an applied setting, this estimator is infeasible in practice but serves as an ideal benchmark: it isolates the amount of variation in parameter estimates attributable to sampling variability alone.

2. Naive Estimate: These estimates use the potentially misclassified version of the variable, $\tilde{Z}$, without correction. This corresponds to the standard practice of estimating models directly from imperfect predicted data, leading to bias and inconsistency.

3. Validation Only Estimate: These estimates are based solely on the validation sample. Since multiple imputation requires such a sample, it will often be possible to estimate the model of interest using it alone, though at the cost of reduced statistical power.[22] Except insofar as the reduced sample size makes asymptotic approximations unreliable, the validation only estimates should share the same expectation as the complete data estimates, differing primarily by having much wider confidence intervals. This comparison highlights the benefit of multiple imputation in combining information from the large, imperfect dataset to achieve narrower intervals without sacrificing consistency.

4. Multiple Imputation Estimate: These estimates use imputations generated from the HCMM-LD model. Unless otherwise noted, each estimate is based on 10 imputations. Details specific to each simulation design are provided in the sections that follow.

Completely Random Misclassification in the Outcome of a Simple Binary Logit

Unlike classical error in the outcome of a linear regression, misclassification in the outcome of a logit model induces asymptotic bias, even when the probability of misclassification is independent of all covariates (Michalek and Tripathi, 1980). To demonstrate the severity of this bias under even modest rates of misclassification, I simulate 500 datasets generated according to:

[22]When Z is the dependent variable and the validation data sampling is stratified on $\tilde{Z}$, however, multiple imputation becomes more than a simple power improvement, since the validation data are selected on (a potentially misclassified version of) the dependent variable. This still satisfies the ignorability assumption required for multiple imputation.

$$x \sim \mathcal{N}(0,1), \quad e \sim \text{Gumbel}(0,1), \quad y = \mathbf{1}(x + e > 0)$$

$$N = 5000, \quad N^* = 500, \quad \mathbb{P}(y \neq \tilde{y}) = 0.1$$

To estimate the HCMM-LD imputation model, I set $\{k^{(\mathcal{Y})}, k^{(\mathcal{X})}, k^{(\mathcal{Z})}\} = \{5, 5, 5\}$, iterate 50 times before making my first imputation, and then make 10 iterations between each of 10 imputations. Although these settings are lower than those recommended for applied work, the simplicity of this data generating process ensures rapid convergence.

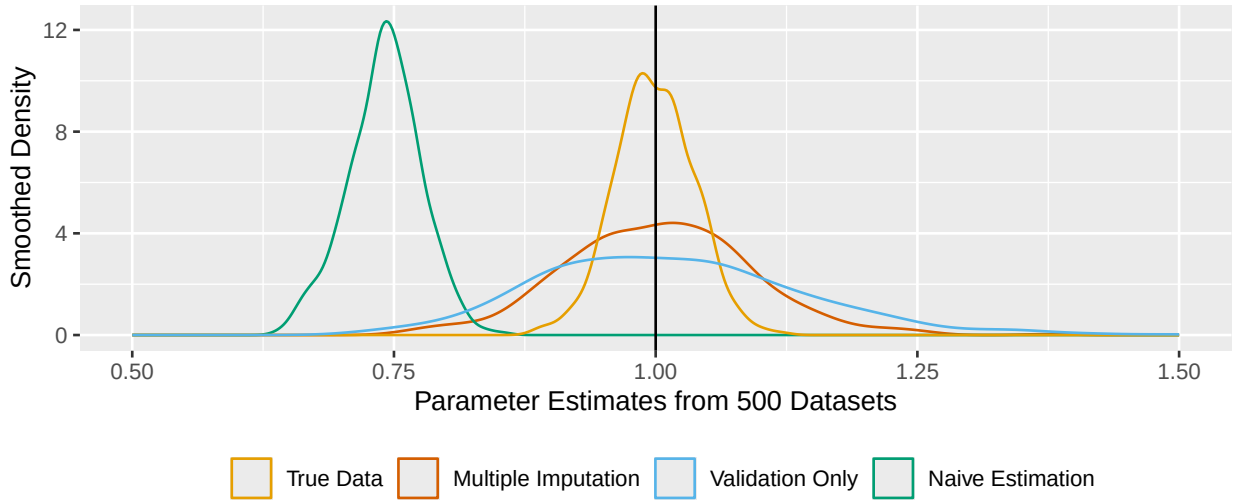


Figure 2: Density of Parameter Estimates from Simple Logit DGP with Completely Random Misclassification

Figure 2 displays the distribution of parameter estimates across the 500 simulated datasets for each of the four estimation methods. As expected, parameter estimates from the Naive estimation are severely biased with an average bias of -0.259, compared to 0.013 for the Validation Only estimates and 0.002 for the Multiple Imputation estimates. Among the latter two, the Multiple Imputation estimates exhibit lower variance. Empirical coverage rates across the 500 simulations are 0.948, 0.946, and 0.926 for the True Data, Validation Only, and Multiple Imputation methods, respectively—each close to the nominal rate of .95.

To test robustness under more extreme misclassification, I repeat the simulation with the

probability of misclassification tripled to $\mathbb{P}(y \neq \tilde{y}) = 0.3$. This represents severe measurement error—since y is binary with $\mathbb{E}(y) = 0.5$, a misclassification rate of 0.5 would render $\tilde{y}$ pure noise. As shown in in Figure 3, the Multiple Imputation estimate loses some of its variance advantage relative to the Validation Only estimate, both remain unbiased. The empirical coverage rate for the Multiple Imputation estimator increases slightly to 0.946.

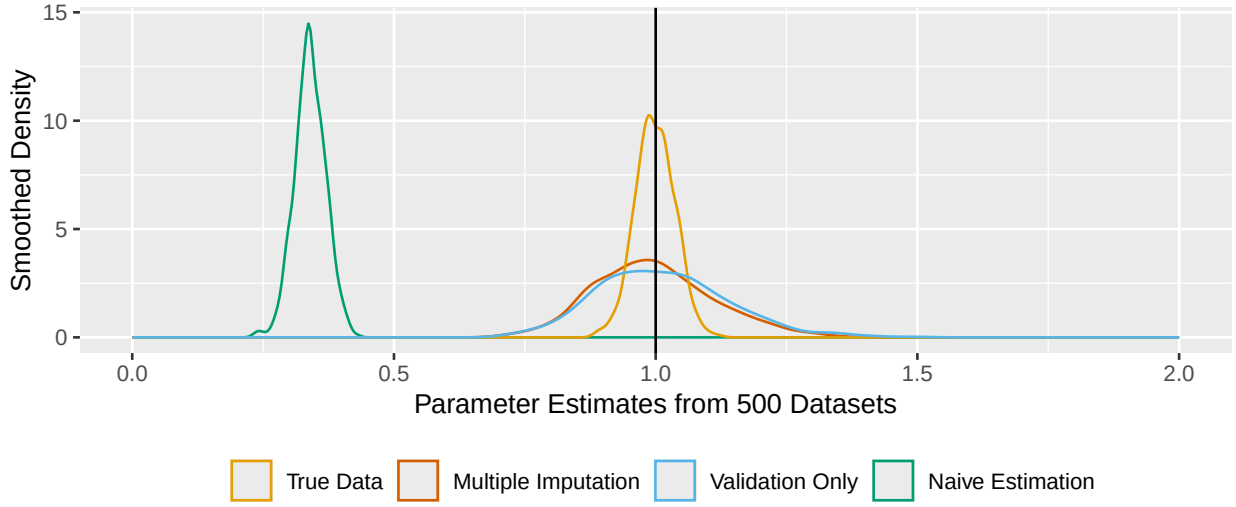


Figure 3: Density of Parameter Estimates from Simple Logit DGP with Completely Random and Severe Misclassification

Conditionally Random Misclassification in the Outcome of a Simple Binary Logit

The previous simulations examined the simplest case of completely random misclassification—an error structure independent of any covariates. A more realistic challenge arises when the probability of misclassification depends on the independent variable of interest. To study this case, I use the same data generating process as before, except that the probability of being misclassified now depends, on x: $\mathbb{P}(y \neq \tilde{y}|x) = 0.2\frac{e^x}{1+e^x}$. This specification yields an overall misclassification rate of 0.1, identical to the first simulation, but with the probability of misclassification increasing in x. Consequently, true values of $y = 1$ are more likely to be misclassified as $\tilde{y} = 0$ than vice versa. Estimation of the imputation model follows the same

procedure as before, except that I increase the number of pre-imputation iterations of the Gibbs Sampler to 100 due to the more complex data structure.

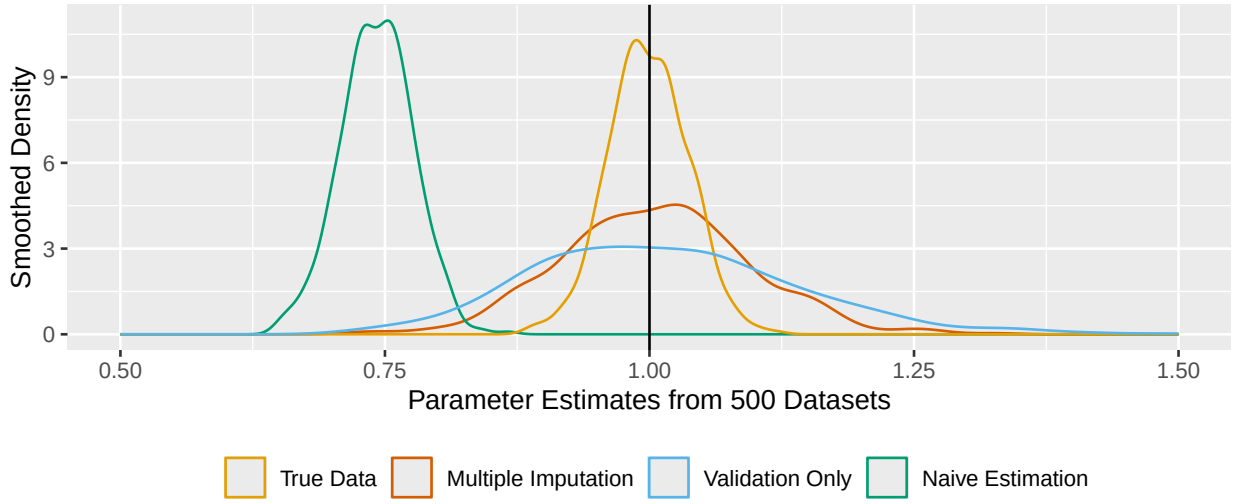


Figure 4: Density of Parameter Estimates from Simple Logit DGP with Dependent Misclassification

Figure 4 presents distributions of parameter estimates for each of the four estimation strategies across 500 simlulated datasets. The Validation and Multiple Imputation estimates again exhibit minimal average bias, while the Multiple Imputation estimator retains its advantage of lower variance. Empirical coverage remains close to nominal for all estimators, with Multiple Imputation at 0.92.

The above specification links misclassification to x in a way that makes errors more likely when y is relatively easy to predict—that is, at higher values of x. Now I explore a scenario where prediction error is highest near the decision boundary, i.e. when $x \approx 0$: $\mathbb{P}(y \neq \tilde{y}|x) = \frac{\sqrt{2}}{10} e^{-\frac{x^2}{2}}$, which again yields an overall misclassification rate of 0.1.[23]

Figure 5 shows that the results remain congruent with the earlier simulations. Both the Validation Only and Multiple Imputation estimators are unbiased, and the Multiple Imputation estimates again show tighter dispersion. Empirical coverage for the Multiple Imputation estimator rises to 0.944.

[23] By the law of total probability, $\mathbb{P}(y \neq \tilde{y}) = \int_{-\infty}^{\infty} \frac{\sqrt{2}}{10} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.1$

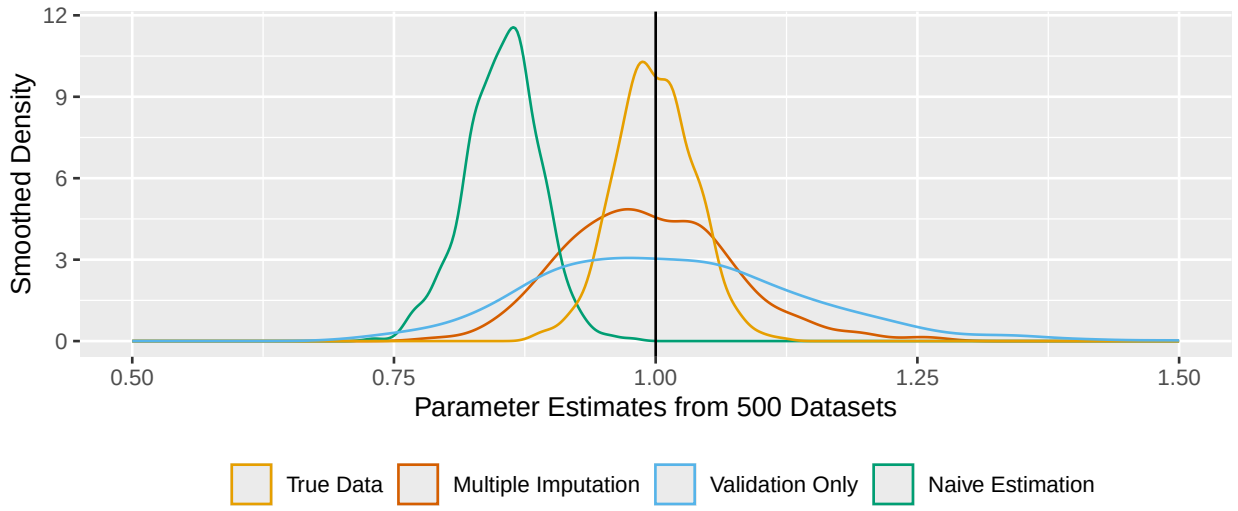


Figure 5: Density of Parameter Estimates from Simple Logit DGP with Dependent Misclassification Around x=0

Non-ignorable Validation Sampling

All previous simulations used validation data that perfectly satisfied the assumption of ignorable validation sampling. In real applications, validation samples of convenience may not satisfy this condition. These simulations are designed to assess whether multiple imputation remains worthwhile when only non-ignorable validation data are available. Specifically, they cover specifications under which multiple imputation is theoretically biased and inconsistent. The goal is to assess the magnitude of this bias in comparison to a naive estimate that applies no correction.

For these simulations, I use the following data generating process:

$$x \sim \mathcal{N}(0,1), \quad e \sim \mathcal{N}(0,1), \quad z \sim \text{Bernoulli}(0.5), \quad y = \mathbf{1}(x + \rho z + e > 0)$$

$$N = 5000, \quad N^* = 500, \quad \mathbb{P}(y \neq \tilde{y} \mid z) = \begin{cases} 0.1 & z = 1 \\ 0.3 & z = 0 \end{cases}$$

Here, ρ governs the correlation between z and y. To simulate non-ignorability, I construct

the validation data from a sample of 100 observations for which $z = 0$ and 400 observations for which $z = 1$. For each iteration, I estimate three models: a linear model of y on x (the "error-in-Y" model), a linear model of x on y (the "error-in-X" model) and a logit model of y on x. Each model is estimated four ways: using the "true data" with the true values of y, using the "naive data" with only the misclassified version of y, using multiple imputation with an ignorable imputation model (in which z is observed in imputation), and finally non-ignorable multiple imputation (estimated using only $\tilde{y}$ and x). The simulation is repeated 100 times for each distinct value of $\rho \in \{0, 0.1, \ldots, 1.9, 2\}$. Figure 6 plots the means of each parameter estimate and 95% confidence interval of that mean for each value of ρ.

Across all models and all values of ρ, the multiple imputation estimates based on either ignorable or non-ignorable specifications of the imputation model recover parameter estimates much closer to the true data estimates than the naive models. The ignorable validation results consistently outperform the non-ignorable results, but neither ever performs worse than the naive estimate. Of course, even higher values of ρ or more extreme forms of selection into the validation sample would change these results, and could potentially even make the non-ignorable multiple imputation estimate worse than the naive estimates. Still, the violations of ignorability in these simulations are non-trivial, and the results suggest that the multiple imputation approach remains reasonably robust to deviations from this assumption.

Taken together, the simulation results demonstrate that the HCMM-LD–based multiple imputation approach performs consistently well across a range of data-generating processes. Even when the probability of misclassification depends on covariates or when the validation sample violates ignorability, the method substantially reduces bias relative to naive estimators and maintains near-nominal coverage. These findings provide strong motivation to evaluate the approach in a real-world setting, where data limitations and complex dependence structures pose additional challenges.

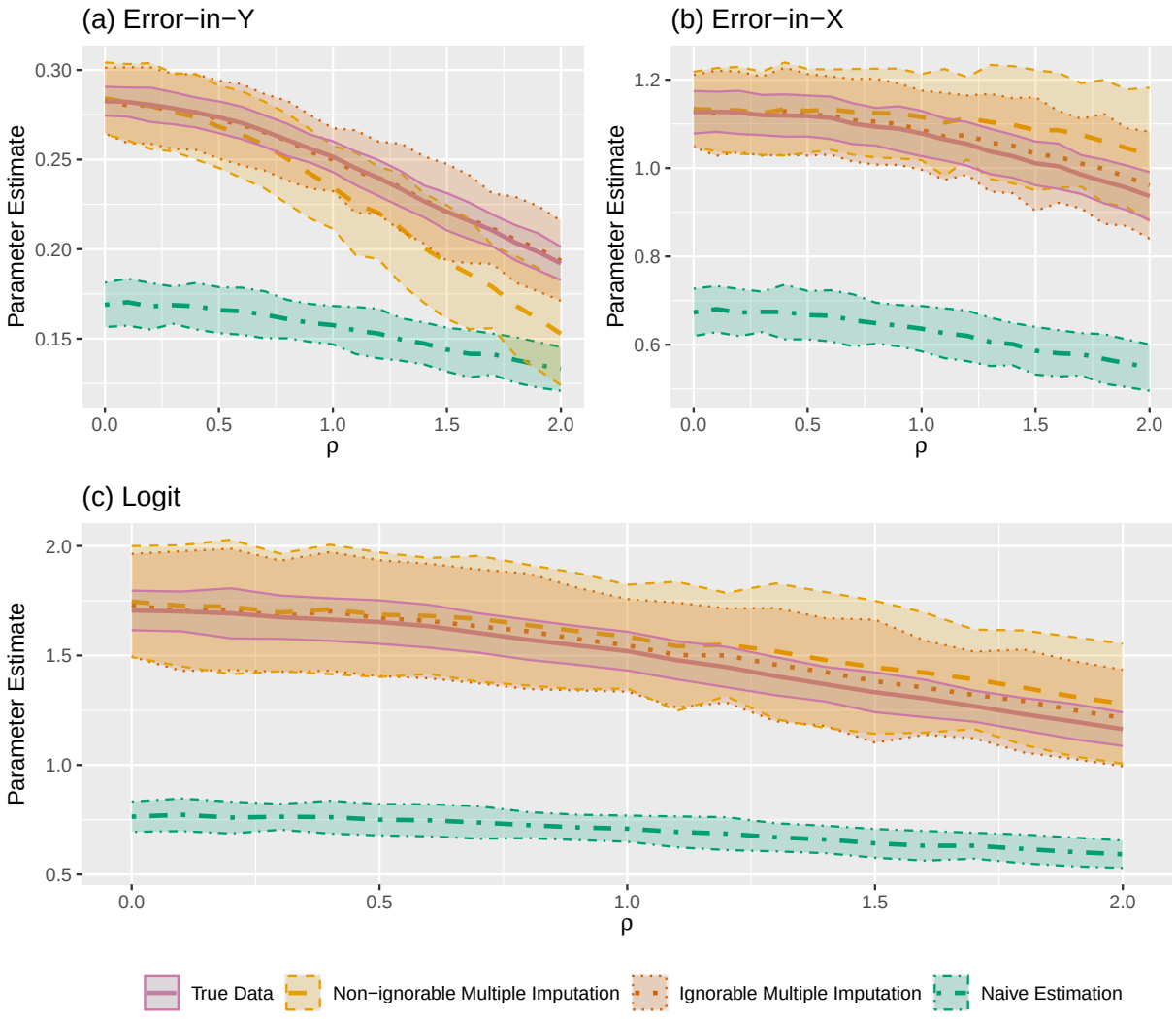


Figure 6: Results from Non-ignorable Sampling Simulations

Empirical Application: Demand for Short-run Water in Kern County

Satellite-based crop datasets, such as the USDA's Cropland Data Layer (CDL), are now ubitquitous in applied agricultural and environmental economics. These predicted datasets enable large-scale studies of crop choice and resource use but are known to contain classification errors that can bias downstream econometric analyses. I evaluate the accuracy of the CDL for Kern County, California, a major agricultural region at the southern end of the Central Valley, and show how misclassification in the CDL affects estimated relationships between water availability and crop choice.

Kern County is an ideal setting for this exercise. It is one of the most agriculturally productive counties in the United States, ranking third nationally in total crop sales in the most recent USDA Census of Agriculture. Due in large part to the county's heavy emphasis on diverse specialty crops, the CDL is known to have accuracy issues in this region (Espinoza et al., 2023). At the same time, Kern County's Deparment of Agricultural and Measurement Standards provides comprehensive, field-level ground truth data, a rare resource among U.S. counties.[24] Created in parnership with local growers, largely in support of a pesticide use reporting program, this dataset enables direct comparison between true crop choices and CDL predictions. Figure 7 shows a map of Kern County fields and high-level crop categories in 2022.

Data Sources and Accuracy Assessment

The CDL is generated from multiple remote sensing sources, including NASA/USGS Landsat imagery (USDA NASS, 2024). Raw data spectral readings from these satellites are georeferenced and calibrated for terrain, atmospheric conditions, solar angles, and sensor

[24] At the time of this writing, 2020-2023 are directly downloadable from their website, and 2018 and 2019 are available through archived copies via Wayback Machine. 2024 and 2025 are only available in provisional form.

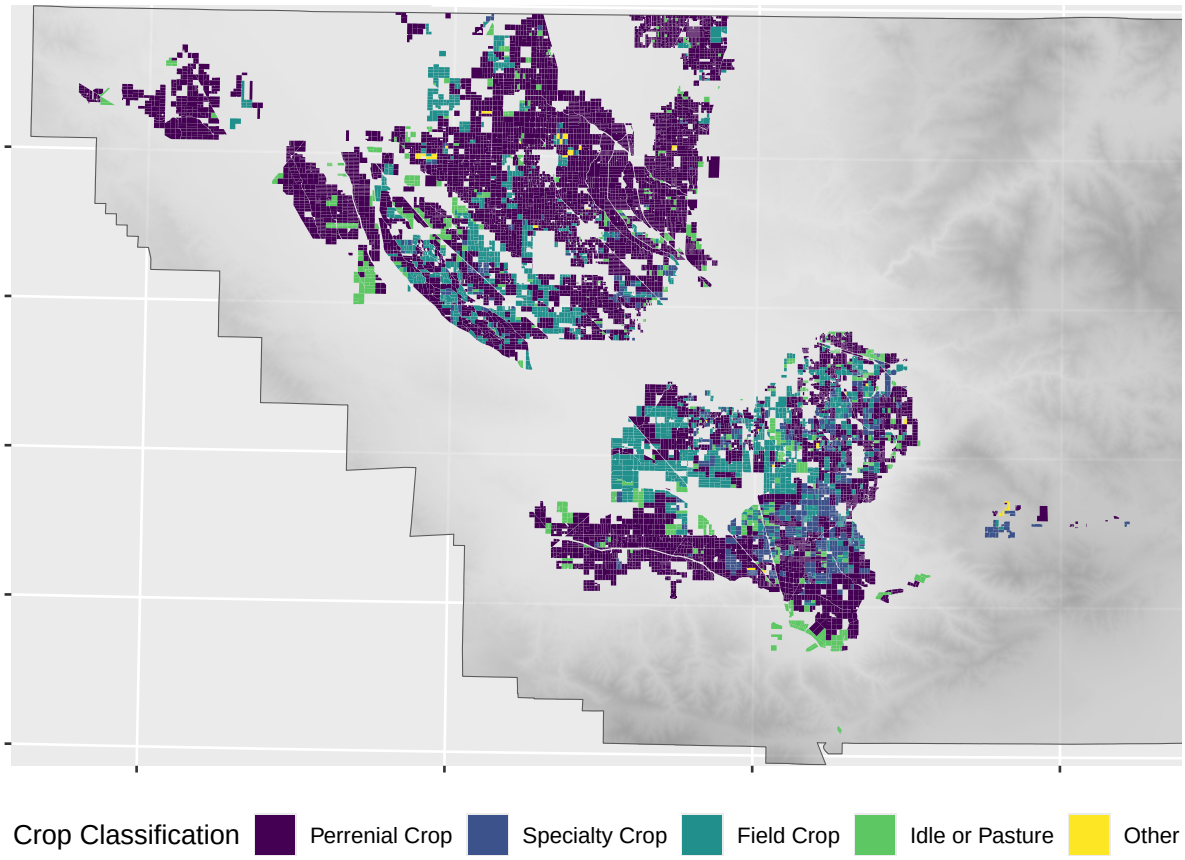


Figure 7: Map of High-Level Crop Categories on Kern County Fields, 2022

NOTES – Shapefiles for fields are the USDA Common Land Units. Crop categories are aggregated from granular crop data provided by the Kern County Department of Agriculture and Measurement Standards.

characteristics. The quality of these adjustments depend on the quality of other inputs, including the model of surface terrain, scientific understanding of how atmospheric and solar conditions affect the signal, and the precision of location and angle measurement. Each of these components is imperfect, and the degree of imperfection can be related to conditions on the ground like topographic roughness or prevailing weather. The CDL interprets this raw data using a decision tree, trained on Farm Services Agency (FSA) Common Land Unit data, a non-random sample generated from crop support program administration and therefore skewed toward subsidized crops (USDA NASS, 2024). For California as a whole in 2023, the CDL reports an overall accuracy of 80.7% for FSA-covered crops, but this masks substantial heterogeneity: widespread crops like rice achieve 98.2% producer accuracy, while specialty crops such as cucumbers or watermelons can be below 40%. User accuracies follow a similar pattern.[25]

The CDL is a raster dataset, whereas the Kern County dataset is a shapefile, and the two use different crop categorizations. A typical approach for linking these data in multiple imputation would assign each shapefile polygon a crop based on the pixels it overlaps. In this application, however, I avoid directly privileging the Kern County shapefiles to prevent a scenario where the CDL appears inaccurate simply because it is evaluated against the Kern County data's own boundaries. Instead, before carrying out any multiple imputation, I link both datasets to the USDA Common Land Unit (CLU) shapefiles, which represent single, contiguously-managed fields—the natural unit for modeling crop choice. For each field, the CDL value is assiged as the plurality of underlying pixels, while the Kern County value is assigned as the plurality of acreage from intersecting field polygons. To reconcile differences in crop classification schemes, I aggregate crops to higher-level categories (e.g., broccoli and cabbage are grouped as "Cole Crops"). This aggregation will only make the CDL appear more accurate, as misclassifications between crops within the same category are treated as

[25]Producer accuracy is the percent of actual fields of a crop that are correctly classified. Even relatively common crops for California like onions can have strikingly low producer accuracies (63.9 percent). User accuracy is the percent of classified fields of a crop that are correctly classified and is also called reliability.

correct.

Field-level comparison between the CDL and the Kern County ground truth data reveals substantial misclassification. Figure 8 shows a confusion matrix for these two sets of data for 2022. Each cell is a count of fields, cells along the diagonal of this matrix correspond to fields correctly classified in the CDL, all other cells represent misclassifications. Percentages across the top show producer accuracies—the percentage of fields actually growing a crop category that are classified as growing that crop category in the satellite data.

Overall accuracy of the CDL in Kern County in 2022 is 71.6 percent by field count, rising only to 72.3 percent by acreage. Some crop categories never appear in the CDL despite being planted on dozens of fields (e.g. cole crops), others are routinely misclassified with producer accuracies well below 50 percent (e.g. carrots). Idle and pasture land is particularly difficult to classify correctly, a common feature in crop satellite data, since distinguishing between grassland pasture and grain crop production has more to do with management than attributes of the plants visible from satellite imagery. Many perennial crops are routinely misclassified as idle and vice versa, perhaps because immature orchards look more like bare fields than fully developed orchards to a machine classifier's top-down view.

Crop Choice Model

Using these data, I examine how year-to-year variation in water deliveries to Kern County water districts affects farmers' crop choices. In California, and particularly in Kern County, most agricultural water is delivered through water districts, which either hold water rights directly or maintain long-term contracts with the Central Valley Project and State Water Project. These two projects are extensive systems of dams, canals, and pumps that move water from the wetter north to the drier, intensively farmed south. Direct water rights provide stable access to pre-specified volumes of surface water, while project contracts provide access to a stable *share* of project water in each year. These institutions create both long-run

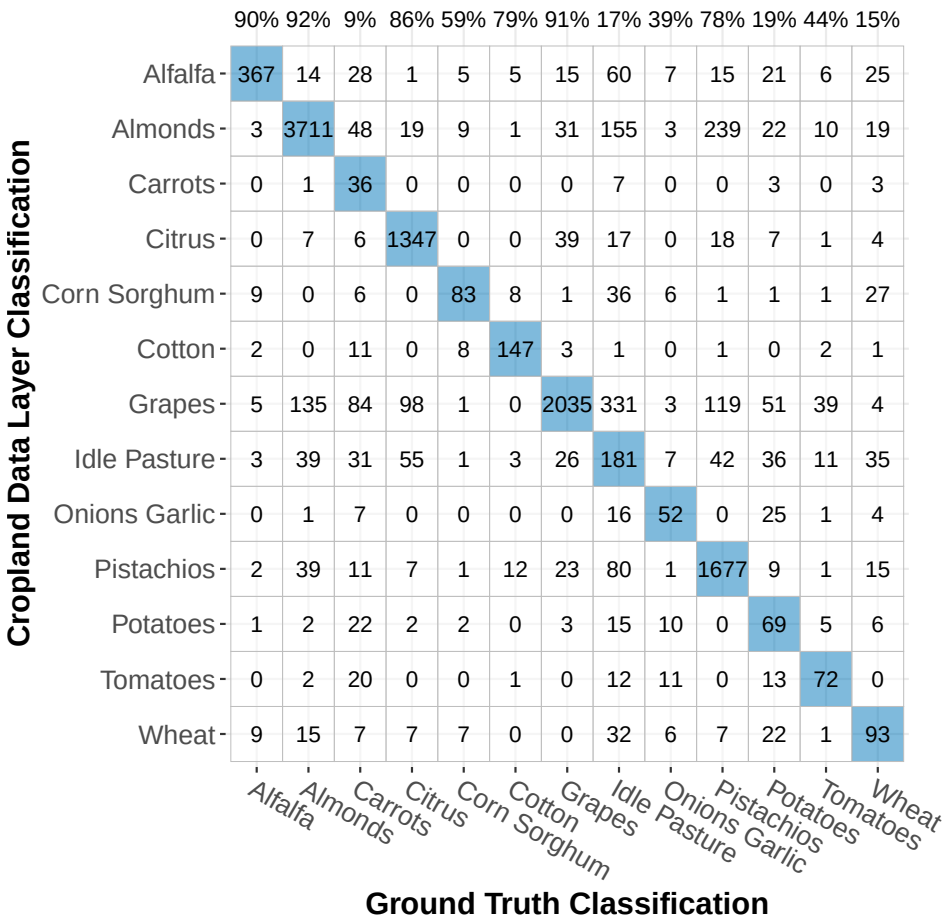


Figure 8: Confusion Matrix for Kern County Fields, 2022

NOTES – Data from the Kern County Department of Agriculture and Measurement Standards serves as ground truth. The ground truth data and the CDL raster data are intersected with USDA Common Land Unit Fields to produce these field counts. Individual crops are aggregated until categories perfectly match between the two datasets. Cell values correspond to counts of individual fields. Percentages across the top are producer accuracies. Crop categories making up less than one percent of Kern County field acreage are excluded; this causes minor disagreement between the producer accuracy percentages (which do include these categories) and values shown in the matrix. Correct classifications fall along the main diagonal, all other cells indicate misclassification.

differences in water availability across water districts and substantial year-to-year variation within districts. Long-run differences are clearly well-known to local farmers, and short-run variation, largely driven by variation in winter snow pack in the Sierra Nevada, is largely foreseeable by the time of planting.

Farms typically use water from their district allocation preferentially, with groundwater serving as a backstop. This water is cheap for farmers—water districts price water like a cost-covering utility rather than charging for the water's real marginal cost.[26] During the study period, the main cost of groundwater extraction was electricity for pumping, a non-trivial cost given that Kern County's over-extracted groundwater is often hundreds of feet below the surface.[27]

A farmer deciding what to plant faces substantially different water availability constraints depending on what water district they live in and snow pack conditions over the preceding winter. The feasibility and profitability of various planting choices depend on the water requirements of that crop, how much of that requirement can be met with district water, and how costly additional backstop water will be. Short-run variability in district water deliveries impacts cropping decisions and is an important component of how drought impacts agriculture. Understanding how short-run water variability influences crop choice is crucial for evaluating drought impacts and informing adaptive water management strategies.

Given the diversity of crops grown in Kern County, a fully granular crop choice model is infeasible. I aggregate crops to four broad categories: perennial crops (e.g. almonds, oranges, wine grapes), specialty crops (e.g. carrots, lettuces, watermelons), field crops (e.g. grains, potatoes, dry beans), and fallow or pasture land. Parcels with non-agricultural or unconventional agricultural uses (e.g. developed land, nursery greenhouses) are excluded. This aggregation eliminates some misclassification error, but as the earlier confusion matrix

[26]The water districts themselves do not pay the marginal cost of water either, since they take their water from rights they own or contracts with the two projects.

[27]Groundwater pumping costs scale with the square of the depth to groundwater.

showed, substantial measurement error remains in the CDL even at this higher level. Figure 9 shows the confusion matrix for these higher level aggregates.

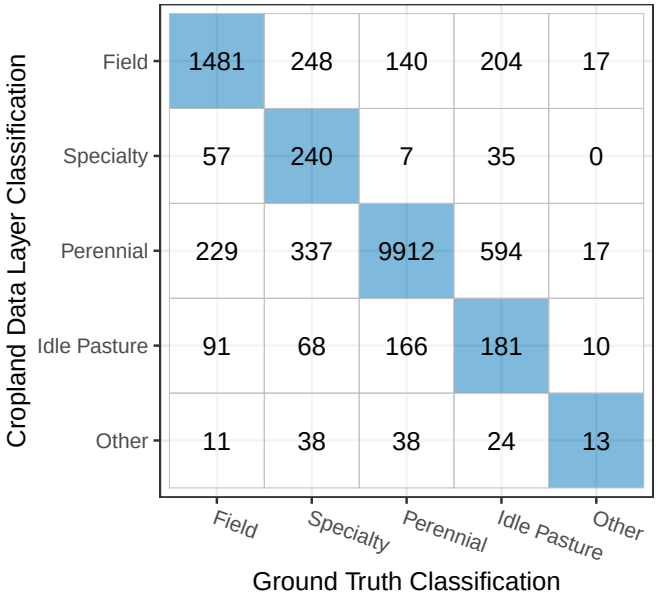


Figure 9: Confusion Matrix for Kern County Fields, 2022

NOTES – Data from the Kern County Department of Agriculture and Measurement Standards serves as ground truth. The ground truth data and the CDL raster data are intersected with USDA Common Land Unit Fields to produce these field counts. Cell values correspond to counts of individual fields. Correct classifications fall along the main diagonal, all other cells indicate misclassification.

I model the farmer's reaction to short-run water availability by specifying latent per-acre profit, where the crop $j \in \{Perennial, Specialty, Field, Fallow\}$ planted on field i in water district d in year t is chosen to maximize y_{ijt}:

$$y_{ijt} = \beta\ p_{jt} + \gamma_j \text{Water Delivered per Acre}_{dt} + \theta_{jd} + e_{ijt}$$

where Fallow is the holdout category. Since reliable price series are unavailable for most specialty crops, I proxy p_{jt} using revenue per acre calculated from Kern County's annual crop reports, aggregated to the high-level crop categories with acreage weights. Water deliveries

include water delivered to district d by the Central Valley Project and State Water Project. Collection of this data follows the procedures set forth by Hagerty (2024). District $\times$ crop-specific intercepts capture time invariant crop-specific costs in a way that allows differences across districts. This captures both long-run differences in district water availability and the costs of groundwater pumping, since depth to groundwater is highly spatially correlated and districts are, for the most part, geographically contiguous. My data cover cropping choices from 2018-2023 for 14917 fields across 14 water districts.[28] Though this is a relatively short panel, owing to the availability of the Kern County data, this timeframe does span both dry and wet years.

A farmer chooses crop j when $y_{ijt} > y_{ikt}$ for all other k, i.e.

$$\mathbb{P}(j = a) = \mathbb{P}(y_{iat} > y_{ibt} \ \forall \ b)$$

$$= \mathbb{P}(\beta(p_{at} - p_{bt}) + (\gamma_a - \gamma_b)(\text{Water Delivered per Acre}_{dt}) + (\theta_{ad} - \theta_{bd}) > e_b - e_a \ \forall \ b)$$

Assuming e_{ijt} is distributed Gumbel$(0, 1)$, this is a multinomial logit model, paralleling a random utility framework for a consumer's problem. Analagous to computing willingness to pay in consumer choice models, the ratio γ_j/β estimates the marginal revenue product of short-run water availability for crop j.

I estimate the crop choice model using four approaches:

1. Complete True Data: The model estimated using the full Kern County dataset.
2. Complete Satellite Data: The model estimated using exclusively CDL data.
3. Multiple Imputation (Ignorable): The model estimated using a random 20 percent sample of the Kern County data and the complete CDL data with multiple imputation. I set $\{k^{(\mathcal{Y})}, k^{(\mathcal{X})}, k^{(\mathcal{Z})}\} = \{10, 20, 3\}$ (calibrated to ensure truncation is not excessive). The

[28] The included water districts are Arvin-Edison, Semitropic, Wheeler Ridge-Maricopa, Kern Delta, North Kern, Cawelo, Shafter-Wasco, Belridge, Rosedale-Rio Bravo, Buena Vista, Kern-Tulare, Berrenda Mesa, Delano-Earlimart, and Tehachapi-Cummings.

model is iterated until convergence before generating 100 imputations. The imputation model includes indicators for district and year but omits the price proxy, which is a deterministic function of year and crop, which are already nonparametrically included. This variable is reintroduced in estimation of the crop choice model.

4. Multiple Imputation (Non-ignorable): Identical to the previous model, except that the validation sample is a non-ignorable sample of fields adjacent to roads. This mimics a plausible form of non-ignorability that could appear in real world validation data. Roughly 57% of fields in Kern County fall within the 100 meter distance to road cutoff for being eligible for inclusion in the non-ignorable validation sample. No measure of road proximity is used in the imputation step.

Table 1 reports the estimated share of fields in each crop category in 2022 under each method. Both multiple imputation models reproduce crop shares that closely match the true data, outperforming the satellite only estimates.

Table 1: Percent of Kern County Fields in Each Crop Category (2022)

	True Data	Satellite Data	Multiple Imputation	Multiple Imputation
	(1)	(2)	(3)	(4)
Perennial Crops	71.5%	78.3%	71.2%	72.5%
	(0.4)	(0.4)	(0.4)	(0.4)
Specialty Crops	13.4%	15.3%	13.7%	13.3%
	(0.3)	(0.3)	(0.3)	(0.2)
Field Crops	7.2%	2.6%	6.7%	6.5%
	(0.2)	(0.1)	(0.2)	(0.3)
Idle or Pasture	7.9%	3.9%	8.1%	7.7%
	(0.2)	(0.2)	(0.3)	(0.3)
Validation Data			Ignorable	Non-ignorable
Validation Share	1	0	0.2	0.2
Imputations	NA	NA	100	100

Table 2 reports the estimated coefficients for key variables in the crop choice model. Relative

to the model estimated with full true data, the CDL only estimates exhibit sign reversal and spurious significance. The multiple imputation results, by contrast, align more closely with the true data. Estimates from the true data fall within the confidence interval for both multiple imputation estimates and vice versa.

Table 2: Crop Choice Model Results for Key Variables

	True Data	Satellite Data	Multiple Imputation	Multiple Imputation
	(1)	(2)	(3)	(4)
Revenue per Acre ($1000s)	0.006	−0.018***	0.015	0.011
	(0.005)	(0.006)	(0.022)	(0.014)
Delivery per Acre (Field, af)	0.019	−0.087**	0.059	0.077
	(0.034)	(0.037)	(0.207)	(0.116)
Delivery per Acre (Specialty, af)	0.074**	0.088**	0.065	0.073
	(0.035)	(0.042)	(0.230)	(0.120)
Delivery per Acre (Perennial, af)	0.017	−0.012	0.050	0.022
	(0.025)	(0.027)	(0.166)	(0.099)
Validation Data			Ignorable	Non-ignorable
Unique Field × Years	72535	72151	Varies	Varies
Validation Share	1	0	0.2	0.2
Imputations	NA	NA	100	100

Note: $p < 0.001$***, $p < 0.05$**, $p < 0.1$*. Each estimate is from a multinomial logit crop choice model. Each model also contains a water district × crop-specific intercept.

Table 3 shows results for the marginal revenue product (MRP) of one additional acre-foot per acre of short-run water delivered to the water district. Before comparing estimates, note that differences in MRP across crop categories are expected, even in the true data. In a simple production model with no frictions, all producers would use water until its MRP was equal to its marginal cost, which would also be equal to the equilibrium price of water.

However, the district-level water variable reflects deliveries to the water district, normalized by total acreage, rather than actual on-farm deliveries. Within a district, the water consumption

Table 3: Marginal Revenue Product of One Additional acre-foot per Acre of Short-run Water Availability

Marginal Revenue Product of Water	True Data	Satellite Data	Multiple Imputation	Multiple Imputation
Crop Category	(1)	(2)	(3)	
Field Crops	\$3,210	\$4,970	\$3,930	\$7,000
Specialty Crops	\$12,700	-\$5,200	\$4,330	\$6,640
Perennial Crops	\$2,930	\$700	\$3,330	\$2,000
Validation Data			Ignorable	Non-ignorable
Validation Share	1	0	0.2	0.2
Imputations	NA	NA	100	100

Note: Each estimate is the ratio of the crop-specific water delivery per acre variable and the price variable in a multinomial crop choice model. Standard errors cannot be computed via the Delta method because this ratio is not a continuously differentiable function of the parameter estimates (the partial derivative with respect to the price coefficient is undefined at zero, which is not a rejected null hypothesis in the true data or multiple imputation models).

of idle fields and perennials may not adjust much to annual variation in water availability, while annuals and specialty crop growers can more flexibly adjust to take advantage of available surplus. As a result, one additional acre-foot per acre at the district level may correspond to multiple additional acre-feet effectively available for individual farmers, inflating estimated MRPs relative to market prices.

Second, California's water markets are far from frictionless. Beyond the typical transaction costs involved in any one-off, specialized trade, water trades involve delivery costs for using the state canal system (which frequently face capacity constraints) and administrative costs associated with state review from multiple agencies. Across California, these costs are enough to stymie hundreds of millions of dollars of potential gains from trade annually (Hagerty, 2024). The large fixed transfer costs and uncertainty surrounding reviews make the market for short-term water trades very thin (Regnacq et al., 2016).

Finally, Kern County's local conditions, including high water demand, deep and depleting

aquifers, and limited surface water, result in some of the state's highest observed water prices, consistent with high MRPs.

Due to the small and insignificant coefficient estimates on revenue per acre ($\hat{\beta}$) in the choice model regressions, the derived MRP ratios ($\hat{\gamma}_j/\hat{\beta}$) are necessarily volatile. Still, across all crop categories, the ignorable multiple imputation estimates are consistently closer to the true data than the naive CDL estimates, confirming that multiple imputation can substantially correct bias from misclassification. The relative bias reduction ranges from 53.2% for specialty crops to 82.1% for perennials.[29] Even the non-ignorable multiple imputation estimates are closer for all crops aside from field crops.

The empirical results highlight the potential for multiple imputation to serve as a bridge between high-quality microdata and large-scale predicted data. By leveraging flexible imputation modeling, the approach recovers not only accurate descriptive statistics but also valid structural parameters from highly flawed predicted satellite data.

Conclusion

Predicted categorical data, such as crop classification from satellite imagery or sentiment labels inferred from natural language processing of text, are now integral to empirical research across economics and the broader social sciences. Yet these data inevitably contain misclassification, and naively treating predictions as truth can induce substantial biases in downstream analyses. This paper develops a general, implementable method for correcting such biases using multiple imputation informed by a limited validation sample. The method requires only a small subset of correctly labeled observions and does not depend on knowing or specifying the form of the misclassification process.

By using a hierarchical Dirichlet process mixture model, a form of Bayesian nonparametric

[29] I calculate percent bias reduction using the formula $100 \times (1 - (|MRP^{MI} - MRP^{True}|/|MRP^{Naive} - MRP^{True}|))$.

model, for imputation, the proposed approach flexibly captures the joint distribution of predicted and true categorical variables, as well as the dependence of this joint distribution on other covariates. In both simluations and an empirical application estimating the marginal revenue product of short-run water in Kern County agriculture, the multiple imputation method substantially reduces bias relative to analyses that rely on predicted data alone. Even when validation samples are limited or non-representative, the approach delivers meaningful improvements over naive estimation. These improvements are realized without requiring the researcher to retrain the prediction model or alter the code used for model estimation.

To take proper advantage of the research opportunities that predicted data make available, addressing misclassification must become standard practice in applied work. The approach presented here offers a practical and widely applicable solution: it treats predicted data as informative but uncertain, and integrates that uncertainty into inference through multiple imputation.

Appendix 1: Gibbs Sampler for the HCMM-LD

Before introducing the Gibbs Sampler, I first cover the remaining priors and hyperparameters that are not fully described in the main text.

As priors for the parameters in the X component, I use

$$\psi_s^{(j)} \overset{\text{iid}}{\sim} \text{Dirichlet}(\gamma_{s1}^{(j)}, \ldots, \gamma_{sd_j}^{(j)})$$

with each hyperparameter γ set equal to 1 to initialize. Posteriors are insensitive to this choice versus other minimally informative priors.

For the Y component, I use

$$\{(B_r, \Sigma_r)\} \overset{\text{iid}}{\sim} \text{MatN}(B_0, I, T_B) \times \text{IW}(q+2, \Sigma)$$

$$(B_0, \Sigma) \sim \text{MatN}(0, I, \sigma_{0\beta}^2 I) \times W\left(q+1, \frac{1}{q+1} I\right)$$

$$T_B = \text{diag}(1/\tau_1, \ldots, 1/\tau_q)$$

with priors for each $\tau_1, \ldots, \tau_q$ drawn independently from $G(0.5, 0.5)$ and $\sigma_{0\beta}^2 = 10$. Each of these choices match those made by Murray and Reiter (2016) and are intended to be uninformative.

With these preliminaries, I now describe the Gibbs Sampler. Note that this only contains slight variations from Murray and Reiter (2016), mostly to simplify the process for the context of estimating only on the complete cases.

1. Draw top-level component membership

For each observation, sample Z_i from

$$\mathbb{P}(Z_i = z | S_i = s, R_i = r) \propto \lambda_z \phi^{(\mathcal{X})}_{zs} \phi^{(\mathcal{Y})}_{zr}$$

2. Draw continuous and categorical component membership

For each observation sample s_i and r_i from

$$\mathbb{P}(S_i = s | Z_i = z, X_i = x_i, \cdot) \propto \phi^{(\mathcal{X})}_{zs} \prod_{j=1}^{p} \psi^{(j)}_{s x_{ij}}$$

$$\mathbb{P}(R_i = r | Z_i = z, Y_i = y_i, X_i = x_i, \phi^{(\mathcal{Y})}_z, B_h, \Sigma_h) \propto \phi^{(\mathcal{Y})}_{zr} N(y_i; D(x_i) B_r, \Sigma_r)$$

3. Update the within-component prior distributions and draw parameters

Draw component parameters for each $1 \leq s \leq k^{(\mathcal{X})}$ and $1 \leq j \leq p$ from

$$\psi^{(j)}_s \sim \text{Dirichlet}\left(\gamma_{s1} + \sum_{i=1}^{n} \mathbf{1}(S_i = s, X_{ij} = 1), \ldots, \gamma_{sd_j} + \sum_{i=1}^{n} \mathbf{1}(S_i = s, X_{ij} = d_j)\right)$$

and for each $1 \leq r \leq k^{(\mathcal{Y})}$ and $1 \leq v \leq q$ let B_{rv} be the vth column of B_r and sample

$$B_{rv} \sim N((\tau_v I + \mathbf{D}_r' \mathbf{D}_r / \tilde{\sigma}^2_{rv})^{-1} (\tau_v B_{0v} + \mathbf{D}_r' \tilde{\mathbf{y}}_{rv} / \tilde{\sigma}^2_{rv}), (\tau_v I + \mathbf{D}_r' \mathbf{D}_r / \tilde{\sigma}^2_{rv})^{-1})$$

where $\mathbf{D}_r$ is the matrix obtained by stacking the vector $\{D(x_i) : R_i = r\}$, $\tilde{\mathbf{y}}_{rv}$ is the vector obtained by concatenating $\{y_{iv} - \tilde{\mu}_{iv} : R_i = r\}$ and $\tilde{\mu}_{iv}$ and $\tilde{\sigma}^2_{rv}$ are parameters of the

conditional normal distribution for $(Y_{iv}|Y_{i/v}, R_i = r)$, where

$$(Y_{iv}|Y_{i/v}, R_i = r) \sim N(D(X_i)B_{rv} + \tilde{\mu}_{iv}, \tilde{\sigma}^2_{rv})$$

and finally for each $1 \leq r \leq k^{\mathcal{Y}}$ sample

$$\Sigma_r \sim IW(d + \sum_{i=1}^{n} \mathbf{1}(R_i = r), \Sigma + S_r)$$

where

$$S_r = \sum_{i:R_i=r} (Y_i - D(x_i)B_r)(Y_i - D(x_i)B_r)'$$

These parameter draws will be used in the next iteration of Step 2.

4. Update and draw hyperparameters

For each entry of B_0 sample

$$B_{0jv} \sim N\left((k^{(\mathcal{Y})}\tau_v + 1/\sigma^2_{0\beta})^{-1}\tau_v \sum_{r=1}^{k^{(\mathcal{Y})}} B_{rjv}, (k^{(\mathcal{Y})}\tau_v + 1/\sigma^2_{0\beta})^{-1}\right)$$

and for $1 \leq v \leq q$ sample

$$\tau_v \sim G\left(\frac{a_\tau + k^{(\mathcal{Y})}p^*}{2}, \frac{b_\tau + \sum_{r=1}^{k^{(\mathcal{Y})}}(B_{rv} - B_{0v})'(B_{rv} - B_{0v})}{2}\right)$$

These parameter draws will be used in the next iteration of Step 3.

5. Draw stick-breaking parameters

Resample λ from

$$\xi_z^{(\mathcal{Z})} \sim \text{Beta}\left(1 + m_z, \alpha + n - \sum_{l=1}^{z} m_l\right)$$

for $1 \leq z \leq k^{(\mathcal{Z})} - 1$, with $\xi^{(\mathcal{Z})}_{k^{(\mathcal{Z})}} = 1$, and where $m_z = \sum_{i=1}^{n} \mathbf{1}(Z_i = z)$ and set $\lambda_z = \xi^{(\mathcal{Z})}_z \prod_{l<z} \left(1 - \xi^{(\mathcal{Z})}_l\right)$.

For $1 \leq z \leq k^{(\mathcal{Z})}$, iterate over $1 \leq s \leq k^{(\mathcal{X})} - 1$ and $1 \leq r \leq k^{(\mathcal{Y})} - 1$ sampling

$$\xi^{(\mathcal{X})}_{zs} \sim \text{Beta}\left(1 + t^{(\mathcal{X})}_{zs}, \beta^{(\mathcal{X})} + m_z - \sum_{l=1}^{s} t^{(\mathcal{X})}_{zl}\right)$$

$$\xi^{(\mathcal{Y})}_{zr} \sim \text{Beta}\left(1 + t^{(\mathcal{Y})}_{zr}, \beta^{(\mathcal{Y})} + m_z - \sum_{l=1}^{r} t^{(\mathcal{Y})}_{zl}\right)$$

where $t^{(\mathcal{X})}_{zs} = \sum_{i=1}^{n} \mathbf{1}(Z_i = z)\mathbf{1}(S_i = s)$ and $t^{(\mathcal{Y})}_{zr}$ is defined comparably. Generate $\phi^{(\mathcal{X})}_{zs}$ and $\phi^{(\mathcal{Y})}_{zr}$ from the ξs in the same was as in initialization.

These parameters will be used in the next iteration of Step 1.

6. Draw stick-breaking hyperparameters

Let $\mathcal{Z}_{occ}$ be the set of occupied top-level clusters with n_{occ} equal to the cardinality of $\mathcal{Z}_{occ}$. Sample

$$\alpha \sim G(a_0 + k^{(\mathcal{Z})} - 1, b_0 - \log(\lambda_{k^{(\mathcal{Z})}}))$$

$$\beta^{(\mathcal{X})} \sim G\left(a^{(\mathcal{X})} + n_{occ}(k^{(\mathcal{X})} - 1), b^{(\mathcal{X})} - \sum_{z \in \mathcal{Z}_{occ}} \log\left(\phi^{(\mathcal{X})}_{zk^{(\mathcal{X})}}\right)\right)$$

$$\beta^{(\mathcal{Y})} \sim G\left(a^{(\mathcal{Y})} + n_{occ}(k^{(\mathcal{Y})} - 1), b^{(\mathcal{Y})} - \sum_{z \in \mathcal{Z}_{occ}} \log\left(\phi^{(\mathcal{Y})}_{zk^{(\mathcal{Y})}}\right)\right)$$

with the standard choices $a^{(\mathcal{X})} = a^{(\mathcal{Y})} = b^{(\mathcal{X})} = b^{(\mathcal{Y})} = 0.5$. Following the implementation of Kim et al. (2024), I set a maximum value for the Gamma rate parameters of 10 to prevent floating point issues with taking logs of extremely small probabilities.

These hyperparameters will be used in the next iteration of Step 5.

To assess convergence, I compute the log-likelihood of the ground truth data for successive iterations of the model and estimate multiple chains of the Gibbs Sampler. I then monitor

for levelling-off of the log-likelihood with mixing among chains.

Appendix 2: Consistency of the HCMM-LD

Let $y \in \mathbb{R}^p$ and $x \in \mathcal{X} = \{0,1\}^d$, the finite discrete space of indicators corresponding to dummy variable representations of all discrete variables in the model. Let $f_0(x,y)$ be the true joint density of (X,Y), $g_0(y|x)$ the true conditional density, and $h_0(x)$ the true marginal probability mass function. By Bayes' theorem, $f_0(x,y) = g_0(y|x)h_0(x)$.

Let $f_n(x,y)$ be the posterior draw from the HCMM-LD after observing n samples, with $g_n(y|x)$ and $h_n(x)$ the resulting conditional density and marginal probability mass functions. As described in the main text and by Murray and Reiter (2016), two existing proofs from the literature are useful here and are invoked as assumptions.

A1. L_1-consistency of the conditional $g_n(y|x)$ (Wu and Ghosal, 2010)[30]

$$\sum_{x \in \mathcal{X}} ||g_n(y|x) - g_0(y|x)||_1 h_0(x) \rightarrow_{P_{f_0}} 0$$

A2. L_1-consistency of the marginal $h_n(x)$ (Dunson and Xing, 2009)

$$||h_n(x) - h(x)||_1 \rightarrow_{P_{f_0}} 0$$

and additionally

A3. The model prior Π has full Kullback-Leibler (K-L) support at $f_0(x,y)$. That is, for every $\epsilon > 0$, $\Pi(f : D_{KL}(f_0||f) < \epsilon) > 0$.

I aim to prove the L_1 consistency of the joint $f_n(x,y)$, i.e.

$$\sum_{x \in \mathcal{X}} \int |f_n(x,y) - f_0(x,y)| \, dy \rightarrow_{P_{f_0}} 0$$

[30] Note that the finiteness of $\mathcal{X}$ is critical here. Wu and Ghosal (2010) show that $||g_n(y|x) - g_0(y|x)||_1 \rightarrow_{P_{f_0}} 0$ for any fixed x. It is only because $\mathcal{X}$ is finite and $0 \leq h_0(x) \leq 1$ that this assumption follows from their result.

The limit expression can be rewritten

$$\sum_{x \in \mathcal{X}} \int |g_n(y|x)h_n(x) - g_0(y|x)h_0(x)| \; dy$$

which, by the triangle inequality

$$\leq \sum_{x \in \mathcal{X}} \int |g_n(y|x) - g_0(y|x)|h_n(x) \; dy \; + \sum_{x \in \mathcal{X}} \int g_0(y|x) \; dy \; |h_n(x) - h_0(x)|$$

Noting that $\int g_0(y|x) \; dy = 1$, it is clear that the second term $\rightarrow_{P_{f_0}} 0$ by A2. Define the first term as A_n. I add $||g_n(y|x) - g_0(y|x)||_1(h_0(x) - h_0(x))$ to the summand of A_n and rearrange. I then apply the triangle inequality and use the fact that $g_n(y|x)$ and $g_0(y|x)$ are both densities, implying $||g_n(y|x) - g_0(y|x)||_1 \leq 2$, to give

$$A_n \leq \sum_{x \in \mathcal{X}} ||g_n(y|x) - g_0(y|x)||_1 h_0(x) + 2 \sum_{x \in \mathcal{X}} |h_n(x) - h_0(x)|$$

Both terms $\rightarrow_{P_{f_0}} 0$ by A1 and A2. QED.

What remains to be shown is which underlying distributions $f_0(x, y)$ are within the K-L support of the model prior and therefore satisfy A3. An exhaustive description of all such distributions is beyond the scope of this paper, but appealing to existing results for the marginal and conditional distributions of the model again serves to demonstrate support for a wide class of joint densities.

Note that $D_{KL}(f_0(x, y)||f(x, y))$ can be decomposed as

$$D_{KL}(f_0(x, y)||f(x, y)) = \sum_{x \in \mathcal{X}} h_0(x) \left[\log \frac{h_0(x)}{h(x)} \right] + \sum_{x \in \mathcal{X}} h_0(x) \int_{\mathbb{R}^p} g_0(y|x) \left[\log \frac{g_0(y|x)}{g(y|x)} \right] dy$$

Mixtures of normal distributions are well-known universal approximators for distributions with smooth densities. A number of papers establish different criteria under which the K-L support of a Dirichlet process mixtures of normal distributions will support a density for

continuous variables—a typical requirement is that the true density is continuous and that its tails are adequately small, even weaker criteria are outlined below. Dunson and Xing (2009) establish the K-L support for any finite product multinomial distribution offered by Dirichlet process mixtures of Dirichlet distributions. From these two facts taken together, there exists a wide variety of mixed densities $f = g \cdot h \in \text{Support}(\Pi)$ such that both terms in the above decomposition are less than $\epsilon/2$ and the total K-L divergence is less than ϵ.

For example, one set of sufficiency conditions would be:

1. Positivity of Discrete Support: $h_0(x) > 0$ for all $x \in \mathcal{X}$
2. $g_0(y|x)$ is bounded above
3. $g_0(y|x)$ has finite Shannon entropy
4. $g_0(y|x)$ has mild local variation (no extreme spikes or oscillations), i.e. for some $\delta > 0$, $\int_{\mathbb{R}^p} g(y|x) \log\left(\frac{g_0(y|x)}{\phi_\delta(y)}\right)\, dy < \infty$ where $\phi_\delta(x) = \inf_{||t-y||<\delta} g_0(t,x)$ for all $x \in \mathcal{X}$
5. $g_0(y|x)$ does not have very heavy tails, i.e. for some $\rho > 0$, $\int_{\mathbb{R}^p} g(y|x)||y||^{2(1+\rho)p}\, dy < \infty$ for all $x \in \mathcal{X}$

Condition 1 is not necessary to prove $h_n(x) \to h(x)$ and is therefore not covered in Dunson and Xing (2009); easy modification to $\mathcal{X}$ can accommodate situations where certain combinations of discrete variable values are not possible. Conditions 2-5 are directly from Wu and Ghosal (2010) and only require a set of mild regularity conditions, the violation of which make density approximation of any kind very difficult. Reassuringly, Kleijn and van der Vaart (2006) show that even if the true density $f_0(x,y)$ is not in the support of Π, $f_n(x,y)$ will converge to a density $f^*(x,y) \in \text{Support}(\Pi)$ that minimizes $D_{KL}(f_0(x,y)||f^*(x,y))$, offering some sense of robustness even in the unlikely case that a true density fails these mild regularity conditions.

Of course, the ultimate aim of estimating this model is not to obtain a consistent estimate of the joint density, but to obtain a consistent estimate of the conditional density $m_0(x_a|x_b, y)$, where x_a and x_b are a partition of $x \in \mathcal{X}$ and x_a are the indicators corresponding to the

discrete variable for which we are generating multiple imputations. Fortunately, the L_1 convergence of this conditional density with respect to the true underlying density of the conditioning variables $s_0(x_b, y)$ is a direct consequence of the L_1 convergence of the joint distribution. Letting m_0 and m_n be the true and sample conditional distributions of x_a, s_0 and s_n the analogous marginal distributions of x_b and y, and f_0 and f_n the analogous joints,

$$\sum_{\mathcal{X}_a}\sum_{\mathcal{X}_b}\int_{\mathcal{Y}}|m_n - m_0|s_0\ dy = \sum_{\mathcal{X}_a}\sum_{\mathcal{X}_b}\int_{\mathcal{Y}}|(m_n s_n - m_0 s_0) - (m_n s_n - m_n s_0)|\ dy \leq$$

$$||f_n - f_0||_1 + \sum_{\mathcal{X}_a}\sum_{\mathcal{X}_b}\int_{\mathcal{Y}} m_n|s_n - s_0|\ dy = ||f_n - f_0||_1 + ||s_n - s_0||_1 \leq 2||f_n - f_0||_1 \rightarrow_{P_{f_0}} 0$$

where $\mathcal{X}^a$ and $\mathcal{X}^b$ are the subspaces of $\mathcal{X}$ corresponding to the elements x_a and x_b of x. The conditional distributions from which I draw in the imputations are therefore consistent.

Appendix 3: Asymptotic Unbiasedness of the Multiple Imputation Estimator

By Assumption 4, the complete-data estimator $\hat{\beta}(Z, X, Y)$, which is not feasible, admits an influence function form, i.e.

$$\hat{\beta}(Z, X, Y) - \beta = \frac{1}{N} \sum_{i=1}^{N} T(z_i, x_i, y_i) + o_p\left(\frac{1}{\sqrt{N}}\right)$$

Without loss of generality, reorder the data so that the first $N - N^*$ observations are those that are imputed and the final N^* are those that are fully observed. Then for the m-th imputed dataset,

$$\hat{\beta}(Z^m, X, Y) - \beta = \frac{1}{N} \sum_{i=1}^{N} T(z_i^m, x_i, y_i) + o_p\left(\frac{1}{\sqrt{N}}\right)$$

and the overall multiple imputation estimator will be

$$\hat{\beta}^{MI}(Z^+, \tilde{Z}, X, Y) - \beta = \frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{N} \sum_{i=1}^{N} T(z_i^m, x_i, y_i)\right] + o_p\left(\frac{1}{\sqrt{N}}\right)$$

The difference between the expansion of $\hat{\beta}^{MI}(Z^+, \tilde{Z}, X, Y)$ and $\hat{\beta}(Z, X, Y)$ is

$$\Delta^m := \hat{\beta}^{MI}(Z^+, \tilde{Z}, X, Y) - \hat{\beta}(Z, X, Y) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N-N^*} [T(z_i^m, x_i, y_i) - T(z_i, x_i, y_i)] + o_p\left(\frac{1}{\sqrt{N}}\right)$$

It is a well-known property of influence functions that $\mathbb{E}(T(z_i, x_i, y_i)) = 0$, so it suffices to show that $\mathbb{E}(T(z_i^m, x_i, y_i)) \to 0$. This expectation can be written,

$$\mathbb{E}[T(z_i^m, x_i, y_i)] = \mathbb{E}[\mathbb{E}[T(z_i^m, x_i, y_i) \mid Z^+, \tilde{Z}, X, Y]]$$

The internal expectation is equal to

$$\int_{\tilde{Z}\times X\times Y} \sum_{Z} T(z_i, x_i, y_i) g_n(z_i \mid \tilde{z}_i, x_i, y_i)\ dP(\tilde{Z}, X, Y)$$

where $P(\tilde{Z}, X, Y)$ is the true marginal probability distribution of $\tilde{Z}$, X, and Y, and g_n is the estimate of the conditional probability distribution $g_{Z|\tilde{Z},X,Y}$ from the imputation model, shown to be L_1 consistent in the previous appendix. I proceed by defining a function $G(x_i, y_i)$ that dominates the integrand,

$$\sum_{Z} T(z_i, x_i, y_i) g_n(z_i \mid \tilde{z}_i, x_i, y_i) \leq \max_j |T(z_j, x_i, y_i)| := G(x_i, y_i)$$

with the inequality a result of g_n being a probability mass function.

$$G^2(x_i, y_i) = \max_j T^2(z_j, x_i, y_i) \leq \sum_j T^2(z_j, x_i, y_i)$$

By assumption, the variance of $T(z, x, y)$ is finite, and given the discreteness of z, this implies the expectation of $G(x_i, y_i)$ must also be finite.

$$\mathbb{E}[G^2(x_i, y_i)] \leq \mathbb{E}\left[\sum_j T^2(z_j, x_i, y_i)\right] < \infty \implies \mathbb{E}(G(x_i, y_i)) = C < \infty$$

The existence of a function $G(x_i, y_i) \geq \sum_Z T(z_i, x_i, u_i) g_n(z_i \mid \tilde{z}_i, x_i, y_i)$ for any $n \in \mathbb{N}$ and which has a finite expectation implies that $\sum_Z T(z_i, x_i, y_i) g_n(z_i \mid \tilde{z}_i, x_i, y_i)$ is uniformly integrable. Uniform integrability combined with convergence in probability implies that $\mathbb{E}[T(z_i^m, x_i, y_i)] \to \mathbb{E}[T(z_i, x_i, y_i)]$, which further implies that

$$\mathbb{E}[\Delta^m \mid Z^+, X, Y] \to 0$$

In other words, the multiple imputation estimator converges to the same value as the complete data estimator.

References

Dennis J. Aigner. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1):49–59, 1973. doi: 10.1016/0304-4076(73)90005-5.

Olanrewaju Akande, Fan Li, and Jerome Reiter. An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71(2):162–170, 2017. doi: 10.1080/00031305.2016.1277158.

Jennifer Alix-Garcia and Daniel L. Millimet. Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation. *Journal of the Association of Environmental and Resource Economists*, 10(5):1335–1367, 2023. doi: 10.1086/723723.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

John Barnard and Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999. doi: 10.1093/biomet/86.4.948.

Erich Battistin, Michele De Nadai, and Barbara Sianesi. Misreported schooling, multiple measures and returns to educational qualifications. *Journal of Econometrics*, 181(2):136–150, 2014. doi: 10.1016/j.jeconom.2014.03.002.

Lane F. Burgette and Jerome P. Reiter. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076, 2010. doi: 10.1093/aje/kwq260.

Rossella Calvi, Arthur Lewbel, and Denni Tommasi. LATE with missing or mismeasured treatment. *Journal of Business & Economic Statistics*, 40(4):1701–1717, 2022. doi: 10.1080/07350015.2021.1970573.

David Card. The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica*, 64(4):957–979, 1996. doi: 10.2307/2171852.

Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *Review of Economic Studies*, 72(2):343–366, 2005. doi: 10.1111/j.1467-937X.2005.00335.x.

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel. A note on the closed-form identification of regression models with a mismeasured binary regressor. *Statistics & Probability Letters*, 78(12):1473–1479, 2008a. doi: 10.1016/j.spl.2007.12.024.

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel. Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments. *Economics Letters*, 100(3):381–384, 2008b. doi: 10.1016/j.econlet.2008.03.015.

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel. Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statistica Sinica*, 19(3):948–968, 2009.

Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(3):449–460, 2000. doi: 10.1111/1467-9868.00243.

Jef de Slegte, Filip Van Droogenbroeck, Bram Spruyt, Sam Verboven, and Vincent Ginis. The use of machine learning methods in political science: An in-depth literature review. *Political Studies Review*, 23(3):764–784, 2024. doi: 10.1177/14789299241265084.

Melissa Dell. Deep learning for economists. *Journal of Economic Literature*, 63(1):5–58, 2025. doi: 10.1257/jel.20241733.

Francis J. DiTraglia and Camilo García-Jimeno. Identifying the effect of a mis-classified, binary, endogenous regressor. *Journal of Econometrics*, 209(2):376–390, 2019. doi: 10.1016/j.jeconom.2019.01.007.

Dave Donaldson and Adam Storeygard. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198, 2016. doi: 10.1257/jep.30.4.171.

David B. Dunson and Chuanhua Xing. A nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009. doi: 10.1198/jasa.2009.tm08439.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using Large Language Model annotations for the social sciences: A general framework for using predicted variables in downstream analyses. *Working Paper*, 2024.

Vicky Espinoza, Lorenzo Ade Booth, and Joshua H. Viers. Land use misclassification results in water use, economic value, and ghg emission discrepancies in california's high-intensity agriculture region. *Sustainability*, 15(8), 2023. doi: 10.3390/su15086829.

Luisa Turrin Fernholz. *von Mises Calculus for Statistical Functionals*. Springer-Verlag, 1983.

Christian Fong and Matthew Tyler. Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484, 2021. doi: 10.1017/pan.2020.38.

Harley Frazis and Mark A. Loewenstein. Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables. *Journal of Econometrics*, 117(1):151–178, 2003. doi: 10.1016/S0304-4076(03)00121-0.

Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics*, 35(2):697–723, 2007. doi: 10.1214/009053606000001271.

Nick Hagerty. Transaction costs and the gains from trade in water markets. *Working Paper*, 2024.

J. A. Hausman, Jason Abrevaya, and J. M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239–269, 1998. doi: 10.1016/S0304-4076(98)00015-3.

Yingyao Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008. doi: 10.1016/j.jeconom.2007.12.001.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001. doi: 10.1198/016214501750332758.

Meha Jain. The benefits and pitfalls of using satellite data for causal inference. *Review of Environmental Economics and Policy*, 14(1):157–169, 2020. doi: 10.1093/reep/rez023.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *Working Paper*, 2025.

Hiroyuki Kasahara and Katsumi Shimotsu. Identification of regression models with a misclassified and endogenous binary regressor. *Econometric Theory*, 38(6):1117–1139, 2022. doi: 10.1017/S0266466621000451.

Michael P. Keane and Robert M. Sauer. Classification error in dynamic discrete choice models: Implications for female labor supply behavior. *Econometrica*, 77(3):975–991, 2009. doi: 10.3982/ECTA7642.

Hang J. Kim, Juhee Lee, Young-Min Kim, and Jared Murray. *synMicrodata: Synthetic Microdata Generator*, 2024. URL https://CRAN.R-project.org/package=synMicrodata. R package version 2.1.0.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2):837–877, 2006. doi: 10.1214/009053606000000029.

Dan M. Kluger, Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates. Prediction-powered inference with imputed covariates and nonuniform sampling. *Working Paper*, 2025.

Brent Kreider. Regression coefficient identification decay in the presence of infrequent classification errors. *Review of Economics and Statistics*, 92(4):1017–1023, 2010. doi: 10.1162/REST_a_00044.

Arthur Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551, 2007. doi: 10.1111/j.1468-0262.2006.00756.x.

Kerri Lu, Dan M. Kluger, Stephen Bates, and Sherrie Wang. Regression coefficient estimation from remote sensing maps. *Remote Sensing of Environment*, 330, 2025. doi: 10.1016/j.rse.2025.114949.

Aprajit Mahajan. Identification and estimation of regression models with misclassification. *Econometrica*, 74(3):631–665, 2006. doi: 10.1111/j.1468-0262.2006.00677.x.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

Bruce D. Meyer and Nikolas Mittag. Misclassification in binary choice models. *Journal of Econometrics*, 200(2):295–311, 2017. doi: 10.1016/j.jeconom.2017.06.012.

Joel E. Michalek and Ram C. Tripathi. The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *Journal of the American Statistical Association*, 75(371):713–721, 1980. doi: 10.2307/2287675.

Jonas M. Mikhaeil, Andrew Gelman, and Philip Greengard. Hierarchical Bayesian models to mitigate systematic disparities in prediction with proxy outcomes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, forthcoming. doi: 10.1093/jrsssa/qnae142.

Daniel L. Millimet. The elephant in the corner: A cautionary tale about measurement error in treatment effects models. *IZA Discussion Papers*, 5140, 2010.

Jared S. Murray. Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159, 2018. doi: 10.1214/18-STS644.

Jared S. Murray and Jerome P. Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016. doi: 10.1080/01621459.2016.1174132.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. 1994. doi: 10.1016/S1573-4412(05)80005-4.

Pierre Nguimkeu, Augustine Denteh, and Rusty Tchernis. On the estimation of treatment effects with endogenous misreporting. *Journal of Econometrics*, 208(2):487–506, 2019. doi: 10.1016/j.jeconom.2018.10.005.

Margaret Sullivan Pepe and Thomas R. Fleming. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413):108–113, 1991. doi: 10.1080/01621459.1991.10475009.

Jonathan Proctor, Tamma Carleton, and Sandy Sum. Parameter recovery using remotely sensed variables. *NBER Working Paper No. 30861*, 2023.

Ashesh Rambachan, Rahul Singh, and Davide Viviano. Program evaluation with remotely sensed outcomes. *Working Paper*, 2025.

Charles Regnacq, Ariel Dinar, and Ellen Hanak. The gravity of water: Water trade frictions in California. *American Journal of Agricultural Economics*, 98(5):1273–1294, 2016. doi: 10.1093/ajae/aaw051.

Jerome P. Reiter and Trivellore E. Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471, 2007. doi: 10.1198/016214507000000932.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996. doi: 10.2307/2291635.

Luke C. Sanford, Megan Ayers, Matthew Gordon, and Eliana Stone. Adversarial debiasing for unbiased parameter recovery. *Working Paper*, 2025.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2): 639–650, 1994.

Yajuan Si and Jerome P. Reiter. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521, 2013. doi: 10.3102/1076998613480394.

Adrian L. Torchiana, Ed Rosenbaum, Paul T. Scott, and Eduardo Souza-Rodrigues. Improving estimates of transitions from satellite data: A hidden Markov model approach. *Review of Economics and Statistics*, 107(2):426–441, 2025. doi: 10.1162/rest_a_01301.

USDA NASS. Cropland data layers FAQs. 2024. URL https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php.

Tino Werner. Asymptotic linear expansion of regularized M-estimators. *Annals of the Institute of Statistical Mathematics*, 74:167–194, 2021. doi: 10.1007/s10463-021-00792-5.

Yuefang Wu and Subhashis Ghosal. The L_1-consistency of Dirichlet mixturies in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 10(101):2411–2419, 2010. doi: 10.1016/j.jmva.2010.06.012.

David Wuepper, Wyclife Agumba Oluoch, and Hadi Hadi. Satellite data in agricultural and environmental economics: Theory and practice. *Agricultural Economics*, 56(3):493–511, 2025. doi: 10.1111/agec.70006.

Takahide Yanagi. Inference on local average treatment effects for misclassified treatment. *Econometric Reviews*, 38(8):938–960, 2019. doi: 10.1080/07474938.2018.1485833.