Remotely Incorrect?
Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-Garcia*

Oregon State University

Daniel L. Millimet†

Southern Methodist University & IZA

November 25, 2022

Abstract

Research relying on remotely sensed data on land use and deforestation has exploded in recent years. While satellite-based measures have clear advantages in terms of coverage, the presence of measurement error within these products is often overlooked. Here, we detail the econometric implications of these errors when analyzing the determinants of binary measures of deforestation or forest cover. We then discuss estimators that exploit knowledge of the remote sensing process to obtain consistent estimates. Finally, we assess our estimators via simulation and an impact evaluation of a conservation program in Mexico. We find that both geography and characteristics of the raw data can lead to systematic under-reporting of deforestation. However, accounting for these sources of error, which are common across many satellite-based metrics, can limit the bias from misclassification.

JEL: C18, C21, Q23, Q28

Keywords: Deforestation, Satellite, Remote Sensing, Measurement Error, Misclassification

*Corresponding author. The authors are grateful for helpful comments from Chris Barrett, Xiang Bi, John Gibson, Pedro Sant'Anna, Laura Schechter, and seminar participants at the OARES virtual seminar series, University of Delaware Economics Department, UC Berkeley Department of Agricultural and Resource Economics, the Triangle Resource and Environmental Economics Seminar, the Center for Environmental Economics – Montpellier, the Aix Marseille School of Economics, and the Economics Department at Southern Methodist University. We also thank conference attendees at the Econometric Society European Winter Meetings, International Association of Applied Econometrics conference, NBER EEE Summer Meetings, Western Economic Association International conference, and TWEEDS conference. E-mail: jennifer.alix-garcia@oregonstate.edu.

†E-mail: millimet@smu.edu.

1 Introduction

Deforestation is a persistent challenge in low and middle income countries, where governments struggle to balance the twin goals of poverty alleviation and greenhouse gas reduction (Sims and Alix-Garcia, 2017). The IPCC calculates emissions from agriculture, forestry, and land use to be just under one-quarter of total anthropogenic greenhouse gas emissions (IPCC, 2019). Further, deforestation is associated with mass extinction and the loss of watershed function and other essential ecosystem services (e.g., Alston et al., 2013). As a result, indicators of forest cover and deforestation are essential inputs for carbon accounting and the parameterization of climate, biodiversity, and hydrological models (Hansen et al., 2010). Moreover, such data allow for investigations into the behavioral and policy determinants of deforestation and land use more generally.

Historically, data on land use have been obtained from ground surveys. However, this has changed over time. There are now over 2,000 satellites orbiting Earth. In combination with the growth in computer processing power, this has driven an explosion in the availability of data on land use – as well as other environmental attributes – derived from remote sensing methods. This new type of information has been fully embraced by researchers, especially as ease of access has improved. For example, the Global Forest Change data set detailed in Hansen et al. (2013) has more than 7,900 citations, at least 34 of which are in economics journals.

Aside from ease of access, remotely sensed data on land use offers other advantages. First, satellite data can offer global coverage. Second, whereas survey data suffers from many sources of error, such as enumeration errors and response biases, these are necessarily eliminated through satellite collection. That said, absence of familiar types of data error does not imply the absence of all error. Errors in remotely sensed measurements are likely present. As many remotely sensed measures of forest cover or deforestation are binary at their most disaggregate level, classification errors are nonclassical in that they are negatively correlated with the truth (Black et al., 2000). Binary outcomes are often used in land use change studies conducting the analysis at the unit of the remote sensing product (the "pixel") or across a random set of points in space (Andam et al., 2008; Ferraro et al., 2013; Robalino and Pfaff, 2013; Robalino et al., 2017). Furthermore, continuous measures of land use (e.g., percent forest cover or deforested) are typically obtained by aggregating

binary data over a large area. If the underlying binary data suffers from nonclassical misclassification error, then the aggregate, continuous measure will as well.[1] Thus, using continuous measures of forest cover or deforestation does not convert the problem to one of classical measurement error (i.e., mean zero and idiosyncratic). Examples of analysis of land use change publications using aggregations across binary measures include any paper using Hansen et al. (2013), such as Abman and Lundberg (2020) and Salemi (2021). Other deforestation analyses recognize the censoring in the land use change data and use estimators that recognize both the decision to deforest (binary) and the amount deforested (continuous) (Alix-Garcia et al., 2012, 2013).

With this in mind, this paper contributes to our understanding of how to address measurement error in satellite data by demonstrating the existence of misclassification in the data, extending existing estimators to the satellite data context, and applying these estimators to the evaluation of an existing forest conservation program. In Section 2.2 we use two satellite-based measures of forest cover for Mexico near the same time period and based upon imagery from the same type of sensor to document large differences across data sources. While we do not take a stance on which, if either, of these data sources is accurate, it is crucial to know that both are reputable data sources and a researcher investigating forest cover in Mexico could easily use either one without facing scrutiny. The two binary measures for the presence of any forest diverge for roughly 18% of the sample. Moreover, differences in the two measures are correlated with environmental (e.g., slope, elevation, and biome) and sensor attributes.

In Section 3 we review the econometric implications of misclassification when assessing the determinants of a remotely sensed binary measure of deforestation using several binary choice estimators common in the land use literature. We then discuss alternative estimators, representing extensions of the misclassification binary choice model proposed in Hausman et al. (1998). In particular, we consider two extensions. First, we allow for the misclassification rates to depend on covariates as in Lewbel (2000). Here, the covariates capture environmental attributes affecting the accuracy of satellite classifications. Second, we use the scobit family of binary choice models, which nests the logit model as a special case (Nagler, 1994).[2] The scobit introduces an additional

[1] This issue does not arise, however, with continuous measures such as NDVI or EVI.

[2] New Stata commands, `mclogit` and `mcre`, are available at `http://faculty.smu.edu/millimet/code.html`.

shape parameter into the link function. This additional flexibility has proven useful when the outcome is of the rare-events type (Goleţ, 2014), which is often the case for deforestation. Finally, we describe the process by which satellite data moves from sensors to usable data points, detailing the various stages at which misclassification may be introduced. Understanding this process is vital for specifying which covariates should be used to model the misclassification rates in the Hausman et al. (1998) approach.

Finally, in Section 4 we investigate the practical performance of the estimators considered. Specifically, we re-visit the impact of a program of payments for ecosystem services on deforestation in Mexico over the period 2003-2015. The panel data we use here is different from the cross-sectional data we use in Section 2.2. Now, we have only a single measure of *deforestation* to examine. However, the same geographic correlates that affect measurement of the *level* of forest coverage also affect detection of *change* in forest cover. Prior to analyzing the data, we undertake a (limited) Monte Carlo study designed to mimic the panel data. The simulations lead to three primary conclusions. First, ignoring misclassification introduces significant bias. Second, current approaches in the literature designed to deal with misclassification are done in vain; the bias remains. Third, our extensions of the Hausman et al. (1998) estimator perform quite well. In particular, the misclassification logit model is preferred with non-rare events data. With rare events data, the misclassification scobit is preferred. This is salient as deforestation is a rare event. Between 1990 and 2015, 219 million hectares of forest were lost (FAO, 2016); an annual rate of change of 0.13 percent. This implies that only 1.3 out of every 1,000 hectares will have observable forest change in random samples used to examine annual deforestation behavior.

In our application, we also obtain three main findings. First, the satellite-based measure that we use under-reports the true extent of deforestation. In our preferred specification, we find that 15% of all instances of deforestation are missed, but that the false positive rate is essentially zero. Overall, we find about 12% of the observed reports are misclassified. Topography and the availability of images are important determinants of misclassification. Second, in light of the finding of no false positives, we also consider the estimator proposed in Nguimkeu et al. (2019) for comparison. The results are quite comparable; around 21% of the observed reports are estimated to be misclassified. Third, ignoring misclassification can result in bias of the average marginal effects. In particular,

our preferred estimator suggests that the conservation program we examine reduces the probability of deforestation by 0.7 percentage points. In comparison, the estimator most frequently used by researchers currently – what we refer to as the *ad hoc* fixed effects linear probability model – produces an estimate that is attenuated by roughly one-third and not statistically different from zero at conventional levels. The average marginal effects are also biased for the other covariates in the model, especially those that are determinants of misclassification.

In sum, our analysis leads to several recommendations for researchers interested in using remotely sensed data, particularly related to forest cover and deforestation. Most importantly, researchers ought to engage with remote sensing scientists to understand how the data are constructed and the nature of its limitations. Important topics of discussion should be how accuracy of the data might change with geography or across time. It is likely that similar geographic, weather-related, and technical sensor issues affect other remotely sensed outcomes, such as crop classifications or vigor, population measures, and pollution metrics from optical sensors. In these cases, if outcomes are binary, as in the example of deforestation or other land use classifications, estimators based on Hausman et al. (1998), incorporating institutional knowledge on the sources of error, offer a potential improvement over current practices, even in the case of rare events data.

Our analysis also points to several avenues in need of future research. First, we do not consider solutions that might exploit the presence of two error-laden binary (or continuous) measures. While such data may give rise to other estimation methods (e.g., Schennach and Hu, 2013), multiple land use measures are typically unavailable to the researcher. This is the case in our application as well. Second, we do not consider estimators that might exploit the spatial nature of the data to overcome misclassification. Third, we do not consider nonclassical measurement error in continuous, remotely sensed measures of forest cover or change.[3] Finally, while we can speculate, we cannot say how readily our insights generalize to analysis of other remotely sensed phenomena such as nighttime lights, pollution, urbanization rates, population measures, etc.

Despite leaving these issues for future research, our current study contributes to two important literatures. First, we add to the now large number of papers using satellite-based measures of

[3]The binary choice models we consider that address misclassification exploit nonlinearity of the link function for identification. Because such nonlinearity is absent in regression models with continuous outcomes, additional information (such as exclusion restrictions) are likely needed to obtain consistent estimates.

various concepts to explore economic and other research questions. Moreover, while there are at least two reviews focusing on the use of these measures in economics generally (Donaldson and Storeygard, 2016; Jain, 2020), we are among a few papers to document potential sources of mismeasurement as well as investigate possible solutions. Gibson et al. (2021) and Gibson (2020) are two exceptions. Both papers examine nighttime lights data. Torchiana et al. (2020) is a third exception, and examines land use outcomes. The authors applies a hidden Markov model to correct the underlying data prior to estimation in the case where the data is measuring transitions. Finally, Fowlie et al. (2019) highlight the importance of prediction error in satellite-based air-quality estimates, and Michler et al. (2022) show that intentional spatial displacement of household survey observations does not affect estimation of agricultural productivity as a function of satellite based weather measures. Second, while there is a long and rich literature on overcoming measurement error in regression models, ours is the first paper to consider an extended version of the estimator in Hausman et al. (1998) and Lewbel (2000), as well as Nguimkeu et al. (2019), that allows the misclassification rates to depend on covariates applied to satellite data. We are also the first to propose combining misclassification with a scobit model to address misclassification in rare events data.

2 Misclassification in remotely sensed data

2.1 Inside the black box

Understanding the sources of misclassification in remotely sensed land use data requires one to understand how the data, which could be static (e.g., forest cover) or dynamic (e.g., deforestation), are obtained. To begin, each satellite has different technical specifications, including sensor type (e.g., optical, thermal, or radar), frequency of reporting, and spatial resolution (Union of Concerned Scientists, 2020). While many of the steps that we describe here generalize to other types of technology, we focus on optical sensors which are frequently used to produce data on forest cover and land use change. Optical sensors measure reflected energy, and come to the analyst as measurements of different "spectral bands" arranged in a grid (Kennedy et al., 2009). For optically-derived

information, the process for classifying these images into usable data entails (i) accessing images from their storage place in an archive, (ii) pre-processing the images so that they can be entered into an image-classification system (manual, automated, or a hybrid), and (iii) setting rules for translating the spatial and temporal trends in the images into static or dynamic numerical data.

This assembly-line of tasks creates three broad categories of potential errors: errors due to technical limitations of the sensors themselves, those introduced in the pre-processing of images, and errors in the algorithms used to translate reflectancies into usable data. Technical limitations can induce obvious challenges. For example, the image might originate from a satellite with a spatial resolution of one kilometer (100 hectares), while the behavior of interest may operate at a scale of one hectare or less. Another example of a technical limitation arises with the "scan line error" of the Landsat 7 satellite (see Figure A1 in Appendix A). This error leaves swaths of the imagery blank. The missing swaths are then imputed by either mosaicking (stitching together) multiple images from different time periods or directly imputing the missing imagery using predictions based on available data. These sources of error are arguably random conditional on the true measure of the outcome of interest.

However, in addition to technical limitations, the raw images are frequently distorted due to solar, atmospheric, or topographic features (Young et al., 2017). These distortions are ameliorated by pre-processing of the raw images. While these corrections are necessary, they are not infallible (Kennedy et al., 2009). Another source of distortion error arises from the timing of images which may be affected by reduced visibility due to cloud cover. Such disruptions in timing may cause, say, instances of deforestation to appear in the data with a lag.

After collecting and pre-processing the raw images, these (now processed) images are translated into numerical data, such as the presence of forest, using an algorithm. There are numerous ways to conduct this translation. For smaller areas, classification by visual inspection is often possible. For larger areas, machine learning methods based on pixel-by-pixel approaches and others, known as "object-based" approaches, that use broader spatial dimensions are typically employed (Li et al., 2014). The former are currently more common, and these methods can be divided into two further groups: supervised and unsupervised. Supervised classification involves using information from representative sites where information on the ground is known, and then leveraging this information

to establish decision rules for classification of associated pixels. Unsupervised classification divides remote sensing images into classes based on clustering of image values, without substantial use of secondary data sources. Both of these approaches classify each pixel with a single value. Other methods recognize potential heterogeneity within pixels and classify each pixel based on proportions across multiple categories. Newer object-based classifiers segment images into objects (groups of pixels), and these segments provide the unit of classification. Recent approaches also exploit the geographic information of adjacent pixels (e.g., textural analysis) to aid with classification (Li et al., 2014). This process can be used to measure both the state of land cover on the ground as well as changes. For example, classification strategies to uncover deforestation exploit temporal discontinuities in reflectancy values across images to illuminate changes in the state of events on the ground. Regardless, these sources of error likely depend on the true land use measure as well as the quality of the raw images, which depend on cloud cover and topographic features.

As should be clear, the processes of translating initial images into usable data are complex, involve much uncertainty, and are largely algorithmic. As such, accuracy is not assured and can vary across algorithms, as well as across space and time for a given algorithm, depending upon the underlying characteristics of the objects being classified.

2.2 Example: misclassification in forest cover data

This section uses two different, remotely sensed *cross-sectional* data sets on forest cover in Mexico to demonstrate the presence of misclassification error.[4] The two data sets give rise to two measures of forest cover based on similar imagery taken at nearly the same time. However, the two measures are derived from different pre-processing and classification techniques. Importantly, our aim is not to establish which is more accurate, nor to understand where along the assembly line of the data production any differences may materialize. Our objective is to show that – despite both datasets being publicly available and reasonable resources for empirical researchers – the two data sources are classifying the same landscape in demonstrably different ways. Moreover, differences between

[4]Note, this is *not* the data we use in the application in Section 4 as it is not a panel and it contains information on forest cover rather than deforestation. Our interest here is to use these two cross-sectional measures of forest cover to illustrate the presence of misclassification in satellite data and its correlation with features of the environment. In our application, we only have a single data source. As such, we do not consider econometrics solutions to the misclassification problem that require two measures.

the data sources correlate with environmental and sensor attributes.

The first data source is the *Land Use and Vegetation, Series V* (henceforth, GOM).[5] It is part of a series of land cover maps that has been produced periodically by the Mexican government since 1985. The GOM product that we use exploits 2011 images from the Landsat 5 satellite (Government of Mexico, 2014). The Landsat satellites have a resolution of 30 meters. The GOM data classifies forest by type using supervised classification supported by ground-truthing in the field. The data come in what is called "vector" form, which is a series of polygons defined as homogeneous classes rather than data on individually interpretable pixels. There are 59 land use classes in the original data. For our purposes, we reclassify these land use categories into a binary indicator for forest or non-forest. Although the underlying images have a 30 meter resolution, the minimum mappable unit for the analysis (the smallest size that determines whether a feature is captured) is 50 hectares. This dataset has been used in a number of studies of land use change in Mexico, including Aguilar-Tomasini et al. (2020) and Lorenzen et al. (2021).

The second data source is the University of Maryland's *Global Land Cover and Deforestation* data set (henceforth, the Hansen data) (Hansen et al., 2013). We use Hansen's binary classification for forest or non-forest from 2010, which is based upon Landsat 7 imagery. This data is available in "raster" form (in contrast to the vector above), which means that the information comes in a grid of 30 meter pixels, rather than as polygons. A pixel is classified as forested when its canopy cover measure exceeds 50%, a common cutoff.[6] The Hansen data classifies pixels using supervised classification supported by higher resolution imagery as well as previous tree cover layers derived from both Landsat and lower resolution imagery (Hansen et al., 2013). As mentioned in the introduction, this dataset is widely exploited by researchers.

In light of this, the GOM and Hansen measures may differ due to the reclassification of the GOM product and the differences in scale across the two datasets. In addition, there are small differences between Landsat 5 and Landsat 7 (there is no existing Landsat 6). Both have the same spatial

[5]The data are publicly available (Government of Mexico, 2014). See `http://www.conabio.gob.mx/informacion/metadata/gis/usv250s5ugw.xml?_httpcache=yes&_xsl=/db/metadata/xsl/fgdc_html.xsl&_indent=no` for the dataset and `https://www.inegi.org.mx/contenidos/temas/mapas/usosuelo/metadatos/guia_interusosuelov.pdf` for documentation.

[6]It is not uncommon to use canopy cover cutoffs as low as 10% to define forest cover and the Hansen data offers a number of possible cutoff points.

resolution (30 m) and image size (approximately 170 x 183 km), but Landsat 7 has an additional spectral band (U.S. Geological Service, 2021). Finally, differences between the two data sources may be attributable to the different time periods: 2010 versus 2011. However, it seems unlikely that this final difference is much of a factor since the GOM data (from 2011) report significantly more forest cover than the Hansen data (from 2010) and it is unlikely that new forest growth over such a short time span could explain the differences.

To compare the two datasets, we extract the information within a 5 x 5 km grid laid across the contiguous land area of Mexico.[7] This aggregation yields the proportion of forest cover within each (5 x 5 km) cell. We then generate binary indicators of any forest cover, defined using a threshold of 50 ha (0.10 of a cell) for both datasets (based on the minimum threshold for the GOM data). Finally, we measure several attributes of each cell, such as elevation, slope, and forest type. To examine the role of satellite image availability in driving differences in classification, we also include counts of the number of Landsat 5 and Landsat 7 images with less than 25 percent cloud cover available in 2010 and 2011. A greater number of cloud-free images increases the amount of information available to the remote sensor, and is likely to improve the accuracy of final classifications. Figure B1 in Appendix B shows the distribution of these images across Mexico in 2010.

We provide a brief description of our analysis in the interest of brevity. Complete results are provided in Appendix B. Tables B1 and B2 report summary statistics and cross-tabulations, respectively. The Hansen data reports a lower fraction of cells with any forest coverage; the difference is about 15 percentage points. However, the disagreements are not uni-directional. While the two data sources agree 78% of the time, the Hansen data detects some level of forest while the GOM data does not in about 4% of cells. The reverse occurs in 18% of the cells.[8]

Finally, we assess the relationship between the differences in the two data sources and geographic

[7]We engage in this aggregation because it makes the dataset more manageable, and because some aggregation choice had to be made to make the vector (GOM) dataset comparable to the raster (Hansen) dataset. The process of aggregating across space is both necessary and common in the use of satellite imagery; the terrestrial area of the earth requires around 400 billion Landsat pixels to cover it (NASA, 2021). Furthermore, the classification of a single pixel into a given land cover is, in fact, a mini process of aggregation, where land use categories are determined by different spectral thresholds.

[8]Figure B2 shows how the proportion of cells where there is disagreement in classification changes as we apply different cutoff levels to each dataset. The lowest level of classification disagreement occurs with a cutoff of one for both datasets. Divergence is also low when cutoffs for both datasets are quite low (zero for Hansen and less than 0.20 for GOM). In general, divergence is larger with medium-sized cutoffs and smaller on the ends of the distribution.

characteristics. Table B3 reports the standardized beta coefficients from regressions of the absolute value of the divergences, as well as the sum of the two measures, on environmental attributes of the cell and the availability of cloud-free images using a variety of symmetric cutoffs for forest. Thus, the coefficients may be interpreted as the effect of a one standard deviation increase in each covariate on the absolute difference in misclassification errors and combined signal in the two measures, respectively.[9]

The results confirm that the absolute differences are correlated, both statistically and economically, with almost all of the covariates. In particular, the differences between data sources are pronounced where the topography is extreme, as measured with high elevation and slope. It is also the case that the coefficients on the different forest biomes are all positive relative to the omitted category, grasslands and agriculture. The beta coefficients are particularly large for the pine-oak and dry tropical biomes. This is consistent with misclassification that enters during pre-processing – tropical areas tend to have more clouds – and through the classification algorithms used to define forest cover – dry tropical forest tends to be harder to observe during certain seasons because it loses leaves. The disagreements between the two datasets are decreasing in the minimum number of cloud-free images available for either Landsat 7 in 2010 (the basis of the Hansen classification) and Landsat 5 in 2011 (the basis of the GOM data). This suggests that greater image availability may improve agreement between the two datasets. The interaction between image availability and slope is positive, suggesting for the same number of images, higher slope is associated with greater differences in classification. For the 0.10 (50 ha) forest threshold cutoff, at the mean value of slope in the data (8.8), the positive effect of slope on disagreement across the two datasets overwhelms the palliative effect of greater image availability.[10] The results also confirm that the sum of the two outcomes is positively related to the forest biomes – indicating greater forest signal in those areas. The number of cloud-free scenes is negatively correlated with the signal, which may be due to either the broad scale spatial effects mentioned above or the fact that there is more forest in tropical areas

[9] In the absence of sufficiently strong positive covariance between the measurement errors, the reliability ratio of the sum of the two measures will exceed the reliability ratio for either single measure.

[10] We observe that as we increase the threshold towards one, geographic characteristics (slope, elevation) become more important relative to image availability in predicting differences in classification. The effect of ecosystem type on the sum of the indicators is larger for low thresholds, and the impact of image availability is highest for mid-level cutoffs and generally negatively affected by its interaction with slope.

that tend to be cloudy. The interaction between scenes and slope tends to reduce the signal.

3 Empirical methodology

Having established the sources and existence of misclassification error, we now demonstrate how this error affects estimation, and propose a solution. This section first clarifies the data-generating process and appropriate estimators in the absence of misclassification, then shows how misclassification affects the coefficients of interest, and ends by proposing a solution.

3.1 Setup

Our objective is to assess the determinants of a binary measure of deforestation using panel data derived from remotely sensed data. Let y^*_{it} denote the true outcome for location i at time t, where $y^*_{it} \in \{0, 1\}$. The data-generating process (DGP) for y^* is given by

$$\Pr(y^*_{it} = 1 | x_{it}, \omega_i) = F(x_{it}\beta + \omega_i), \tag{1}$$

where x_{it} is a vector of correctly measured, exogenous covariates, ω_i is a location-specific fixed effect (FE), and $F(\cdot)$ is the link function. If $F(\cdot)$ is the identity link function, then (1) is a linear probability model (LPM). If $F(\cdot)$ is the standard normal CDF or the logistic CDF, then (1) is the usual probit or logit model, respectively.

A few comments are warranted. First, location FEs are easily accommodated in the LPM by mean-differencing. We refer to this estimator hereafter as FE-LPM. However, the remaining models are estimated via Maximum Likelihood (ML). In this case, FEs lead to the well-known incidental parameters problem (Lancaster, 2000).[11] A common solution is to assume a correlated random effects (CRE) structure. The CRE structure directly models the dependence between the FEs and

[11] For the logit model, using the conditional likelihood function, where the conditioning is done on the $\sum_t y_{it}$, circumvents the incidental parameters problem. Nonetheless, it is not an ideal solution since the marginal effects cannot be computed without additional assumptions on the FEs.

the location-specific covariates. Specifically, we assume

$$E[\omega_i|x_i] = \overline{x}_i\gamma, \tag{2}$$

where x_i is a vector of location-specific covariates across all time periods and $\overline{x}_i$ is a vector of location-specific means of the covariates. In error form, we have

$$\omega_i = \overline{x}_i\gamma + \eta_i, \tag{3}$$

where η_i is now a location-specific random effect. Substitution of (3) into (1) yields

$$\Pr(y^*_{it} = 1|x_{it}, \omega_i) = F(x_{it}\beta + \overline{x}_i\gamma + \eta_i), \tag{4}$$

which can be estimated using random effects binary choice models or traditional binary choice models with robust standard errors. Hereafter, we refer to estimators adopting this strategy as CRE estimators.

Second, it is well-known that the usual binary choice models perform very poorly when there are proportionately few occurrences of ones (or, conversely, zeros) in the data (King and Zeng, 2001). Such outcomes are referred to as rare events. It is quite possible that this may be a fair characterization of deforestation data in many applications. As mentioned above, the global average deforestation rate between 1990 and 2015 implies that only 1.3 out of every 1,000 hectares will have observable forest change in a random sample of forest. Brazil, a critical country for forest-based climate mitigation and biodiversity conservation, lost tree cover at a rate of 0.63% in 2020 (3.29MHa of forest). The Democratic Republic of Congo has one of the highest deforestation rates in the world, at 0.66% in 2020 (WRI, 2022). Applications that examine annual deforestation in these settings can expect to find only 6 or 7 out of every thousand pixels experiencing deforestation. Naturally, the rarity of these events decreases with the level of spatial or temporal aggregation, although in settings where deforestation events are highly clustered, observed deforestation might be still be rare even across larger spatial units of aggregation. In the application that we examine in Section 4, for example, 20 percent of the polygons register deforestation at any time during the sample years.

However, within the 32 states in the sample, this percentage ranges from 0.9 to 30 percent.

While several alternatives for modeling rare events data have been proposed in contexts outside of deforestation, we focus here on the scobit model (Nagler, 1994). In the scobit model, the link function is given by

$$F(\cdot) = 1 - \frac{1}{[1 + \exp(x_{it}\beta + \omega_i)]^{\alpha}}, \tag{5}$$

where α is an unknown shape parameter. The scobit model corresponds to the logit model when $\alpha = 1$. The scobit model may potentially perform better with rare events data because the link function is no longer symmetric when $\alpha \neq 1$. For example, Goleţ (2014) finds that the scobit performs very well when modeling rare corporate bankruptcies.

3.2 Misclassification

When y^* is not observed by the researcher, but rather a misclassified version, y, then all of the preceding estimators will be inconsistent. To see this in the FE-LPM, we introduce the following measurement error equation

$$y_{it} = y^*_{it} + \mu_{it}, \tag{6}$$

where $\mu_{it} \in \{-1, 0, 1\}$ is the measurement error. Since μ_{it} can only take on the values of 0 or -1 if $y^*_{it} = 1$, and can only take on the values of 0 or 1 if $y^*_{it} = 0$, then it must be that $\text{Cov}(y^*_{it}, \mu_{it}) < 0$. Since the measurement error is negatively correlated with the truth, it is also negatively correlated with the determinants of the truth. Consequently, all covariates become endogenous and the FE-LPM estimates will be biased and inconsistent unless $\beta = 0$.

Moreover, this problem does not vanish if one aggregates misclassified binary data into a continuous outcome. In Appendix C we demonstrate that this is the case, although we also show that aggregation will dampen the negative covariance between the classification errors and the true outcome. However, the reduction in this covariance is a function of the number of units over which aggregation occurs, leading to heterogeneity in the extent of covariation across aggregated units of different sizes. This type of variation across units often occurs in the analysis of deforestation when data are aggregated to producer parcels or administrative units.

Misclassification also leads to inconsistent estimates in the ML models. To see this, we introduce the following misclassification probabilities

$$\Pr(y_{it} = 1 | y_{it}^* = 0, z_{it}) = G_0(z_{it}\theta_0) \tag{7}$$

$$\Pr(y_{it} = 0 | y_{it}^* = 1, z_{it}) = G_1(z_{it}\theta_1), \tag{8}$$

where $G_0(\cdot)$ and $G_1(\cdot)$ are two new link functions, z_{it} are correctly observed covariates, and θ_0 and θ_1 are corresponding vectors of unknown parameters. Equations (7) and (8) reflect the probabilities of false positives and false negatives occurring in the data, respectively. In Hausman et al. (1998), $G_0(\cdot)$ and $G_1(\cdot)$ are each assumed to be a scalar parameter. Thus, in their model, the probability of misclassification depends only on the true value, y_{it}^*. Here, we allow for covariates to also affect the misclassification probabilities as in Lewbel (2000).

Combining (1), (3), (7), and (8), the probability of a one or zero occurring in the *observed* data is given by

$$\Pr(y_{it} = 1 | x_{it}, z_{it}, \mu_i) = G_0(z_{it}\theta_0) + [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i) \tag{9}$$

$$\Pr(y_{it} = 0 | x_{it}, z_{it}, \mu_i) = 1 - G_0(z_{it}\theta_0) - [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i). \tag{10}$$

These probabilities form the basis of the log-likelihood function, given by

$$\begin{aligned} \ln \mathscr{L} = \sum_i \sum_t \{ & y_{it} \ln \{G_0(z_{it}\theta_0) + [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i)\} \\ & + (1 - y_{it}) \ln \{1 - G_0(z_{it}\theta_0) - [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i)\}\}. \end{aligned} \tag{11}$$

Maximizing (11) will yield consistent estimates of the model parameters assuming the full DGP is correctly specified. In contrast, as shown in Hausman et al. (1998), a naïve ML model that ignores misclassification yields inconsistent estimates. While aggregation may possibly reduce the effects of classification errors depending on how the covariates are aggregated (see Appendix C), aggregation of the binary outcome to one that is continuous precludes the use of the estimator in Hausman et al. (1998) as a solution.

In our implementation of the ML estimators, we allow for the link function, $F(\cdot)$, to correspond to the scobit family. When α equals one, we refer to the model as the Misclassification CRE (MC-CRE) Logit; when α is less than one, we refer to the model as the MC-CRE Scobit. However, in all cases we use the standard normal CDF for the link functions in the misclassification probabilities, $G_0(\cdot)$ and $G_1(\cdot)$.

We also consider an additional set of estimators for comparison. Researchers aware of the fact that variables in z_{it} affect the accuracy of satellite data often choose to simply incorporate these in the model as traditional covariates. Thus, we also consider a FE-LPM and traditional logit and scobit models where the set of covariates is augmented to include z_{it}. We refer to these as *ad hoc* estimators. The Ad Hoc FE-LPM is given by

$$y_{it} = x_{it}\beta + z_{it}\theta + \omega_i + \varepsilon_{it} \tag{12}$$

and the Ad Hoc CRE Logit and Ad Hoc CRE Scobit models are based on the following probabilities

$$\Pr(y_{it} = 1|x_{it}, z_{it}, \mu_i) = F(x_{it}\beta + z_{it}\theta + \overline{x}_i\gamma_0 + \overline{z}_i\gamma_1 + \eta_i). \tag{13}$$

Some final comments are necessary. First, identifying the separate effects of covariates on the determinants of y^* and the misclassification probabilities relies on the nonlinearity of the link functions in (9) and (10). As such, if x and z have covariates in common, identification may be tenuous. Lewbel (2000) proves that the model is semiparametrically identified if x contains a continuous covariate with large support not included in z. The discussion in Section 2 suggests interactions between image availability and topography as potential candidates to include in z.

Second, in the scobit model, identification of the shape parameter, α, along with the misclassification probabilities can be difficult, resulting in challenges with convergence. Intuitively, this arises because θ_0, θ_1, and α all make use of the same variation for identification. To see this, consider a particular observation with a high value of the index, $x_{it}\beta_0 + \overline{x}_i\gamma_0$, for a given set of parameter values β_0 and γ_0, but the observed y_{it} is zero. In this case, the estimates of θ_1 can adjust to suggest a higher probability that this observation is misclassified or α can adjust such that the value of

the index is associated with a lower probability of observing an outcome of one. In the logit model allowing for misclassification, this identification concern does not arise since the shape of the link function, $F(\cdot)$, is fixed. To circumvent this issue, we treat α as unidentified and conduct a grid search by setting $\alpha = 0.20, 0.25, ..., 0.95$.[12] The log-likelihood function can then be used as a model selection tool to choose the value of α (from the set contained in the grid search) that best fits the data.[13]

Third, we follow Papke and Wooldridge (2008) and estimate the Ad Hoc and MC-CRE Logit and Scobit models using the traditional logit and scobit probabilities (i.e., ignoring the presence of the random effect, η). However, the standard errors are clustered at the unit level (or higher).

Finally, as the true functional forms for $F(\cdot)$, $G_0(\cdot)$, and $G_1(\cdot)$ are unknown, it is possible that an Ad Hoc version of the model is, in fact, correctly specified. For instance, it is possible that the true functional forms for $\Pr(y^*_{it} = 1|x_{it}, \mu_i)$ and the misclassifications rates lead to (13) with $F(\cdot)$ as the logistic CDF being the correct model. Moreover, in such a case, if x and z do not overlap, it is possible to assess whether the covariates are statistically significant determinants of y^* and misclassification, respectively. However, it would not be possible to estimate $\Pr(y^*_{it} = 1|x_{it}, \mu_i)$, $\Pr(y_{it} = 1|y^*_{it} = 0, z_{it})$, or $\Pr(y_{it} = 0|y^*_{it} = 1, z_{it})$. Nor would the researcher have any method to assess whether this is the case. Nonetheless, it is important to recognize this possibility.

4 Application

4.1 Description

This section applies our promised misclassification correction to an evaluation of a program of payments for environmental services in Mexico. Payment for environmental services (PES) – defined as any voluntary agreement between a buyer and a seller in which the seller receives

[12] This procedure is similar to Altonji et al. (2005). There, the authors wish to estimate a probit model with an endogenous binary covariate using a bivariate probit model. Lacking an exclusion restriction in the model for the endogenous covariate, the authors note that the model is still identified due to the non-linearity of the bivariate normal CDF. Nonetheless, the authors treat the correlation coefficient between the errors as an unindentified parameter and conduct a grid search over different values.

[13] For practitioners wishing to apply this model, we also recommend using from the misclassification-corrected logit model as initial values in a more parsimonious model and then building up to the final specification.

payment for providing some environmental service such as conservation of the forest cover on the seller's land – have been implemented in countries ranging from the United States to Uganda to encourage environmentally-friendly behaviors on forest and farmland. In low and middle income countries, they have become a popular way to support Reducing Emissions from Deforestation and Degradation (REDD) commitments to climate change mitigation (Jack et al., 2008).

Mexico has a relatively long history of PES policies, and previous analyses have shown the programs to be in deterring deforestation, although with significant variation across time and space.[14] Here, we assess the effect of Mexico's Payments for Hydrological Services program between 2003 and 2015. This program is part of a broader national system of PES that is run by Mexico's National Forestry Commission (CONAFOR, for its Spanish acronym). The program compensates landowners who maintain intact forest cover on their properties with the goal of reducing deforestation. Contracts are to either individual or common-property landowners and last five years. Payments to landowners are conditional on maintaining land cover and completing conservation activities. Until recently, participants were able to apply and receive payments multiple times. The program is monitored by a combination of remote sensing and field verification activities.

To evaluate the program, we use administrative information on properties that submitted applications to the program. The unit of analysis is a parcel (polygon) within a property. The reason for this is that applicants may apply and enroll multiple times to the program. In order to avoid double-counting, the analysis polygons were created by dividing applicant parcels into smaller units that preserve their unique application histories. For example, if a landowner submitted a parcel in 2010 and was rejected, and the following year submitted an imperfectly overlapping parcel that was accepted, these two applications would generate three non-overlapping polygons: one polygon with an indicator for rejected in 2010, another that has an indicator for being rejected in 2010 and then accepted in 2011, and a final polygon coded as accepted in 2011. Figure E1 shows a visual representation of these units within various communities with repeated applications. We limit polygons to those between 10 ha and the maximum allowable parcel size for each application cohort (between 2,000 and 6,000 ha, depending upon the year). The lower bound is meant to eliminate "slivers" of overlap between polygons and the upper bound to get rid of potential errors in the polygon bound-

[14] See Alix-Garcia et al. (2019) for discussion of program history and of program impacts.

aries. All of the properties in our analysis are from "ejidos", which is a Mexican tenure system where land is managed by groups of people. We consider the ejido the unit of decision-making. Each ejido has multiple polygons.

For each polygon we calculate a number of covariates that are associated with forest cover change. These include elevation, slope, distance to nearest road, baseline forest cover in 2000, area of the polygon, and whether or not the polygon is located in a majority indigenous municipality. While previous evaluations of this program have used a more complicated set of covariates and different identification strategies (Alix-Garcia et al., 2012, 2015, 2019), our purpose is to illustrate how our proposed accounting for misclassification affects the estimates.

The deforestation and baseline forest area measures come from Hansen et al. (2013), version 1.2 (accessed in 2016). Importantly, the annual forest cover loss does not come from a difference in levels of measured forest, but rather from a separate time-series analysis that detects disturbances in the pixels deemed to have forest cover in 2000. This data is the only available source with annual variation in deforestation during our period of study; alternative estimators exploiting multiple misclassified measures would not be applicable. We define the true outcome, y_{it}^*, an indicator equal to one if any deforestation occurred on polygon i in year t and zero otherwise. The observed outcome, y_{it}, is a binary indicator if any deforestation is recorded within the polygon in a given time period. This particular dataset registers the most recent deforestation event that occurred in the data. Because it does not track forest regrowth on an annual basis, pixels cannot become deforested and then forested. Because we are aggregating deforestation up to the polygon level, however, our data can register repeated deforestation events.

It has been shown that the accuracy of the deforestation data varies across countries and ecosystem types. For example, assessments by the CONAFOR remote sensing team suggest that the Hansen product offers better results for pine forests than for dry tropical ecosystems. The data are likely to understate loss of natural forest because it may classify plantations and agroforestry crops as forested areas, and it may also fail to capture selective logging or very small areas of deforestation. In a comparison between locally calibrated measures of deforestation and the Hansen measures of deforestation in Madagascar, the Hansen data captured only 64% of deforestation due to slash and burn agriculture (Burivalova et al., 2015). Mitchard et al. (2015) compare deforestation

rates measured using 5 m satellite imagery to the Hansen data and find that while classification was reasonably accurate in Brazil, omitting between 16 and 18% of probable deforestation, it missed 80% of the deforestation events in Ghana. Using our misclassification terminology, these studies suggest a high presence of false negatives in the data.

The Payments for Hydrological Services program is not randomly assigned. Rather, determination of beneficiary status requires several steps. First, applications are limited to geographic "eligible zones" determined by CONAFOR. Any applications coming from outside of eligible zones are automatically rejected. Applications from within eligible zones are evaluated according to a variety of criteria. Although the number of criteria have increased over time, variables used in the decision process throughout the program's history include measures of environmental quality (forest type and location in particular water-scarce areas), opportunity cost (deforestation risk as determined by geographic factors), and social criteria (location in marginalized or indigenous municipalities) (Sims et al., 2014; Alix-Garcia et al., 2019). The data contain all applicants, including those that did not end up receiving payments from the program. We refer to successful applicants as program *beneficiaries* and unsuccessful applicants as *non-beneficiaries*.

Our final sample is a balanced panel of 12,272 polygons from 2001–2014, for a total sample size of 171,808. Of these, 6,259 polygons are beneficiaries in at least one year; 29,667 observations in the sample are beneficiaries. Table 2 displays the summary statistics. Over the sample period, 18.2% (20.7%) of polygon-year observations in the non-beneficiary (beneficiary) sub-sample are classified in the Hansen data as experiencing some deforestation. We code this as deforestation being equal to 1.[15] Note that this variable is not equivalent to the rate of deforestation across forests, since polygons have varying sizes. Importantly, the geographic attributes of polygons are correlated with beneficiary status. In particular, beneficiaries tend to be at slightly lower elevation, higher slope, closer to roads and cities, with higher baseline forest cover, and in municipalities with greater indigenous presence. In addition, beneficiaries are often located in Landsat footprints with more cloud-free scenes from Landsat 7 sensors. Thus, even if beneficiary status is not directed correlated with misclassification in the Hansen data, it is likely correlated with other covariates

[15]We note that the scobit estimator can be sensitive to which outcome is assigned the non-zero value. Here, we follow the literature on rare events and assign the less frequent outcome to the non-zero value (e.g., Goleţ, 2014).

that are associated with misclassification.

4.2 Monte Carlo study

Before turning to the analysis of the actual data, we first undertake a (limited) Monte Carlo study intended to assess the performance of the estimators discussed in Section 3. The design of the simulation closely follows the basic structure of the panel data just discussed. In the interest of brevity, we focus our discussion on the average marginal effect (AME) of a binary treatment, although misclassification affects all of the coefficient estimates.

4.2.1 Design

Data are simulated from variants of the following DGP:

$$
\begin{aligned}
y_{it}^{*} &= \text{Bernoulli}(p_{it}), \quad i = 1, ..., N;\ t = 1, ..., T \\
p_{it} &= \frac{\exp(\beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 d_{it} + \omega_i)}{1 + \exp(\beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 d_{it} + \omega_i)} \\
x_{1it} &\overset{\text{iid}}{\sim} 0.01 \cdot \chi^2(25) \\
x_{2it} &\overset{\text{iid}}{\sim} \chi^2(45) \\
z_{it} &\overset{\text{iid}}{\sim} \text{Poisson}(9) \\
d_{it} &= \text{I}\left(-8 - 0.1 x_{1it} + 0.05 x_{2it} + 0.1 z_{it} + 0.5 \omega_i + u_{it} > 0\right) \\
u_{it}, \omega_i &\overset{\text{iid}}{\sim} N(0, 5) \\
\Pr(y_{it} = 1 | y_{it}^{*} = 0, z_{it}) &= \Phi\left(\theta_0 - 0.10 z_{it}\right) \\
\Pr(y_{it} = 0 | y_{it}^{*} = 1, z_{it}) &= \Phi\left(\theta_1 + 0.15 z_{it}\right)
\end{aligned}
$$

where $\text{I}(\cdot)$ is the indicator function taking a value of one if the argument is true and zero otherwise. Here, y_{it}^{*} and y_{it} are the true and misclassified binary outcomes, respectively, x_{1it} and x_{2it} are exogenous continuous covariates, d_{it} is an exogenous binary covariate, ω_i is a unit-specific unobserved effect, and $\Phi(\cdot)$ is the standard normal CDF.

To conform to our application, the distributions of the exogenous continuous covariates, x_{1it} and x_{2it}, align closely with two covariates in the real data: the slope of the land and distance

to the nearest road (divided by 100), respectively. The binary covariate, d_{it}, corresponds to the treatment variable in that the proportion of treated is roughly 20%. Finally, the distribution of the determinant of misclassification, z_{it}, closely mirrors the distribution of the number of cloud-free scenes.

In all designs, we set $\beta_1 = -3$, $\beta_2 = -0.1$, $\beta_3 = -2$, the number of cross-sectional units, N, is 2,000 and the number of time periods, T, is 15.[16] The following parameters are varied:

$$\begin{aligned} \beta_0 &\in \{3.5, 0, -3.5\} \\ \theta_0 &\in \{-1.5, -0.5\} \\ \theta_1 &\in \{-2.5, -1.3\} \end{aligned}$$

The parameter β_0 affects the proportion of ones in the true data. The three values of β_0 map to $\Pr(y_{it}^* = 1)$ being approximately 0.34, 0.14, and 0.04, respectively. The parameter θ_0 governs the false positive rate in the observed data. In our application, we believe false positives are rare. Thus, the two parameter values correspond to false positive rates of roughly 0.01 and 0.09, respectively. Finally, the parameter θ_1 determines the false negative rate in the observed data. In our application, we believe false negatives to be quite common. Thus, the two parameter values correspond to false negative rates of approximately 0.15 and 0.50, respectively.

Our objective is to estimate the average of the observation-specific marginal effects of x_{1it}, x_{2it}, and d_{it}. We report the bias and the root mean squared error (RMSE) for the AMEs based on 200 replications of each set of parameters. We consider seven estimators as described in Table 1.

The True CRE Logit (Estimator 1) applies the correct specification, assuming the CRE approximation to the true FEs is reasonable, to the true data. This serves as the benchmark since this is the best one can do in the absence of misclassification.[17] Second, the MC-CRE Logit (Estimator 6) is the correct specification, assuming the CRE approximation to the true FEs is reasonable, in the presence of misclassification. Third, although the MC-CRE Scobit (Estimator 7) is never the correct model, we evaluate it as an option since it may perform better when the outcome is of the rare

[16] In our application, T is 15 and N is about 20,000. Here, we set N to 2,000 to expedite the computations.

[17] Alternatively, one could estimate a fixed effects logit using the correctly measured data. However, computation of the AMEs is then not straightforward since the fixed effects are conditioned out of the likelihood function.

events type. Moreover, when estimating the MC Scobit, we conduct a grid search over the shape parameter, α, and use the estimates from the model corresponding to the highest log-likelihood function as discussed in Section 3.

4.2.2 Monte Carlo results

In the interest of brevity, the results are relegated to Appendix D. Moreover, we focus our discussion on the estimation of the AME of the binary covariate, d, as this aligns with our application where the parameter of interest is the AME of the binary PES treatment. Results for the other parameters are generally similar. Tables D1 - D3 report the bias and RMSE (both multiplied by 100) of each of the estimators considered across our 12 DGPs. Figure D1 plots the RMSE of the each estimator relative to the True CRE Logit for the treatment effect. Figures D2 and D3 do so for x_1 and x_2, respectively.

The simulations lead to four general takeaways. First, as expected, all estimators ignoring misclassification – CRE Logit, Ad Hoc CRE Logit, FE-LPM and Ad Hoc FE-LPM – do poorly across the majority of the DGPs. For example, with a high proportion of ones and a high degree of misclassification, the relative RMSE of all four estimators exceeds 12, meaning that it is 12 times larger than the RMSE of the benchmark case (see Figure D1). However, the relative performance of these estimators depends on the proportion of ones in the data, as well as the severity of the misclassification. When the proportion of ones is relatively high, the LPM estimators always dominate the CRE Logit estimators. However, as the outcome becomes more rare, the relative performance is much more variable.

Second, the *ad hoc* approach of adding covariates related to misclassification does not improve the performance of the CRE Logit and FE-LPM. More often, the addition of these covariates increases the bias (in absolute value), particularly when the false negative rate is high. The RMSE of the Ad Hoc FE-LPM is smaller than that of the FE-LPM in only five of 12 DGPs. The RMSE of the Ad Hoc CRE Logit is smaller than that of the CRE Logit in only one of 12 DGPs. Thus, despite it being commonplace to control for environmental variables thought to affect the reliability of remotely sensed outcomes, this is not a cure for the misclassification induced.

Third, the bias of the estimators ignoring misclassification is sometimes positive and sometimes

negative; the sign even occasionally varies across the LPM and CRE Logit estimators for the same DGP. This implies that misclassification (as modeled here) does not necessarily lead to attenuation bias. This is consistent with the conclusions in Hausman et al. (1998).

Finally, the estimators that account for misclassification have much smaller bias and RMSE overall. In particular, when the proportion of ones and the false negative rate are high, the MC-CRE Scobit and MC-CRE Logit perform similarly, as well as produce RMSEs that are close to the benchmark and much smaller than the RMSE of the remaining estimators. As the proportion of ones falls, the MC-CRE Scobit tends to outperform the MC-CRE Logit. However, the performance of each estimator worsens. Moreover, while some of the estimators ignoring misclassification perform well when the outcome is of the rare-events type, their performance is highly volatile as the extent and type of misclassification varies.

In sum, our simulation results confirm the ability of the MC-CRE Logit and MC-CRE Scobit to address misclassification, even in the case of rare events. Moreover, while the two estimators perform similarly in non-rare events data, the MC-CRE Scobit performs better in the case of rare events. Thus, our (admittedly limited) simulation exercise suggests that researchers rely on the MC-CRE Scobit. Furthermore, while there are a few instances where the estimators ignoring misclassification perform nearly as well as the estimators addressing misclassification, the vastly inferior performance in the majority of DGPs considered here suggests that researchers should not rely on them in practice.

4.3 Application results

With the guidance offered by the Monte Carlo study, we now turn to the results from our application. Covariates included in x are those listed in Table 2, as well as a binary indicator for beneficiary status and year dummies. Covariates included in z are the number of cloud-free images and the interactions between the number of cloud-free images and slope as well as area of the polygon. We consider two sets of location-specific fixed effects in the estimation. To begin, we use ejido fixed effects, since the ejido is the managerial unit of interest for the land that applied to the program. Ejidos can contain groups of polygons. As the variables in Table 2 are all measured

at the polygon level, use of ejido fixed effects allows some cross-sectional variation to be used in the estimation. However, inclusion of the binary indicator if a polygon ever received PES is meant to address selection into program. Alternatively, we use polygon fixed effects. While this controls for greater unobserved determinants of selection into treatment, it removes much of the variation in the data. All of the covariates in x with the exception of beneficiary status are time invariant within a polygon. Thus, treatment status, mean treatment status over time within a polygon, and a cubic time trend are the only covariates included in x in the polygon fixed effect specifications.[18] For these reasons, we prefer the ejido fixed effect specifications. In all cases, standard errors are clustered at the ejido level.

4.3.1 Ignoring misclassification

Tables 3 and 4 show the results from the estimators that ignore misclassification using ejido and polygon fixed effects, respectively. Two findings stand out. First, for both levels of fixed effects, the FE-LPM and Ad Hoc FE-LPM point estimates on the treatment effect for program beneficiary are the smallest in magnitude and not statistically significant. The treatment effect is the largest for the CRE Logit. All point estimates suggest that beneficiary status decreases the probability of deforestation, although the estimates for the polygon fixed effects tend not to be statistically different from zero. Thus, ignoring misclassification, we find evidence of a beneficial impact of PES on deforestation, particularly when not using a LPM.

Second, the estimated effects of the remaining covariates in the ejido fixed effects estimation are qualitatively similar across the various estimators. The value of the skewness parameter that maximizes the Scobit log-likelihood is small, at 0.20.

4.3.2 Incorporating misclassification

Tables 5 and 6 display results from the models that account for misclassification, based on (9) and (10), and include estimates of the proportion of false positives (G_0) and false negatives (G_1). The covariates included in x are identical to the CRE Logit and Scobit models in Table 3. The covariates

[18] We replace year dummies with a cubic time trend in the polygon fixed effects specification as the models failed to converge otherwise.

included in z are the number of cloud-free images (odd-numbered columns) or the number of cloud-free images and its interaction with the polygon area and average slope (even-numbered columns and column 5). In all cases, a likelihood ratio test rejects the restrictions in the more parsimonious specifications (odd-numbered columns) at the $p < 0.01$ level. Comparing the maximized value of the log likelihood functions indicates that the MC-CRE Logit in column 2 in Tables 5 and 6 fits the data best; however, the MC-CRE Scobit in column 4 in Table 6 achieves nearly identical performance as α is close to one. Thus, the MC-CRE Logit, allowing for the misclassification rate to depend on the number of cloud-free images and its interaction with topography, is the *preferred* estimator in this application.

To start, we briefly analyze the estimated false positive and negative rates. In terms of false positives, the estimated sample average probability is 0.7% according to our preferred ejido MC-CRE Logit in column 2 in Table 5. In contrast, the estimated sample average probability of a false negative is 14.7%. In light of the minimal evidence of false positives in the data, in the remainder of the specifications we constrain the false positive rate to zero to aid identification. The ejido MC-CRE Scobit models, while achieving a worse fit (as measured by the log-likelihood value), produce slightly larger estimates of the false negative rate in Table 5. In the specifications including the full set of covariates in the misclassification probabilities, the estimated sample average false negative rate ranges from 14.7% to 22.8%. Table 6 shows that the inclusion of polygon FE models produces a higher estimated false negative rate of roughly 32.1% in the specifications including the full set of misclassification covariates. Thus, there is substantial evidence of under-reporting of deforestation in the satellite data.

Figure 1 provides further evidence by plotting the density and cumulative density of the observation-specific estimates of false negative probabilities, $\Phi(z_{it}\hat{\theta}_1)$, from the various specifications including the full set of misclassification covariates. With ejido fixed effects, we see that between 20 and 60% of the sample has an estimated false negative rate exceeding 20%, depending on estimator, with the MC-CRE Logit being at the lower end (see Panel (a)). With polygon fixed effects, the estimated false the negative rates are essentially identical across the MC-CRE Logit and MC-CRE Scobit models. For the Logit and Scobit, the median estimated false negative rate exceeds 40% (see Panel (b)).

The estimated AMEs of beneficiary status are generally negative and statistically different from zero across all models. Broadly speaking, the effects indicate a decrease in the polygon-level probability of deforestation of one percentage point. The treatment effects estimated using ejido fixed effects and reported in Table 5 are smaller in magnitude and with lower significance levels than those estimated using polygon fixed effects and reported in Table 6. Comparing our preferred estimator, MC-CRE Logit, to the most commonly used estimator in practice, Ad Hoc FE-LPM, indicates attenuation bias in both cases when misclassification is ignored. The problem is more severe with polygon fixed effects than with ejido level effects: while the ejido MC-CRE Logit treatment effects are about 40% higher, the polygon MC-CRE Logit estimates are more than three times the magnitude of those from the corresponding Ad Hoc FE-LPM.

Failure to account for misclassification results in attenuation bias of the AMEs for several other covariates in the model as well. In particular, comparing the MC-CRE Logit (column 2 in Table 5) to the CRE Logit (column 3 in Table 3), we find the magnitude of the AME is smaller for slope, larger for distance to road, and smaller for being in a majority indigenous municipality. Also notable is that the AME of average slope varies between the odd- and the even-numbered columns. This occurs because the even-numbered columns allow the probability of a false negative to depend on average slope (and polygon area). Thus, failure to account for the effect of topography on misclassification alters the AME of topography on deforestation. Figure E2 shows the interquartile range of the estimates on the included covariates, and highlights the considerable differences in the point estimates of topographic variables in the misclassification-corrected estimators relative to the standard ones.

Finally, the observed proportion of outcomes equal to one in the data is 19%. Combining this proportion with the estimates of the false negative rate in column 2 suggests that the true proportion of outcomes equal to one is roughly 22%. This is in line with the accuracy studies mentioned previously. From the simulation results in Section 4.2, this suggests that the MC-CRE Logit and the MC-CRE Scobit should perform well if the model is otherwise correctly specified.

In sum, our analysis finds evidence of a beneficial effect of PES on deforestation in Mexico, with the effect being reasonably large in magnitude for the majority of the sample (a decline in the polygon level probability of deforestation of about 1 percentage point). In addition, the analysis

confirms the need to address misclassification in remotely sensed, binary measures of deforestation. In this particular case, AMEs from customary models used by researchers are attenuated for the treatment variable and biased for a number of included covariates. The models addressing misclassification also fit the data better, and the estimated levels of false positives and false negatives are generally consistent with accuracy assessments of the Hansen data on deforestation in other countries.

4.3.3 Incorporating misclassification via partial observability probit model

The apparent absence of false positives allows us to explore one further avenue for estimation in the presence of misclassification. Nguimkeu et al. (2019) show that a binary choice model with misclassification can be consistently estimated under the assumption of no false positives using the partial observability probit model developed in Poirier (1980).

The setup is quite similar to our approach. As in (1), true deforestation status is given by

$$y_{it}^{*} = \mathrm{I}(x_{it}\beta + \omega_i + \varepsilon_{it} > 0), \tag{14}$$

where ε_{it} is the error term. An incident of true deforestation is reported in the data if $q_{it} = 1$. Thus, q is one in the absence of a false negative. Consistent with (8), it is determined by

$$q_{it} = \mathrm{I}(z_{it}\tilde{\theta}_1 + \kappa_{it} > 0), \tag{15}$$

where κ_{it} is the error term and $\tilde{\theta}_1 = -\theta_1$. In the absence of false positives, observed deforestation, y_{it} is given by

$$y_{it} = y_{it}^{*} \cdot q_{it} = \mathrm{I}(x_{it}\beta + \omega_i + \varepsilon_{it} > 0, z_{it}\tilde{\theta}_1 + \omega_{it} > 0) \tag{16}$$

Under the assumption of no false positives and the error terms following a bivariate normal distribution with zero means and unit variances, the determinants of both y^* and q can be consistently estimated despite only y being observed using the partial observability bivariate probit model in Poirier (1980).

The results are displayed in column 5 in Tables 5 and 6. These show similar magnitudes and

significance levels to our preferred estimates based on the MC-CRE Logit. The estimated false negative rates are displayed in Figure 1. Here, we see higher (lower) estimated false negative rates from the partial observability model with ejido (polygon) fixed effects.

5 Conclusion

The opportunities for researchers to exploit remotely sensed data to gain new insights are seemingly infinite. In the case of deforestation, these insights are critically important. Changes in land use have far-reaching effects on climate change, biodiversity, and other environmental services. Slowing deforestation requires effective policy interventions. Remotely sensed data allows for empirical evaluation of such interventions by bringing previously unavailable data into the hands of researchers. However, to ensure the evaluations from which such insights are derived are credible requires researchers to properly understand this data source. New satellites with ever-greater resolution and different types of sensors are launched every year, and remote sensing scientists are constantly developing new algorithms to improve the accuracy of the final data products. Yet, with each new technology and translation, new sources of error will undoubtedly arise alongside the possibility to uncover previously unseen dynamics. To fully harness the potential of this information, researchers must engage in conversations across disciplinary boundaries to understand the construction of the data, and avoid the usage of naïve statistical models that fail to account for the nonclassical measurement error that may contaminate the data.

In this paper we have provided evidence of the extent and nature of mismeasurement in commonly used, remotely sensed data on forest cover. Although our focus has been on forest cover and deforestation, some lessons are generalizable. Sensor function, ecological attributes, and topographic features that lead to nonclassical measurement errors in data on forest cover can generate the same systematic errors when measuring other phenomena such as nighttime lights, urban development, air pollution, and more. Moreover, in remotely sensed, binary measures these errors must be nonclassical. Our simulation study reveals that this bias can be significant *and* need not necessarily lead to attenuation.

We have also demonstrated the feasibility and performance, both via simulation and through

an application, of several estimators when analyzing the determinants of a remotely sensed, binary outcome such as deforestation. In our application, failure to address misclassification in the analysis of deforestation in Mexico leads to attenuation bias. Once misclassification is addressed, we find that PES significantly slows deforestation.

While we believe the methods provided here offer a significant advancement over current research practices, much work remains to be done. One such opportunity is presented by the recently released global dataset measuring moist tropical forests (Vancutsem et al., 2021), which expands the landscape of publicly available remote sensing measures for forests and will allow for comparisons across multiple measures for both deforestation and afforestation. Allowing for greater flexibility in the functional forms, via semiparametric approaches or allowing the link functions for the misclassification rates to be asymmetric, may prove fruitful. Most importantly, future work is needed to better understand the nature of measurement error across different types of remotely sensed data, as well as develop a wider array of remedies. Such remedies might exploit multiple measures containing error, or spatial correlation in measurement error or the phenomena of interest. Future research is also needed to develop useful econometric tools when the remotely sensed outcome is continuous.

References

Abman, Ryan, and Clark Lundberg. 2020. Does free trade increase deforestation? the effects of regional trade agreements. *Journal of the Association of Environmental and Resource Economists* 7 (1): 35–72.

Aguilar-Tomasini, María Alejandra, Tania Escalante, and Michelle Farfán. 2020. Effectiveness of natural protected areas for preventing land use and land cover changes of the transmexican volcanic belt, mexico. *Regional Environmental Change* 20 (3): 1–9.

Alix-Garcia, Jennifer, Craig McIntosh, Katharine RE Sims, and Jarrod R Welch. 2013. The ecological footprint of poverty alleviation: evidence from mexico's oportunidades program. *Review of Economics and Statistics* 95 (2): 417–435.

Alix-Garcia, Jennifer M., Elizabeth N. Shapiro, and Katharine R. E. Sims. 2012. Forest conservation and slippage: Evidence from Mexico's National Payments for Ecosystem Services program. *Land Economics* 88 (4): 613–638.

Alix-Garcia, Jennifer M, Katharine RE Sims, Victor Hugo Orozco-Olvera, Laura Costica, Jorge David Fernandez Medina, Sofia Romo-Monroy, and Stefano Pagiola. 2019. *Can environmental cash transfers reduce deforestation and improve social outcomes? A regression discontinuity analysis of Mexico's national program (2011–2014).* The World Bank.

Alix-Garcia, Jennifer M, Katharine RE Sims, and Patricia Yañez-Pagans. 2015. Only one tree from each seed? Environmental effectiveness and poverty alleviation in Mexico's Payments for Ecosystem Services Program. *American Economic Journal: Economic Policy* 7 (4): 1–40.

Alston, Lee J, Krister Andersson, and Steven M Smith. 2013. Payment for environmental services: Hypotheses and evidence. *Annual Review of Resource Economics* 5: 139–159.

Altonji, Joseph G, Todd E Elder, and Christopher R Taber. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal Political Economy* 113 (1): 151–184.

Andam, Kwaw S, Paul J Ferraro, Alexander Pfaff, G Arturo Sanchez-Azofeifa, and Juan A Robalino. 2008. Measuring the effectiveness of protected area networks in reducing deforestation. *Proceedings of the National Academy of Sciences* 105 (42): 16089–16094.

Black, Dan A, Mark C Berger, and Frank A Scott. 2000. Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association* 95 (451): 739–748.

Burivalova, Zuzana, Martin R Bauert, Sonja Hassold, Nandinanjakana T Fatroandrianjafinonjasolomiovazo, and Lian Pin Koh. 2015. Relevance of global forest change data set to local conservation: case study of forest degradation in Masoala National Park, Madagascar. *Biotropica* 47 (2): 267–274.

Donaldson, Dave, and Adam Storeygard. 2016. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives* 30 (4): 171–98.

FAO. 2016. *Global Forest Resources Assessment 2015: how are the world's forests changing?* Food and Agriculture Organization of the United Nations, Rome.

Ferraro, Paul J, Merlin M Hanauer, Daniela A Miteva, Gustavo Javier Canavire-Bacarreza, Subhrendu K Pattanayak, and Katharine RE Sims. 2013. More strictly protected areas are not necessarily more protective: evidence from bolivia, costa rica, indonesia, and thailand. *Environmental Research Letters* 8 (2): 025011.

Fowlie, Meredith, Edward Rubin, and Reed Walker. 2019. Bringing satellite-based air quality estimates down to earth. In *AEA Papers and Proceedings*, volume 109. 283–88.

Gibson, John. 2020. Better Night Lights Data, For Longer. *Oxford Bulletin of Economics and Statistics* 83: 770–791.

Gibson, John, Susan Olivia, Geua Boe-Gibson, and Chao Li. 2021. Which night lights data should we use in economics, and where? *Journal of Development Economics* 149: 102602.

Goleţ, Ionuţ. 2014. Symmetric and asymmetric binary choice models for corporate bankruptcy. *Procedia - Social and Behavioral Sciences* 124: 282–291.

Government of Mexico. 2014. Guía para la interpretación de cartografía Uso del suelo y vegetación Escala 1:250 000 Serie V. *Instituto Nacional de Geografía e Estadística* .

Hansen, Matthew C, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, et al. 2013. High-resolution global maps of 21st-century forest cover change. *Science* 342 (6160): 850–853.

Hansen, Matthew C, Stephen V Stehman, and Peter V Potapov. 2010. Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences* 107 (19): 8650–8655.

Hausman, Jerry A, Jason Abrevaya, and Fiona M Scott-Morton. 1998. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87 (2): 239–269.

IPCC. 2019. Summary for policymakers. In *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.-O. Pörtner, D. C. Roberts, P. Zhai, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. Portugal Pereira, P. Vyas, E. Huntley, K. Kissick, M. Belkacemi, J. Malley, (eds.).

Jack, B Kelsey, Carolyn Kousky, and Katharine RE Sims. 2008. Designing payments for ecosystem services: Lessons from previous experience with incentive-based mechanisms. *Proceedings of the National Academy of Sciences* 105 (28): 9465–9470.

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

Kennedy, Robert E, Philip A Townsend, John E Gross, Warren B Cohen, Paul Bolstad, YQ Wang, and Phyllis Adams. 2009. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sensing of Environment* 113 (7): 1382–1396.

King, Gary, and Langche Zeng. 2001. Logistic regression in rare events data. *Political Analysis* 9 (2): 137–163.

Lancaster, Tony. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95 (2): 391–413.

Lewbel, Arthur. 2000. Identification of the binary choice model with misclassification. *Econometric Theory* 16: 603–609.

Li, Miao, Shuying Zang, Bing Zhang, Shanshan Li, and Changshan Wu. 2014. A review of remote sensing image classification techniques: The role of spatio-contextual information. *European Journal of Remote Sensing* 47 (1): 389–411.

Lorenzen, Matthew, Quetzalcóatl Orozco-Ramírez, Rosario Ramírez-Santiago, and Gustavo G

Garza. 2021. The forest transition as a window of opportunity to change the governance of common-pool resources: The case of mexico's mixteca alta. *World Development* 145: 105516.

Michler, Jeffrey D, Anna Josephson, Talip Kilic, and Siobhan Murray. 2022. Privacy protection, measurement error, and integration of remote sensing and socioeconomic survey data. *arXiv preprint arXiv:2202.05220* .

Mitchard, Edward, Karin Viergever, Veronique Morel, and Richard Tipper. 2015. Assessment of the accuracy of University of Maryland (Hansen et al.) Forest Loss Data in 2 ICF project areas– component of a project that tested an ICF indicator methodology. Technical report, University of Edinburgh.

Nagler, Jonathan. 1994. Scobit: an alternative estimator to logit and probit. *American Journal of Political Science* 38: 230–255.

NASA. 2021. Quotes to note. *Landsat Science website* .

Nguimkeu, Pierre, Augustine Denteh, and Rusty Tchernis. 2019. On the estimation of treatment effects with endogenous misreporting. *Journal of Econometrics* 208: 487–506.

Papke, Leslie E, and Jeffrey M Wooldridge. 2008. Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics* 145: 121–133.

Poirier, Dale J. 1980. Partial observability in bivariate probit models. *Journal of Econometrics* 12 (2): 209–217.

Robalino, Juan, and Alexander Pfaff. 2013. Ecopayments and deforestation in costa rica: A nationwide analysis of psa's initial years. *Land Economics* 89 (3): 432–448.

Robalino, Juan, Alexander Pfaff, and Laura Villalobos. 2017. Heterogeneous local spillovers from protected areas in costa rica. *Journal of the Association of Environmental and Resource Economists* 4 (3): 795–820.

Salemi, Colette. 2021. Refugee camps and deforestation in sub-saharan africa. *Journal of Development Economics* 152: 102682.

Schennach, Susanne M, and Yingyao Hu. 2013. Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association* 108 (501): 177–186.

Sims, Katharine R E, and Jennifer M Alix-Garcia. 2017. Parks versus PES: Evaluating direct and incentive-based land conservation in Mexico. *Journal of Environmental Economics and Management* 86: 8–28.

Sims, Katharine RE, Jennifer Alix-Garcia, Elizabeth Shapiro-Garza, Leah R Fine, Volker C Radeloff, Glen Aronson, Selene Castillo, et al. 2014. Improving environmental and social targeting through adaptive management in Mexico's payments for hydrological services program. *Conservation Biology* 28 (5): 1151–1159.

Torchiana, Adrian L., Ted Rosenbaum, Paul T. Scott, and Eduardo Souza-Rodrigues. 2020. Improving estimates of transitions from satellite data: A hidden markov model approach. Technical report, University of Toronto.

Union of Concerned Scientists. 2020. UCS Satellite Database .

U.S. Geological Service. 2021. What are the band designations for the Landsat satellites? .

Vancutsem, Christelle, Frédéric Achard, J-F Pekel, Ghislain Vieilledent, S Carboni, Dario Simonetti, Javier Gallego, et al. 2021. Long-term (1990–2019) monitoring of forest cover changes in the humid tropics. *Science Advances* 7 (10): eabe1603.

WRI. 2022. Global Forest Watch Country Dashboard, World Resources Institute .

Young, Nicholas E, Ryan S Anderson, Stephen M Chignell, Anthony G Vorster, Rick Lawrence, and Paul H Evangelista. 2017. A survival guide to Landsat preprocessing. *Ecology* 98 (4): 920–932.

Figure 1: Distribution of estimated false negative rates from logit, scobit, and partial observability probit models

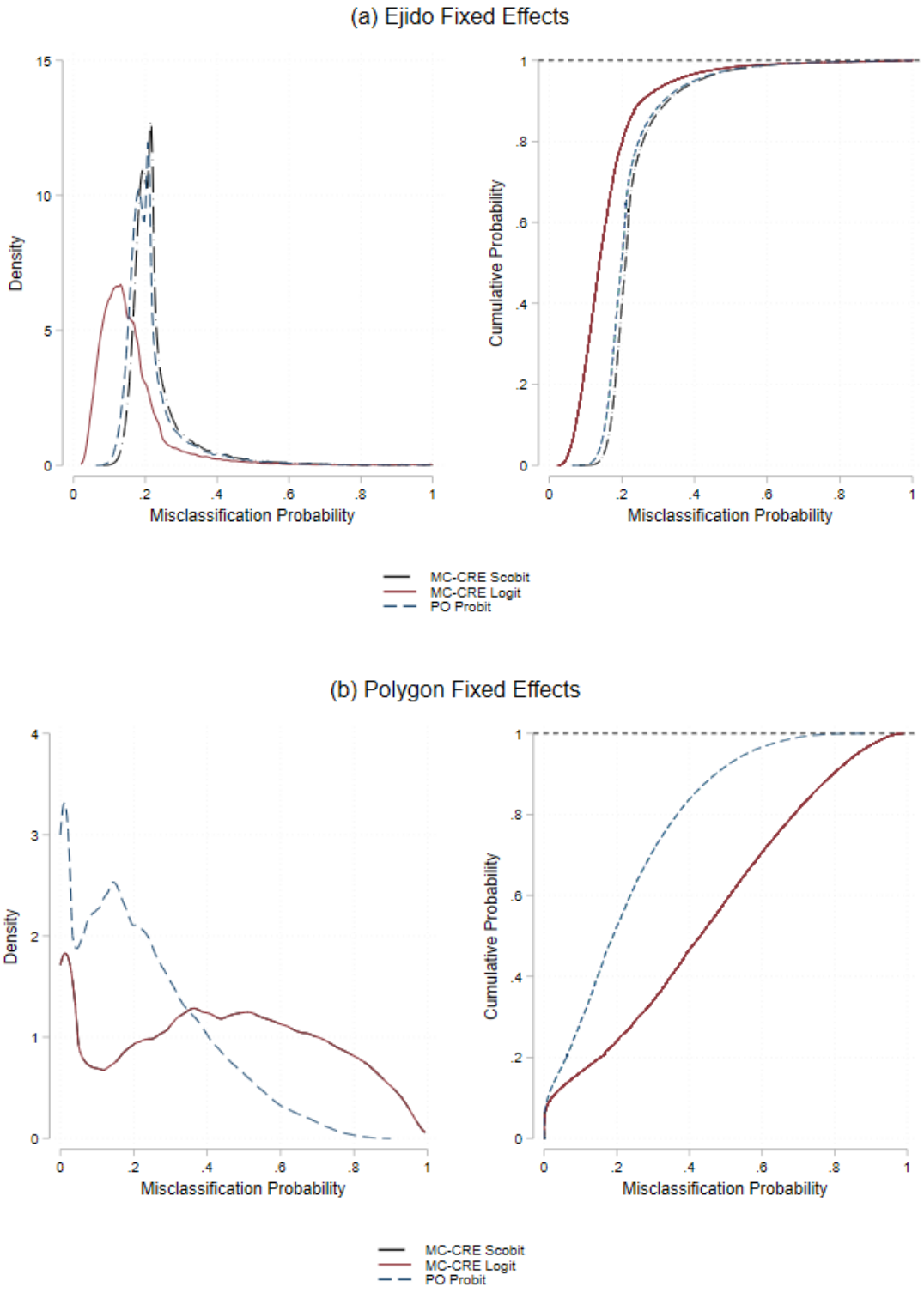

Table 1: Estimators explored in the Monte Carlo study

	Estimator	Dep. var	Covars.	FE	ME
1	True CRE Logit	y_{it}^{*}	x_{1it}, x_{2it}, d_{it}	$x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$	–
2	CRE Logit	y_{it}	x_{1it}, x_{2it}, d_{it}	$x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$	–
3	Ad Hoc CRE Logit	y_{it}	x_{1it}, x_{2it}, d_{it}, z_{it}	$x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$, $z_{i\cdot}$	–
4	FE-LPM	y_{it}	x_{1it}, x_{2it}, d_{it}	unit FEs	–
5	Ad Hoc FE-LPM	y_{it}	x_{1it}, x_{2it}, d_{it}, z_{it}	unit FEs	–
6	MC-CRE Logit	y_{it}	x_{1it}, x_{2it}, d_{it}	$x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$	$\Phi\left(\widetilde{z}_{it}\theta_0\right)$, $\Phi\left(\widetilde{z}_{it}\theta_1\right)$
7	MC-CRE Scobit	y_{it}	x_{1it}, x_{2it}, d_{it}	$x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$	$\Phi\left(\widetilde{z}_{it}\theta_0\right)$, $\Phi\left(\widetilde{z}_{it}\theta_1\right)$

The Column FE indicates how location specific effects are accommodated within the model, the column ME the structure of the misclassification correct, and the column α the assumptions of the shape parameter in the log likelihood function. $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$, $z_{i\cdot}$ are the unit-specific averages of the covariates. $\Phi\left(\widetilde{z}_{it}\theta_0\right)$ is the probability of a false positive and $\Phi\left(\widetilde{z}_{it}\theta_1\right)$ of a false negative, where $\widetilde{z}_{it}$ includes a constant and z_{it}.

Table 2: Summary statistics

	(1) Non-beneficiary land	(2) Beneficiary land	(3) Norm diff
Dependent Variable			
Deforestation (1 = yes)	0.182	0.207	0.044
Covariates, x			
Ever received PES (1 = yes)	0.446	1.000	1.115
Average Elevation (mt)	1528.927	1459.642	-0.049
Average Slope (degree)	15.445	15.349	-0.008
Distance to any road (meters)	4370.923	4020.060	-0.059
Distance to city with > 5,000 people	29.247	27.627	-0.057
Area of Tiny polygon	398.438	350.378	-0.066
Percent forest cover, 2000	0.641	0.723	0.180
Percent of majority indigenous	0.275	0.322	0.074
Covariates, z			
Cloud-free Landsat 7 images	8.788	9.440	0.121
Observations	142,141	29,667	171,808

The sample is divided into those parcels of land that were beneficiaries of a PES payment and those that applied but were rejected. Columns (1) and (2) show means for each group for the years a parcel fell into those categories and column (3) the normalized difference in means.

Table 3: Results from models ignoring misclassification: Ejido fixed effects

	FE-LPM (1)	Ad Hoc FE-LPM (2)	CRE Logit (3)	Ad Hoc CRE Logit (4)	CRE Scobit (5)	Ad Hoc CRE Scobit (6)
Beneficiary (1 = yes)	-0.006	-0.005	-0.008**	-0.007**	-0.005	-0.005
	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Ever received PES (1 = yes)	-0.009*	-0.011**	-0.010*	-0.012**	-0.010**	-0.011**
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Average Elevation (mt)	-0.005***	-0.005***	-0.005***	-0.005***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Average Slope (degree)	-0.279***	-0.490***	-0.293***	-0.350***	-0.295***	-0.334***
	(0.064)	(0.095)	(0.069)	(0.069)	(0.061)	(0.061)
Distance to any road (meters)	-0.010***	-0.011***	-0.011***	-0.011***	-0.010***	-0.010***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Distance to city with > 5,000 people	0.030	0.047	0.037	0.049	0.022	0.034
	(0.079)	(0.081)	(0.072)	(0.075)	(0.063)	(0.065)
Area of Tiny polygon	0.015***	0.021***	0.013***	0.015***	0.016***	0.018***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Percent forest cover, 2000	0.159***	0.162***	0.160***	0.174***	0.134***	0.135***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)
Percent of majority indigenous	0.067***	0.075***	0.051**	0.057***	0.052**	0.053**
	(0.025)	(0.025)	(0.021)	(0.021)	(0.021)	(0.021)
Cloud-free scenes, L7		0.001		0.003***		0.002***
		(0.001)		(0.001)		(0.001)
Cloud-free scenes, L7 x Avg Slope		0.019***				
		(0.006)				
Cloud-free scenes, L7 x Area		-0.000***				
		(0.000)				
α					0.200	0.200
log $\mathscr{L}$			-76303.971	-72550.338	-76060.010	-72288.585

Column headers indicate estimator. Standard errors are clustered at the municipality level. Marginal effects evaluated at sample means are displayed for the logit and scobit models. Fixed effects are at the Ejido level. Time fixed effects also included in all models. * p <.10, ** p< .05, *** p<.01.

Table 4: Results from models ignoring misclassification: Polygon fixed effects

	FE-LPM	Ad Hoc FE-LPM	CRE Logit	Ad Hoc CRE Logit	CRE Scobit	Ad Hoc CRE Scobit
	(1)	(2)	(3)	(4)	(5)	(6)
Beneficiary (1 = yes)	-0.004	-0.003	-0.005	-0.005	-0.006*	-0.005
	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Cloud-free scenes, L7		-0.001		0.003***		-0.001**
		(0.001)		(0.001)		(0.001)
Cloud-free scenes, L7 x Avg Slope		0.000*				
		(0.000)				
Cloud-free scenes, L7 x Area		0.000***				
		(0.000)				
α					0.200	0.200
log $\mathscr{L}$			-86707.221	-76164.080	-86705.241	-86705.241

Column headers indicate estimator. Standard errors are clustered at the municipality level. Marginal effects evaluated at sample means are displayed for the logit and scobit models. Fixed effects are at the polygon level. Time fixed effects also included in all models. * p <.10, ** p< .05, *** p<.01.

Table 5: Results from models allowing misclassification: Ejido fixed effects

	MC Logit		MC Scobit		Partial Observability
	(1)	(2)	(3)	(4)	(5)
Beneficiary (1 = yes)	-0.011*	-0.007*	-0.012*	-0.008*	-0.008**
	(0.006)	(0.004)	(0.006)	(0.004)	(0.003)
Ever received PES (1 = yes)	-0.022**	-0.018***	-0.022**	-0.018***	-0.017***
	(0.009)	(0.006)	(0.009)	(0.006)	(0.004)
Average Elevation (mt)	-0.005**	-0.004**	-0.005**	-0.004**	-0.004***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Average Slope (degree)	-0.583***	-0.416***	-0.581***	-0.487***	-0.496***
	(0.112)	(0.083)	(0.109)	(0.091)	(0.047)
Distance to any road (meters)	-0.020***	-0.014***	-0.019***	-0.015***	-0.014***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.001)
Distance to city with > 5,000 people	-0.004	-0.025	0.003	-0.003	0.003
	(0.118)	(0.094)	(0.111)	(0.087)	(0.047)
Area of Tiny polygon	0.041***	0.031***	0.041***	0.035***	0.035***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)
Percent forest cover, 2000	0.201***	0.154***	0.196***	0.154***	0.151***
	(0.018)	(0.013)	(0.017)	(0.014)	(0.007)
Percent of majority indigenous	0.070*	0.047*	0.068*	0.046*	0.049***
	(0.036)	(0.024)	(0.035)	(0.025)	(0.016)
G_0	0.002	0.007	0.000	0.000	0.000
G_1	0.471	0.147	0.467	0.228	0.215
α			0.900	0.900	
log $\mathscr{L}$	-72234.302	-71714.657	-72250.321	-71812.546	-71834.915
Full MC covars	no	yes	no	yes	yes

Column headers indicate estimator. Standard errors are clustered at the municipality level. Marginal effects evaluated at sample means are displayed. $0 and 1$ are the probability of a false positive and negative, respectively, evaluated at sample means. Column 1 allows the misclassification rates to depend on the number of L7 cloud-free scenes. Column 2 allows the misclassification rates to depend on the number of L7 cloud-free scenes and its interaction with average slope and polygon area. Time fixed effects included in all models. * p <.10, ** p< .05, *** p<.01.

Table 6: Results from models allowing misclassification: Polygon fixed effects

	MC Logit		MC Scobit		Partial Observability
	(1)	(2)	(3)	(4)	(5)
Beneficiary (1 = yes)	-0.141***	-0.014**	-0.170***	-0.014**	-0.012***
	(0.022)	(0.006)	(0.022)	(0.006)	(0.004)
G_0	0.000	0.000	0.000	0.000	0.000
G_1	0.765	0.321	0.764	0.321	0.136
α			0.200	0.900	
log $\mathscr{L}$	-80578.991	-77358.071	-80577.276	-77358.103	-77299.270
Full MC covars	no	yes	no	yes	yes

Column headers indicate estimator. Standard errors are clustered at the municipality level. Marginal effects evaluated at sample means are displayed. $0 and 1$ are the probability of a false positive and negative, respectively, evaluated at sample means. Column 1 allows the misclassification rates to depend on the number of L7 cloud-free scenes. Column 2 allows the misclassification rates to depend on the number of L7 cloud-free scenes and its interaction with average slope and polygon size. Time fixed effects included in all models. * p <.10, ** p< .05, *** p<.01.