Remotely Incorrect?

Jennifer Alix-Garcia[*]
Oregon State University

Daniel L. Millimet[†]
Southern Methodist University & IZA

January 28, 2021

Abstract

The past decade has witnessed an explosion of research, across many disciplines, relying on remotely sensed data. This explosion reflects an increase in computing technology, along with public access to many satellite data sources. While this clearly represents progress, given the numerous advantages of satellite imagery, researchers unfamiliar with the intricacies of the collection process typically overlook an important feature of the data: non-classical measurement error. Here, we detail the potential sources and nature of these errors, propose a solution for the case of a mismeasured, remotely sensed binary outcome, and validate this solution using a Monte Carlo study. We use our estimator to evaluate a conservation program in Mexico. Our analysis yields practical recommendations for researchers going forward.

JEL: C18, C21, Q23, Q28

Keywords: Satellite, Remote Sensing, Deforestation, 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 participants at the OARES virtual seminar series. E-mail: jennifer.alix-garcia@oregonstate.edu.

[†] E-mail: millimet@smu.edu.

1 Introduction

An increase in the number of satellites and growth in computer processing power over the past decade has driven an explosion in the availability of data derived from remote sensing methods. Examples of such data include measures of nighttime light intensity, rainfall, temperature, land use, deforestation, pollution, population, marine health, and more. Researchers are increasingly able to download and use this type of data without consulting with the scientists who have created it and thus are well-acquainted with its pitfalls. For example, since 2012, over 150 economic studies have used nighttime lights data (Gibson 2020). Among the over 3,500 citations of the Global Forest Change dataset (Hansen et al. 2013) since 2013, 34 are in economics journals.

Because satellites operate where surveyors cannot, they can gather data from remote locations and at spatial resolutions vastly superior to data reported at the regional or national level. Moreover, data collection by satellites necessarily eliminates many sources of error, such as enumeration errors and response biases, that often plague traditional data collection procedures. However, absence of familiar types of data error does not imply the absence of all error. Systematic mismeasurement in remotely sensed measurements likely is present. This paper details the potential sources and nature of these errors, proposes a solution for the case of a mismeasured, remotely sensed binary outcome, and validates this solution using a Monte Carlo study. Finally, we apply our estimator to the evaluation of a conservation program in Mexico.

Addressing measurement error in statistical analysis is critical, but this is particularly the case with satellite data. It is well-known that classical measurement error – errors that are idiosyncratic and mean zero – in a continuous covariate in a linear regression framework results in attenuation bias. However, classical measurement error in one covariate may also bias, in either direction, the coefficient estimates on other (correctly measured) covariates that are correlated with the mismeasured covariate. Classical measurement error in a continuous dependent variable in a linear regression model does not lead to bias, but does lead to a loss in precision.

However, there are two reasons to expect measurement error in satellite data to be non-classical. First and foremost, often the variables being generated from the remotely sensed data are not continuous. Examples of this are binary measures of any forest cover or a decrease in forest cover and bounded measures of nighttime light intensity (between zero and a fixed upper threshold), urbanization (a binary classification), or share of forest cover (a fraction between zero and one). With bounded variables, which include binary variables, measurement error cannot be classical as the errors are necessarily negatively correlated with the true value (Black et al. 2000). Second, geographic characteristics may affect the accuracy of remotely sensed images and these same characteristics may affect the phenomenon being measured. This issue results in non-classical measurement error in both continuous and bounded outcomes.

With non-classical measurement error in covariates, coefficient estimates remain biased, but attenuation bias is no longer assured. This same type of error in the dependent variable extends the impact beyond a loss in precision; all coefficient estimates will be biased (Hausman 2001). Importantly, this includes the estimates on otherwise exogenous covariates such as a randomized treatment assignment if the coefficient is non-zero.

To highlight these issues, in the first part of the paper we describe how the construction of measures from optical satellite sensors can lead not only to measurement error, but non-classical measurement error. Moreover, we confirm the non-classical nature of the errors in static measures of land cover – even when continuous – by making use of two satellite-based measures of land cover for Mexico near the same time period and based upon imagery from the same type of satellite. Having two measures allows us to assess their degree of divergence and the correlation between these divergences and aspects of the environment. Our analysis leads to a few key takeaways. First, while the two binary measures for the presence of any forest diverge for roughly 18% of the sample, the continuous measures of the proportion of forest coverage appear much more divergent. This suggests caution when using very granular remotely sensed data. Second, correlations between the differences in both the binary and continuous measures and environmental attributes are statistically and economically significant, confirming the non-classical nature of the measurement error.

In the second part of the paper we focus on the statistical analysis of binary, remotely sensed outcomes such as the presence of forest cover or deforestation.[1] Specifically, we discuss the bias such errors create, including in ad hoc models that control for geographic variables thought to affect the accuracy of remotely sensed data. We then propose a solution based on an extension to the misclassification binary choice model proposed in Hausman et al. (1998). In particular, we allow for the misclassification rates to depend on covariates, as in Lewbel (2000), and we embed this framework in a scobit model of binary choice, which nests the logit model as a special cases (Nagler 1994). The scobit introduces an additional shape parameter into the link function. This additional flexibility has been proven useful when the outcome is of the rare-events type (Goleţ 2014). Binary data are of the rare-events type when the proportion of ones is very low, as may be the case when the data measure forest cover or deforestation.

We assess our proposed solution in a Monte Carlo study designed to mimic our application. The simulations lead to three primary conclusions. First, ignoring measurement error in a binary outcome introduces significant bias. Second, the typical tactic of including attributes that may induce measurement error in remotely sensed data as covariates in the analysis is done in vain; the bias induced by the measurement error remains. Third, our proposed solution performs quite well. In particular, the misclassification logit model is preferred with non-rare events data. With rare events data, the misclassification scobit (with a low value of the shape parameter) is preferred.

In the final part of the paper, we re-visit the impact of a program of payments for ecosystem services on deforestation in Mexico over the period 2001–2014 while accounting for misclassification

[1] We focus on binary outcomes for two reasons. First, as mentioned above, continuous measures of forest cover appear to be much noisier and therefore less useful in practice. Second, as we discuss, the non-linearity of binary choice models allows us to draw on existing solutions to measurement error. We certainly agree that future work should think carefully about addressing non-classical measurement error with continuous, remotely sensed outcomes.

in the presence of deforestation. Applying naïve models ignoring measurement error, as well as our proposed solutions, we find several striking results. First, satellite measures vastly under capture the true extent of deforestation. In our preferred specification, we find that half of all instances of deforestation are missed. In other words, there is a high rate of false negatives. However, there is little evidence that reported deforestation is erroneous – if forest loss is detected in the data, it is likely to have occurred. Overall, then, we find about 17% of the observed reports are misclassified.

Second, ignoring misclassification often results in attenuation bias of the average marginal effects. In particular, our preferred estimator suggests that the conservation program examined reduces the probability of deforestation by 1.4% on average. In contrast, the most commonly used estimator currently, what we refer to as the ad hoc fixed effects linear probability model, produces an estimate that is one-third the best estimate in magnitude and not statistically different from zero at conventional levels. The average marginal effects are also significantly attenuated for all of the included covariates.

Third, we find that cloud coverage and topography are important determinants of data accuracy when it comes to remotely sensed measures of deforestation. Failure to allow slope to affect reported measures of deforestation through misclassification significantly attenuates the estimates of the direct effect of slope on deforestation. In other words, not only does addressing misclassification lead to a larger (in absolute value) effect of slope on deforestation, but allowing the misclassification rates to depend on slope leads to an even larger effect.

In sum, our analysis leads to several recommendations for researchers interested in using remotely-sensed metrics. Most importantly, researchers ought to engage with remote sensing scientists to understand how the data are constructed and the nature of its limitations. In addition, researchers should address the non-classical measurement error in the data. When the remotely sensed data is being used to construct a binary outcome, the misclassification estimators investigated here offer an improvement over current practices, even in the case of rare events data. This is especially true when comparing the performance of the misclassification estimators to linear probability models.

Our contributions are both descriptive and methodological. First, while there are now large numbers of papers using satellite-based measures of various concepts, and at least two reviews focusing on the use of these measures in economics (Donaldson & Storeygard 2016, Jain 2020), we are among the few papers to both document the potential sources of measurement error as well as investigate possible solutions. Two exceptions to this are Gibson et al. (2019) and Gibson (2020). Both papers examine nighttime lights data. The former broadly reviews the two main data sources for nighttime lights measurement, describing the technical details of the sensors and giving recommendations on how to best choose between sources, aggregate data, and understand output. The latter establishes that early nighttime lights sources suffer from mean-reverting errors, describes the sources of these errors, and establishes a protocol for improving the accuracy of the more recent (VIIRS) dataset. By contrast, we focus more broadly on how processing and interpretation of raw imagery can result in non-classical error and propose estimation strategies that take this error into account.

On the methodological side, to our knowledge, ours is the first paper to consider an extended version of the estimator in Hausman et al. (1998) and Lewbel (2000) 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. Our approach can also be combined in a model such as that proposed in Nguimkeu et al. (2019) to address measurement error in a binary treatment if one is using a binary, remotely sensed variable as a covariate.

The rest of the paper proceeds as follows. Section 2 provides an overview of how remotely sensed data are constructed, as well as how systematic error may be introduced. Section 3 discusses the econometric problems arising from measurement error in binary measures of land cover derived from satellite data along with several potential solutions. Section 4 presents a Monte Carlo simulations to evaluate the finite sample performance of various estimators. Section 5 illustrates the practical importance of addressing measurement error in the context of the evaluation of a payments for ecosystem service program in Mexico. Finally, Section 6 concludes.

2 Where do errors come from?

Users of economic data are well-aware of measurement problems associated with survey outcomes (Hausman 2001, Groves et al. 2011, Meyer et al. 2015) and economic aggregates from government agencies (Mankiw & Shapiro 1986). However, it may not be obvious to many social scientists that measures derived from optical, thermal, or radar sensors mounted on satellites can also have systematic sources of bias. To fix ideas, it is useful to begin with a description of how information from satellites is converted into usable static or dynamic data available to researchers. We use optical sensors as an example, but many of the steps that we describe here generalize to other types. Optical sensors are used to measure reflected energy, and come to the analyst as measurements of different "spectral bands" arranged in a grid (Kennedy et al. 2009).

There are presently over 2,000 satellites orbiting Earth. Each satellite has different technical specifications, including sensor type, frequency of reporting, and spatial resolution (Union of Concerned Scientists 2020). For optically-derived information, which is frequently used to produce land cover and land use change classifications, the basic process for classifying a satellite image into data that can be used by researchers begins by accessing images from their storage place in an archive, pre-processing these images so that they can be entered into an image-classification system (manual, automated, or a hybrid), and finally setting rules for translating the spatial and temporal trends in the satellite images into a metric.

This assembly-line of tasks creates three broad categories of potential errors: errors due to technical limitations of the sensors themselves, errors introduced in the pre-processing of images, and errors in the algorithms used to translate the signal from pixels in the image into data usable by an economist. Technical limitations can induce obvious challenges. For example, the product one is using 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 image is then completed in one of two ways: mosaicking (stitching together) multiple images from different time periods or directly imputing the missing imagery using predictions based on available data.

Even absent technical limitations of the machinery itself, the raw images are frequently distorted due to solar, atmospheric, and topographic features (Young et al. 2017), and often require significant pre-processing before they can be classified. While these corrections are necessary, they can introduce errors (Kennedy et al. 2009). Clouds frequently obscure visibility, and the timing of changes on the ground, for example, deforestation, may be delayed until images clear of cloud cover are available. Finally, before outcomes based on remote sensing reach the economic researcher, the original (now processed) signals are translated into a usable measure – such as particulate matter, land use, deforestation, population – using an algorithm.

There are an infinite number of ways to conduct this translation. For example, for relatively smaller areas, classification by visual inspection – comparing pixels from one image to areas from a higher resolution image, for example – is possible. For larger areas, there are classification methods based on pixel by pixel approaches and others that use broader spatial dimensions, which are known as "object-based" (Li et al. 2014). The former are currently more common, and these methods can be divided into two further groups, supervised and unsupervised. A 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 classifications divide 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 one value. There also exist methods that try to recognize potential heterogeneity within pixels and yield a classification for each pixel that states the possible proportion in given category. 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 (for example, textural analysis) to aid with classification (Li et al. 2014).

At their heart, the classification processes for remotely sensed data are algorithms. Like all such processes, the accuracy of a given model's predictions depends on the strength of the algorithm. Furthermore, any given process can have varying accuracy across space and time depending upon the underlying characteristics of the objects being classified. For example, it is well-known by remote sensing experts, though not to all end users of the data, that a data source widely-used in the detection of deforestation is more accurate in temperate than in tropical forests (Hansen et al. 2013), in areas of larger clearing (Burivalova et al. 2015), and in more homogeneous landscapes (Mitchard et al. 2015).

To illustrate the systematic nature of errors in the final data product employed by researchers, we examine two different measures of forest cover in Mexico. These two measures are based on similar imagery taken at around the same time, but they differ in the processing and classification of these images.

The first dataset is the "Land Use and Vegetation, Series V," (henceforth, GOM) and 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, and updates the previous Series IV map, which used a compilation of images from a different satellite between 2007 to 2010 (Government of Mexico 2014). The Landsat satellites all have a resolution of 30 meters. The GOM dataset classifies forest by type using supervised classification supported by ground-truthing in the field. This dataset comes in what is called "vector" form, which means that instead of being available as individual interpreted pixels, it is instead a series of polygons defined as homogeneous classes. There are 59 land use classes in the original dataset. 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 ha. The data are publicly available (Government of Mexico 2011, ARD et al. 2002).[2]

The second measure comes from the University of Maryland's Global Land Cover and Deforestation data set; hereafter, 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 as a "raster" dataset (in contrast to the vector above), which means that the information comes in a grid of 30 m cells, rather than as polygons. A pixel is classified as forested when its canopy cover measure exceeds 50%, a common cutoff.[3] The classification was done using a supervised algorithm with comparisons based upon higher resolution imagery as well as previous tree cover layers derived from both Landsat and lower resolution imagery (Hansen et al. 2013).

As mentioned above, the GOM and Hansen measures may differ due to the reclassification of the GOM product and the differences in scale of the two datasets. In addition, there are small differences between Landsat 5 and Landsat 7 (there is no Landsat 6). Both have the same spatial resolution (30 m) and image size (approximately 170 x 183 km), but Landsat 7 has an additional spectral band (U.S. Geological Service 2021). Imagery from these two satellites is quite frequently combined in remote sensing analyses (Kovalskyy & Roy 2013).

In order to make comparisons across the two datasets, we extract the information within a 5 x 5 km grid laid across the contiguous land area of Mexico. We engage in this aggregation because it makes the dataset more managable, and because some aggregation choice had to be made in order to make the vector (GOM) dataset comparable to the raster (Hansen). The process of aggregating

[2] See `Addurl.com`.

[3] It is not uncommon to use canopy cover cutoffs as low as 10% to define forest in global datasets, and the Hansen dataset offers a number of possible cutoff points.

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.

This aggregation yields the proportion of forest cover within each cell in 2010. We convert these continuous measures into binary indicators of any forest cover, with a threshold of 50 ha (based upon the processing of the GOM data) for both datasets. It would be possible to count smaller areas of forest as detectable in the Hansen dataset, but we use the same threshold in order to make the indicators more directly comparable. Below we discuss how different thresholds affect agreement between the two datasets. Finally, we observe several attributes of each cell, such as elevation, slope, and forest type. In order 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 A2 shows the distribution across Mexico of these images in 2010.

Table 1: Summary of measures in 5 km grid cells across Mexico

	Mean	SD	Obs
Proportion forest, 2011, GOM	0.360	0.397	79954
Proportion forest, 2010, Hansen	0.195	0.295	79954
Binary forest, 2011, GOM	0.581	0.493	79954
Binary forest, 2010, Hansen	0.482	0.500	79954
Mean elevation, 1000s m	1.022	0.820	79954
Mean slope	8.810	7.326	79954
Sd(slope)	7.198	4.917	79954
Dry tropical biome	0.189	0.391	79954
Moist tropical biome	0.137	0.344	79954
Pine/oak biome	0.231	0.421	79954
Mangrove biome	0.018	0.134	79954
Dry woodlands or grasslands	0.419	0.493	79954
Number of cloud-free scenes L7, 2010	12.124	3.899	79954
Number of cloud-free scenes L5, 2011	7.567	5.286	79954
Minimum of Landsat scenes, 2010-2011	7.299	4.942	79954

Table shows simple summary statistics for all variables used in analysis. L7 indicates the Landsat 7 satellite.

Table 1 shows means and standard deviations of the variables, including the measures of forest cover from the two sources. Three striking differences emerge. First, the GOM data reports considerably more forest cover, as a proportion of land, than the Hansen data; nearly 17 percentage points more. This may be because of the larger minimum mappable unit of the GOM data, but also due to differences anywhere along the processing line between the raw image and the final dataset. Second, the Hansen data reports a slightly lower fraction of cells with at least some forest coverage; the difference is about 10 percentage points. Together, these imply that there is greater divergence in the continuous measure of proportion forest than the binary measure of forest presence. Finally, the divergence between the two data sources is not uni-directional. Specifically, the Spearman rank correlation, which assesses the direction of correlation, is 0.78; the Pearson correlation coefficient, which measures both the direction and the magnitude of the relationship between the two variables, is 0.63. This higher Spearman correlation suggests a non-linear relationship.

To further investigate the divergence between the two sources, Table 2 shows the cross-tabulation of the binary measure of any forest in a cell across the two sources. The two data sources concur over 81% of the time. In about 4% of cells, the Hansen data detects some level of forest while the GOM data does not; the reverse occurs in 14% of the cells. Appendix Figure B4 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 1 for both datasets. Divergence is also low when cutoffs for both datasets are quite low (0 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. We present these relationships not to establish which data is more accurate, but rather to demonstrate that they are classifying the same landscape in different ways.

Table 2: Cross-tabulation of measures of any forest across Mexico

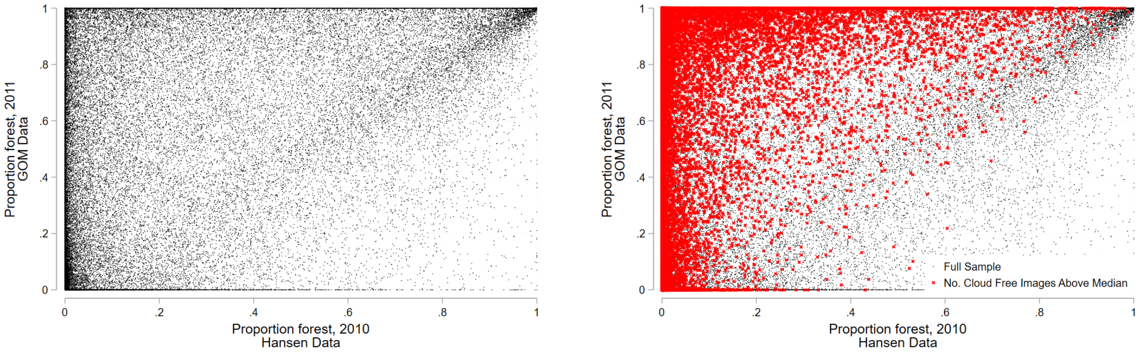
	Hansen data		
	0	1	Total
GOM data			
0	37.9	4.0	41.9
1	13.9	44.2	58.1
Total	51.8	48.2	100.0

Table shows cross-tabulation for two data sources used in analysis. Cells show percentages in each category.

To examine the differences visually, the left panel in Figure 1 presents a scatterplot of the continuous outcomes across the entirety of both data sets. If the two data sources were identical, all data points would lie along the 45 degree line. However, the figure makes it clear that this is far from the case. In particular, the Hansen data contains a significant mass of observations with low, but non-zero, forest cover. In contrast, the GOM data reports relatively larger forested areas. Nonetheless, a nontrivial share of the data also lies below the 45 degree line, indicating that the GOM (Hansen) data are not simply over- (under-)measuring forest cover.

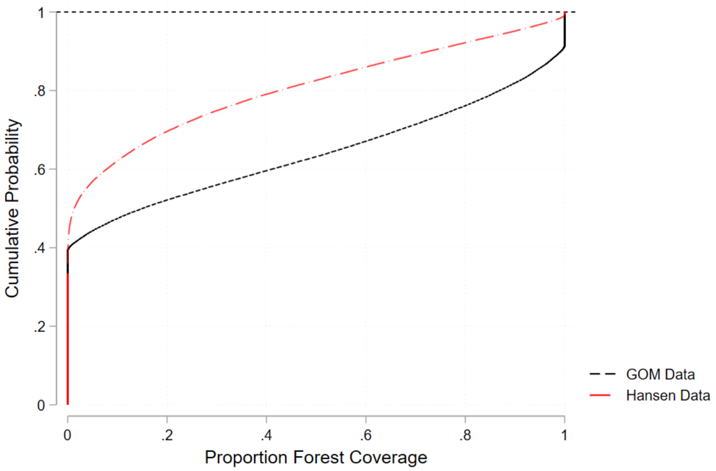
To assess whether these differences are explained by the differential availability of imagery due to cloud coverage, the right panel in Figure 1 identifies data points where the number of Landsat

Figure 1: Scatterplot of forest cover in 5 km cells across Mexico



7 cloud free images is above the median. Even in this sub-sample, the divergence between the data sources is stark. These differences are further highlighted by examining the empirical CDFs in Figure 2, which also shows that the Hansen data is dominated by smaller proportions of forest – nearly 70 percent of the forested proportions in Hansen are less than 0.20, as opposed to 50 percent of the GOM cells.

Figure 2: Empirical CDFs of forest cover in 5 km cells across Mexico



Next, we compute reliability statistics for the continuous measures of forest cover following the approach in Abowd & Stinson (2013). The results are shown in Table 3. Each row supposes that the truth is a weighted average of the GOM and Hansen data, with the weights listed in the first column. Given this "truth," the variance of the signal and measurement error contained in each data source is presented. The former is equal to $\mathrm{V}ar\left(y^*\right)$, where $y^* = \omega GOM + (1-\omega)Hansen$ is the "truth". The latter is equal to $\mathrm{V}ar\left(\mu\right)$, where $\mu = y - y^*$ is the measurement error associated with data source y. The final column displays the reliability statistic of each data source, given by $1 - \mathrm{V}ar\left(\mu\right)/\mathrm{V}ar\left(y^*\right)$. Note, the reliability statistic may be negative, as it for the Hansen data in the first row, if the measurement error is non-classical. In this case, the variance in the Hansen data, which is relatively small due to the large mass of data points close to zero (see Figure 1), is dwarfed by the variance of the measurement error if the GOM measure is accurate. Even if the truth is the equally-weighted average of both measures, then each measure has a reliability statistic below 0.85.

Table 3: Reliability statistics

Truth Model	Variance		Variance	Variance of ME		Reliability Statistic	
Weight(GOM,Hansen)	GOM	Hansen	of Signal	GOM	Hansen	GOM	Hansen
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1,0	0.158	0.087	0.158	0.000	0.097	1.000	-0.112
0.9,0.1	0.158	0.087	0.142	0.001	0.078	0.994	0.099
0.5,0.5	0.158	0.087	0.098	0.024	0.024	0.847	0.722
0.1,0.9	0.158	0.087	0.085	0.078	0.001	0.504	0.989
0,1	0.158	0.087	0.087	0.097	0.000	0.388	1.000

Finally, we assess how the divergence in what is measured by the two data sources is related to geographic characteristics of the land. Differences in measured outcomes that are systematically related to physical characteristics suggest that measurement errors may be nonclassical, leading to bias in statistical analyses. Table 4 explores the correlations between the difference in continuous or binary outcomes with geographic features as well as the availability of satellite imagery to support the classification process. The table displays beta coefficients. Most importantly, we find that the divergences are correlated, both statistically and economically, with 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 high for the pine-oak and tropical biomes. This is consistent with measurement error that enters during pre-processing – tropical areas tend to have more clouds – and through the classification algorithms used to define forest cover.

The differences 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. In interpreting the image availability effect, it is important to note that Landsat scenes are quite large – 185 km square on average – areas which can pick up ecosystem and other broad scale spatial effects.

Table B1 shows these same regressions using different cutoff thresholds for forest cover in the two datasets. We observe that as we increase the threshold towards 1, geographic characteristics (slope, elevation) become more important relative to image availability.

As we stated at the outset, we take no stance on the relative reliability of the GOM and Hansen

Table 4: Differences in measurement of forest cover using two Landsat-based sources

	Proportion of 5 km cell Abs(GOM - Hansen) (1)	Indicator of any forest Abs(GOM - Hansen) (2)
Ln(elevation, 1000s)	0.052*** (0.004)	0.131*** (0.006)
Mean slope	0.233*** (0.000)	-0.497*** (0.000)
Dry tropical biome	0.509*** (0.004)	0.281*** (0.007)
Moist tropical biome	0.290*** (0.004)	0.198*** (0.007)
Pine/oak biome	0.488*** (0.003)	0.217*** (0.006)
Mangrove biome	0.164*** (0.006)	0.100*** (0.013)
Minimum of Landsat scenes, 2010-2011	-0.055*** (0.000)	-0.180*** (0.001)
Minimum of Landsat scenes x slope mean	0.113*** (0.000)	0.370*** (0.000)
Mean DV	0.210	0.178

Column headers indicate dependent variables. The unit of analysis is 5 x 5 km grid cells across the entire landscape of Mexico. State fixed effects are included. Standard errors are robust, and the estimator is OLS. Beta coefficients are displayed. * p <.10, ** p< .05, *** p<.01.

data sources. We do, however, know that both are indicative of the types of data sources increasingly being used by researchers. As such, their divergences should cause researchers great pause. The differences here appear more extreme than similar exercises that compare data on self-reported income with linked administrative earnings records, where neither data source is seen as infallible (Kapteyn & Ypma 2007, Gottschalk & Huynh 2010, Abowd & Stinson 2013). Thus, we now turn to possible econometric remedies when estimating the determinants of a binary outcome, such as forest cover, measured via satellite imagery.

3 Empirics

3.1 Setup

In light of our application, we focus on estimating the determinants of a binary outcome in the presence of panel 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 assumed to be given by

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

where x_{it} is a vector of correctly measured, exogenous covariates, μ_i is a location–specific fixed effect (FE), and $F(\cdot)$ is referred to as the link function. If $F(\cdot)$ is the identity link function, then $F(\cdot) = x_{it}\beta + \mu_i$ and (1) is a linear probability model (LPM). The estimating equation in this case is

$$y^*_{it} = x_{it}\beta + \mu_i + \varepsilon_{it}. \tag{2}$$

If $F(\cdot)$ is the standard normal cumulative distribution function (CDF) or the logistic CDF, then (1) is the usual probit or logit model, respectively. A less well-known alternative model that we also consider is known as a skewed logit or scobit model (Nagler 1994). In the scobit model, $F(\cdot)$ is defined as

$$F(\cdot) = 1 - \frac{1}{[1 + \exp(x_{it}\beta + \mu_i)]^{\alpha}}, \tag{3}$$

where α is an unknown parameter. The scobit model corresponds to the usual logit model when $\alpha = 1$.

A few comments are warranted. First, location FEs are easily accomodated in the LPM by either first-differencing or mean-differencing the data and then estimating the transformed model via Ordinary Least Squares (OLS). 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).[4] A common solution in applied analysis is to assume a correlated random effects (CRE) structure. The CRE structure directly models the dependence between the FEs and the location-specific covariates. Specifically, we assume

$$E[\mu_i|x_i] = \overline{x}_i\gamma, \tag{4}$$

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

$$\mu_i = \overline{x}_i\gamma + \eta_i, \tag{5}$$

where η_i is now a location-specific random effect. Subsitution of (5) into (1) yields

$$\Pr(y^*_{it} = 1|x_{it}, \mu_i) = F(x_{it}\beta + \overline{x}_i\gamma + \eta_i), \tag{6}$$

which can be estimated using random effects binary choice models or traditional binary choice

[4] 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 invoking additional assumptions on the FEs.

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 & Zeng 2001). Such outcomes are referred to as rare events. One reason for the poor performance of probit and logit models in this case is because the link function, $F(\cdot)$, is symmetric. This implies that the probability approaches zero and one at the same rate as the index, $x_{it}\beta + \mu_i$, approaches $-\infty$ and $+\infty$, respectively. The probit and logit also possess the property that the marginal effects of the covariates are maximized for observations with an initial probability of y^*_{it} being one of one-half. This may also contribute to the poor performance of these models in the case of rare events.

While several alternatives for modeling rare events data have been proposed, we focus here on the the scobit model. It may potentially perform better with rare events data because the link function is no longer symmetric when $\alpha \neq 1$. In particular, as $\alpha \to 0$, the probability of y^*_{it} being one conditional on x_{it} and μ_i falls.[5] This implies lower probabilities of the value one occurring unless the index, $x_{it}\beta + \mu_i$, is quite large. Goleţ (2014) finds that the scobit does very well when modeling rare corporate bankruptcies.

Finally, ML estimation of the scobit models produces consistent estimates of the parameters if the DGP is correct. The LPM, while convenient and popular, is unlikely to produce consistent estimates (Horrace & Oaxaca 2006).

3.2 Measurement Error

When y^* is not observed by the researcher, but rather a mismeasured 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

$$y_{it} = y^*_{it} + \omega_{it}, \tag{7}$$

where $\omega_{it} \in \{-1, 0, 1\}$. However, since ω_{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 the case that $\text{Cov}(y^*_{it}, \omega_{it}) < 0$ in the presence of measurement error. Substituting (7) into (2) yields the following estimating equation

$$y_{it} = x_{it}\beta + \mu_i + \omega_{it} + \varepsilon_{it} \tag{8}$$

Since $\text{Cov}(y^*_{it}, \omega_{it}) < 0$, it follows that $\text{Cov}(x_{it}\beta + \mu_i + \varepsilon_{it}, \omega_{it}) < 0$, leading to biased estimates. In words, since the measurement error is negatively correlated with the truth, it is also negatively correlated with the determinants of the truth. Consequently, all of the covariates in (8) become

[5] The scobit model is similar to the Generalized Extreme Value (GEV) regression model proposed in Calabrese & Osmetti (2013). The GEV model also introduces an additional free parameter into the link function to allow for asymmetry.

endogenous.[6]

To see the inconsistency resulting from measurement error in the ML models, we introduce the following misclassification probabilities

$$\begin{aligned} \Pr(y_{it} = 1 | y^*_{it} = 0, z_{it}) &= G_0(z_{it}\theta_0) & (9) \\ \Pr(y_{it} = 0 | y^*_{it} = 1, z_{it}) &= G_1(z_{it}\theta_1), & (10) \end{aligned}$$

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 (9) and (10) 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, say α_0 and α_1. 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). As discussed in Section 2, with remotely sensed data the probability of misclassification may depend on weather variables, such as cloud cover, or geographic variables, such as the slope of the land.

Combining (1), (5), (9), and (10), the probability of a one or zero occurring in the observed data is given by

$$\begin{aligned} \Pr(y_{it} = 1 | x_{it}, z_{it}, \mu_i) &= G_0(z_{it}\theta_0) + [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i) & (11) \\ \Pr(y_{it} = 0 | x_{it}, z_{it}, \mu_i) &= 1 - G_0(z_{it}\theta_0) - [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i) & (12) \end{aligned}$$

A naïve ML model that ignores measurement error uses the following (incorrect) probabilities to construct the likelihood function:

$$\begin{aligned} \Pr(y_{it} = 1 | x_{it}, z_{it}, \mu_i) &= F(x_{it}\beta + \overline{x}_i\gamma + \eta_i) & (13) \\ \Pr(y_{it} = 0 | x_{it}, z_{it}, \mu_i) &= 1 - F(x_{it}\beta + \overline{x}_i\gamma + \eta_i). & (14) \end{aligned}$$

This model will yield inconsistent estimates. However, deriving the likelihood function based on (11) and (12) will yield consistent estimates assuming the full DGP is correctly specified. Specifically, the log-likelihood function is

$$\begin{aligned} \ln \mathcal{L} = & \sum_i \sum_t \{ y_{it} \ln \{ G_0(z_{it}\theta_0) + [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i) \} & (15) \\ & + (1 - y_{it}) \ln \{ 1 - G_0(z_{it}\theta_0) - [1 - G_0(z_{it}\theta_0) - G_1(z_{it}\theta_1)]\, F(x_{it}\beta + \overline{x}_i\gamma + \eta_i) \} \}. \end{aligned}$$

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-

[6] One exception to this occurs if $\beta = 0$. However, note that even if one element of β, say β_k, equals zero but the corresponding covariate, x_k is correlated with other elements of x with non-zero coefficients, then the estimate of β_k will still be biased.

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 one additional set of estimators for comparison. Researchers aware of the effect of weather and geographic variables, z_{it}, on the accuracy of satellite data often choose to simply control for 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 now given by

$$y_{it} = x_{it}\beta + z_{it}\theta + \mu_i + \varepsilon_{it} \tag{16}$$

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

$$\Pr(y_{it} = 1|x_{it}, z_{it}, \mu_i) = F(x_{it}\beta + z_{it}\theta + \overline{x}_i\gamma_0 + \overline{z}_i\gamma_1 + \eta_i). \tag{17}$$

A few final comments pertaining to identification and estimation in the ML models allowing for misclassification 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 (11) and (12). 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 not included in z.

Second, in the scobit model, identification of the shape parameter, α, along with the misclassification probabilities is also tenuous. 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 constrain it to different values.[7] By doing a grid search over α, we can assess sensitivity of the results to changes in α. Moreover, we can compare values of the log-likelihood functions for model selection.

Third, we follow Papke & Wooldridge (2008) and estimate the MC-CRE Scobit models using the traditional scobit probabilities (i.e., ignoring the presence of the random effect, η). However, we adjust the standard errors in order to allow for arbitrary serial correlation by clustering at the

[7] This procedure is analogous 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 first-stage for the endogenous covariate, they note that the model is still identified due to the non-linearity of the bivariate normal CDF. Nonetheless, they treat the correlation coefficient between the errors as an unindentified parameter and conduct a grid search over different values.

unit level (or higher).

Finally, given our discussion in Section 2 that made use of two potentially mismeasured versions of the same object of interest, one might be tempted to consider econometric methods that exploit access to such data (e.g., Black et al. 2000, Browning & Crossley 2009). We do not pursue this here for the main reason that we only have access to a single source of remotely sensed data in our application. However, this could be a valuable avenue for future research.

4 Monte Carlo Study

This section presents a Monte Carlo study intended to assess the performance of standard approaches for estimating treatment effects in a panel setting relative to estimators that correct for misclassification of the outcome variable. The design of the simulation closely follows the structure of the data in our application in section 5. We focus our presentation of the results related to the estimation of the treatment effect, although the measurement error has implications for all of the coefficient estimates in the regressions.

4.1 Design

Data are simulated from variants of the following DGP:

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

where Bernoulli$(\cdot)$ is the Bernoulli distribution, Poisson$(\cdot)$ is the Poisson distribution, χ^2 is the Chi-squared distribution, and I$(\cdot)$ is the indicator function taking a value of one if the argument is true and zero otherwise. Here, y_{it}^* is the true binary outcome, x_{1it} and x_{2it} are exogenous continuous covariates, d_{it} is an exogenous binary covariate, and μ_i is a unit-specific unobserved effect. Note, if y_{it}^* is observed, then a FE logit model is the correct specification.

To add misclassification, y_{it}^* is unobserved to the researcher; y_{it} is observed instead. $\Pr(y_{it}\neq y_{it}^*|y_{it}^*=0,z_{it})$ is the (conditional) probability of a false positive, where $\Phi(\cdot)$ is the standard normal

CDF. $\Pr(y_{it} \neq y_{it}^* | y_{it}^* = 1, z_{it})$ is the (conditional) probability of a false negative. These probabilities depend on a covariate, z_{it}. With y_{it} observed in lieu of y_{it}^*, a FE logit model no longer produces consistent estimates of β.

The data-generating process is designed 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 used in our application below: the slope of the land and distance to the nearest road, respectively. The binary covariate, d_{it}, corresponds to the treatment variable in our application 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. In our application, T is 15 and N is about 20,000. Here, we use N equal to 2,000 to expedite the computations. 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 $\Pr\left(y_{it}^* = 1\right)$ 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 marginal effects of x_{1it}, x_{2it}, and d_{it}. The true marginal effects are given by

$$\begin{aligned} ME(x_{1it}) &= \Lambda\left(W_{it}\right)\left[1 - \Lambda\left(W_{it}\right)\right]\beta_1 \\ ME(x_{1it}) &= \Lambda\left(W_{it}\right)\left[1 - \Lambda\left(W_{it}\right)\right]\beta_2 \\ ME(d_{it}) &= \Lambda\left(\beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 + \mu_i\right) - \Lambda\left(\beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \mu_i\right) \end{aligned}$$

where $\Lambda(\cdot)$ is the logistic CDF and $W_{it} \equiv \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 d_{it} + \mu_i$. The corresponding average marginal effect (AME) of each covariate is the average of these observation-specific marginal effects.

We report the bias and the root mean squared error (RMSE) for the AMEs based on 500 replications of each set of parameters. The following estimators are considered:

1. True CRE Logit: y_{it}^* on x_{1it}, x_{2it}, d_{it}, $x_{1i\cdot}$, $x_{2i\cdot}$, and $d_{i\cdot}$, where $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$ are the unit-specific averages of the covariates.

2. CRE Logit: y_{it} on x_{1it}, x_{2it}, d_{it}, $x_{1i\cdot}$, $x_{2i\cdot}$, and $d_{i\cdot}$, where $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$ are the unit-specific averages of the covariates.

3. Ad CRE Hoc Logit: y_{it} on x_{1it}, x_{2it}, d_{it}, z_{it}, $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$, and $z_{i\cdot}$, where $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$, $z_{i\cdot}$ are the unit-specific averages of the covariates.

4. FE-LPM: y_{it} on x_{1it}, x_{2it}, d_{it}, and fixed effects.

5. Ad Hoc FE-LPM: y_{it} on x_{1it}, x_{2it}, d_{it}, z_{it}, and fixed effects.

6. MC-CRE Logit: y_{it} on x_{1it}, x_{2it}, d_{it}, $x_{1i\cdot}$, $x_{2i\cdot}$, and $d_{i\cdot}$, where $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$ are the unit-specific averages of the covariates and the probability of a false positive is modeled as $\Phi\left(\widetilde{z}_{it}\theta_0\right)$ and a false negative as $\Phi\left(\widetilde{z}_{it}\theta_1\right)$ and $\widetilde{z}_{it}$ includes a constant and z_{it}.

7. MC-CRE Scobit: y_{it} on x_{1it}, x_{2it}, d_{it}, $x_{1i\cdot}$, $x_{2i\cdot}$, and $d_{i\cdot}$, where $x_{1i\cdot}$, $x_{2i\cdot}$, $d_{i\cdot}$ are the unit-specific averages of the covariates and the probability of a false positive is modeled as $\Phi\left(\widetilde{z}_{it}\theta_0\right)$ and a false negative as $\Phi\left(\widetilde{z}_{it}\theta_1\right)$ and $\widetilde{z}_{it}$ includes a constant and z_{it}. We constrain the parameter α to be 0.25, 0.50, and 0.75.

To summarize the key comparisons, the True CRE Logit (Model 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.[8] Second, the MC-CRE Logit (Model 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 (Model 7) is never the correct model, we evaluate it as an option since it may perform better when the outcome is of the rare events type. Moreover, when estimating the MC Scobit, we fix the shape parameter, α, at various values rather than estimate it given the identification concerns discussed in Section 3.

4.2 Results

In the interest of brevity, and aligning with our application where the parameter of interest is the AME of a treatment, we focus our discussion on the estimation of the AME of the binary covariate, d. The full set of results are provided in Appendix C and are generally similar. Table 5 reports the bias and RMSE (both multiplied by 100) of each of the estimators considered across our 12 DGPs. Recall, as β_0 declines, the proportion of ones in the data falls from roughly 0.34 to 0.17 to 0.04. Thus, lower values of this parameter lead to the outcome being more in line with the rare event type. Moreover, the true AME also depends on the value of β_0, varying from roughly -0.13 to -0.07 to -0.03 as the parameter declines. In Panels A and B (C and D), the false positive

[8] 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.

rate is about 0.01 (0.09). In Panels A and C (B and D), the false negative rate is about 0.15 (0.50). Finally, Figure 3 plots the RMSE of each estimator relative to the RMSE of the True CRE Logit for each DGP.

In terms of bias, a few interesting patterns arise. First, the True CRE Logit consistently underestimates the true AME on average, although the bias is small (as the bias reported in the table is multiplied by 100). Second, the CRE Logit and Ad Hoc CRE Logit have the largest bias (in absolute value) when the proportion of false negatives is high (i.e., $\theta_1 = -1.3$) or when the proportion of ones is high (i.e., $\beta_0 = 3.5$). In these cases, the bias for these two models is an order of magnitude higher than it is in the True model. Only when the proportion of false negatives is relatively low and the proportion of ones is relatively low (i.e., $\theta_1 = -1.3$, $\beta_0 = 0, -3.5$) do the CRE Logit and Ad Hoc CRE Logit outperform the FE-LPM and Ad Hoc FE-LPM. That said, all four estimators do consistently poorly.

Third, 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 (i.e., $\theta_1 = -1.3$). Fourth, 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 of Hausman et al. (1998).

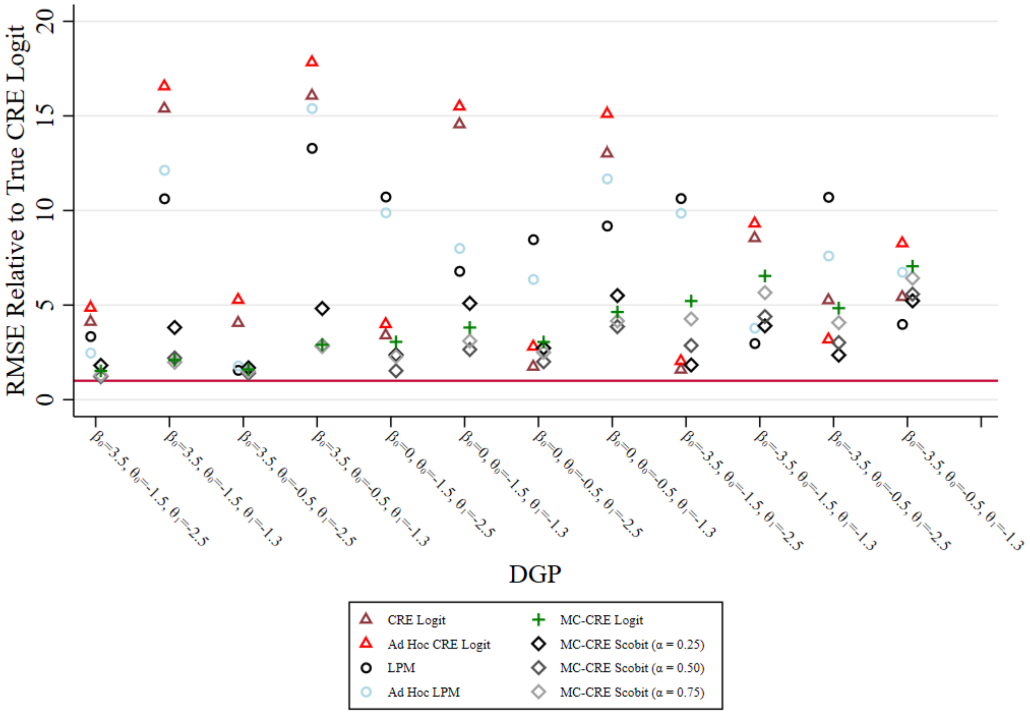
Fifth, the estimators that account for misclassification have much smaller bias overall. In particular, when the proportion of ones is high (i.e., $\beta_0 = 3.5$), the MC-CRE Scobit with α equal to 0.50 or 0.75 tends to produce the smallest bias. As the proportion of ones falls (i.e., $\beta_0 = 0, -3.5$), the MC-CRE Scobit with α equal to 0.25 or 0.50 tends to produce the smallest bias. Finally, the direction of the bias changes with α; the bias is consistently negative for the MC-CRE Logit and transitions to consistently positive when $\alpha = 0.25$.

Table 5: Monte carlo results: AME(d)

Design	True CRE Logit	CRE Logit	Ad Hoc CRE Logit	LPM	Ad Hoc LPM	MC-CRE Logit	MC-CRE Scobit ($\alpha = 0.25$)	MC-CRE Scobit ($\alpha = 0.50$)	MC-CRE Scobit ($\alpha = 0.75$)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
I. Bias									
A. θ_0 = -1.5, and θ_1 = -2.5									
$\beta_0 = 3.5$	-0.191	1.666	1.985	-1.279	-0.864	-0.359	0.599	0.133	-0.170
$\beta_0 = 0$	-0.036	0.813	0.969	-2.670	-2.457	-0.608	0.511	-0.005	-0.368
$\beta_0 = -3.5$	-0.023	0.147	0.236	-1.524	-1.409	-0.565	0.139	-0.188	-0.414
B. θ_0 = -1.5, and θ_1 = -1.3									
$\beta_0 = 3.5$	-0.233	6.519	7.019	4.478	5.122	-0.264	1.418	0.458	-0.010
$\beta_0 = 0$	-0.040	3.596	3.829	1.631	1.939	-0.641	1.134	0.126	-0.366
$\beta_0 = -3.5$	-0.023	1.227	1.340	0.335	0.478	-0.625	0.352	-0.178	-0.459
C. θ_0 = -0.5, and θ_1 = -2.5									
$\beta_0 = 3.5$	-0.216	1.656	2.196	-0.214	0.452	-0.273	0.431	0.029	-0.164
$\beta_0 = 0$	-0.031	0.140	0.565	-2.044	-1.495	-0.498	0.427	0.002	-0.300
$\beta_0 = -3.5$	-0.026	-0.676	-0.272	-1.500	-1.013	-0.456	0.157	-0.135	-0.327
D. θ_0 = -0.5, and θ_1 = -1.3									
$\beta_0 = 3.5$	-0.222	6.768	7.515	5.582	6.476	-0.123	1.608	0.324	0.008
$\beta_0 = 0$	-0.036	3.233	3.761	2.239	2.879	-0.543	1.097	0.159	-0.290
$\beta_0 = -3.5$	-0.024	0.689	1.143	0.364	0.878	-0.458	0.391	-0.070	-0.313
II. Root Mean Squared Error									
A. θ_0 = -1.5, and θ_1 = -2.5									
$\beta_0 = 3.5$	0.418	1.721	2.032	1.401	1.036	0.638	0.759	0.508	0.540
$\beta_0 = 0$	0.252	0.858	1.007	2.705	2.495	0.771	0.598	0.386	0.577
$\beta_0 = -3.5$	0.146	0.232	0.298	1.554	1.441	0.762	0.268	0.418	0.623
B. θ_0 = -1.5, and θ_1 = -1.3									
$\beta_0 = 3.5$	0.425	6.534	7.033	4.515	5.154	0.891	1.621	0.936	0.840
$\beta_0 = 0$	0.248	3.608	3.841	1.685	1.984	0.945	1.262	0.657	0.770
$\beta_0 = -3.5$	0.145	1.242	1.354	0.433	0.552	0.951	0.568	0.639	0.823
C. θ_0 = -0.5, and θ_1 = -2.5									
$\beta_0 = 3.5$	0.427	1.734	2.253	0.669	0.766	0.683	0.719	0.598	0.637
$\beta_0 = 0$	0.250	0.436	0.701	2.117	1.591	0.762	0.682	0.501	0.625
$\beta_0 = -3.5$	0.147	0.772	0.467	1.572	1.117	0.710	0.346	0.443	0.598
D. θ_0 = -0.5, and θ_1 = -1.3									
$\beta_0 = 3.5$	0.422	6.789	7.535	5.619	6.507	1.226	2.036	1.206	1.189
$\beta_0 = 0$	0.251	3.264	3.789	2.305	2.930	1.162	1.380	0.968	1.043
$\beta_0 = -3.5$	0.147	0.797	1.213	0.586	0.990	1.036	0.767	0.819	0.943

Notes: Results based on 500 repetitions. Bias and Root Mean Squared Error are each multiplied by 100. All models include include d, x_1, and x_2 as covariate. Ad hoc models include z as an additional covariate. True Logit uses correctly measured outcomes; all other models use misclassified outcomes. CRE = Correlated Random Effects. LPM = Linear Probability Model. See text for further details.

Figure 3: Monte carlo results: AME(d)



Notes: Markers show the ratio of RMSE of the indicated model to the RMSE of the CRE Logit model using the data free of measurement error. Values of β_0 decrease along the x-axis, resulting in lower presence of ones. Within each value for β_0, θ_0 increases also decreases from left to right, decreasing the rate of false positives. Within values of β_0 and θ_0, values of θ_1 decreases from left to right, lowering the rate of false negatives.

In terms of RMSE of the estimators ignoring misclassification – CRE Logit, Ad Hoc Logit, LPM, and Ad Hoc LPM – several findings emerge. First, the performance of all four estimators is quite poor when the proportion of ones in the data is reasonable (i.e., $\beta_0 = 3.5, 0$). For example, with a high proportion of ones and a high degree of misclassification (i.e., $\beta_0 = 3.5$, $\theta_0 = -0.5$, $\theta_1 = -1.3$), 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 3). Second, when the proportion of ones in the data is quite low (i.e., $\beta_0 = -3.5$), the performance of the four estimators is not as poor and, occasionally, quite good.

Third, the relative performance of the estimators that do not account for misclassification 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 (i.e., $\beta_0 = 3.5$), the LPM estimators always dominate the CRE Logit estimators. However, as the outcome becomes more rare (i.e., $\beta_0 = 0, -3.5$), the LPM estimators continue to dominate the CRE Logit estimators only when the false negative rate is very high (i.e., $\theta_1 = -1.3$).

Fourth, the "ad hoc" estimators do not consistently perform better than their counterparts that do not control for z. The RMSE of the Ad Hoc FE-LPM is smaller than that of the FE-LPM in 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 common practice to control for environmental variables thought to affect the reliability of remotely sensed outcomes, this is in no way a cure for the measurement error induced.

Regarding the estimators addressing misclassification, again several results stand out. First, the MC estimators generally outperform the estimators ignoring misclassification, often quite substantially. For example, in the DGP referenced above with a high proportion of ones and a high degree of misclassification (i.e., $\beta_0 = 3.5$, $\theta_0 = -0.5$, $\theta_1 = -1.3$), the relative RMSE of all four estimators is below five; below three for the MC-CRE Logit and MC-CRE Scobit with $\alpha = 0.50, 0.75$. Second, the performance of the MC-CRE Logit and MC-CRE Scobit with α equal to 0.50 or 0.75 deteriorate as the proportion of ones in the data fall. However, the performance of the MC-CRE Scobit with α equal to 0.25 varies non-monotonically with the proportion of ones in the data.

Third, the relative performance of the estimators depends critically on the proportion of ones in the data. When the proportion is relatively high (i.e., $\beta_0 = 3.5$), the MC-CRE Logit and MC-CRE Scobit with α equal to 0.50 or 0.75 performs best. When the proportion of ones is modest (i.e., $\beta_0 = 0$), then the MC-CRE Scobit with α equal to 0.50 marginally outperforms the other estimators. Lastly, when the proportions of ones is quite low (i.e., $\beta_0 = -3.5$), then the MC Scobit with α equal to 0.25 consistently perform best.

A final comment is warranted as it pertains to the results from the simulations involving rare events (i.e., $\beta_0 = -3.5$). As mentioned previously, the estimators ignoring misclassification occasionally perform very well in this instance, while the performance of MC-CRE Logit and MC-CRE Scobit with $\alpha = 0.50$ or 0.75 decline relative to their performance in DGPs with a higher proportion of ones. That said, in our view, the results do not suggest reliance on the estimators ignoring misclassification when analyzing rare events. We reach this conclusion because of the extreme variation in performance of individual estimators across the DGPs considered. In contrast, the MC-CRE Scobit with $\alpha = 0.25$ consistently performs well across all DGPs considered when the outcome measures a rare event. For example, when the false negative rate is very high (i.e., $\theta_1 = 0, -1.3$), then the LPM performs very well but the CRE Logit can perform very poorly. However, when the false negative rate is relatively low (i.e., $\theta_1 = -2.5$), the FE-LPM performs poorly while the CRE Logit or the Ad Hoc CRE Logit may perform well. As such, the volatility in the performance of the estimators ignoring misclassification suggests they should not be relied upon by researchers.

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. Since the true proportion of ones in the data is unknown in the presence of misclassification, the results suggest estimating a range of MC-CRE Scobit models to complement the MC Logit model. When the true proportion of ones is suspected to be quite low, the MC-CRE Scobit estimator with a low value of α is recommended. 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.

5 Application

5.1 Description

In this section we use a dataset that combines administrative information on properties that applied to Mexico's Payments for Hydrological Services program between 2003 and 2015. This program is part of a broader national system of payments for environmental services that is run by Mexico's National Forestry Commission (CONAFOR, for its acronym in Spanish). 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. They last 5 years and payments 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.

The unit of analysis in our application 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 polygons: one rejected in 2010, one rejected in 2010 and accepted in 2011, and one accepted in 2011. Figure A3 shows a visual representation of these units within various communities with repeated applications. We limit polygons to those between 20 and 2000 hectares. 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 boundaries, since the program did not accept applications greater than 2000 hectares per landholder, except in the case of specially negotiated contracts that are not subject to the usual program rules.

Within each of these polygons we calculate a number of covariates that are associated with land use 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. 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 here is simply to illustrate how our proposed measurement error solution affects estimation results, not to perfectly estimate the probability of deforestation or program impacts.

The deforestation and baseline forest area measures come from Hansen et al. (2013), version 1.2 (accessed in 2016). In section 2 we use the layer of this data that provides a measure of forest cover in 2010. For the application we use the annual loss of forest cover from the same data. The annual forest cover loss does not come from a difference in levels of measured forest, but rather from a time-series analysis that detects disturbance of pixels assessed as having forest in 2000. This data is the only available source with annual variation in land-cover during our period of study (2000-2015). We define an indicator equal to one if any deforestation took place in a polygon within a given year.

The deforestation data were intended for global/regional change analysis, not change analysis at the small parcel-level. It has been shown that the accuracy of the classification algorithm varies across different countries and ecosystem types. For example, assessments by the CONAFOR remote sensing team suggest that the Hansen product offers better results in Mexico when the percentages of forest cover are below 30 or above 60 percent. 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 understate selective logging–an important source of forest degradation –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 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 data contain all applicants, including those that did not end up receiving payments from the program. The screening process for applications requires multiple steps. First, applications have always been 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 these criteria have increased over time, for the duration of the program variables used for targeting have included measures of environmental quality (forest type and location in particular water-scarce areas), opportunity cost (deforestation risk as determined by geographic factors), and social criterion (location in marginalized or indigenous municipalities) (Sims et al. 2014, Alix-Garcia et al. 2019).

Table 6 shows that geographic variables may be correlated with beneficiary status. In particular, those that were integrated into the program tend to be at lower elevation, slightly lower slope, closer to roads and cities, with higher baseline forest cover, and in municipalities with greater indigenous presence. In addition, beneficiary land is often located in Landsat footprints with fewer scenes from both the Landsat 5 and Landsat 7 sensors. Finally, over the entire period, the average percentage of deforested land within the polygons is quite small – between 0.07 and 0.09 percent – and the percent with any deforestation over the entire analysis period is 33 percent for beneficiary parcels and 29 percent for non-beneficiary parcels.

5.2 Results

Table 7 shows the results from the models that ignore misclassification. FE-LPM models include municipality fixed effects and CRE models augment the model with municipality-level means of the covariates. Coefficient estimates are reported for the LPM models, while AMEs are reported for the remaining models. Standard errors are clustered at the level of the municipality.

Table 6: Weighted means of covariates according to beneficiary status

	Non-beneficiary land	Beneficiary land	Norm diff
	(1)	(2)	(3)
Percent polygon deforested, Hansen	0.073	0.090	0.026
Any deforestation, Hansen	0.294	0.334	0.061
Percent forested, 2000	0.593	0.745	0.345
Average Elevation (mt)	1669.997	1527.920	-0.111
Average Slope (degree)	15.972	15.827	-0.014
Distance to any road (meters)	5636.097	4649.078	-0.139
Distance to city with $>$ 5,000 people	33.834	30.424	-0.109
Percent of majority indigenous	0.228	0.343	0.184
Mean cloud-free L5 scenes, 2000-2015	5.090	3.591	-0.245
Mean cloud-free L7 scenes, 2000-2015	9.663	9.036	-0.182
Observations	10966	9959	20925

The sample is divided into those parcels of land that ever received a PES payment and those that applied but were rejected. Columns (1) and (2) show area-weighted means and column (3) the area-weighted normalized difference in means.

Two findings stand out. First, the FE-LPM and Ad Hoc FE-LPM point estimates on the treatment effect for program beneficiary are the smallest in magnitude and of only marginal statistical significance. The treatment effect is the largest (nearly three times the size of the LPM point estimates) for the Ad Hoc CRE Scobit. All point estimates suggest that beneficiary status decreases the probability of deforestation.

Second, the estimated effects of the remaining covariates are qualitatively similar across the various estimators. The three minor exceptions are the smaller and marginally statistically significant effects of slope in the Ad Hoc Scobit models with α equal to 0.50 or 0.75, the smaller effect of the percent deforested in 2000 according to the LPM models, and the fact that the effect of the percent of majority indigenous is statistically significant only in the Ad Hoc Scobit models (although the point estimates are similar across all estimators).

Table 8 displays results from the models that account for misclassification error, and include estimates of the proportion of false positives (G_0) and false negatives (G_1). The odd-numbered columns allow the probability of a false negative to depend on the number of L7 cloud-free scenes, whereas the even numbered columns also allow the probability to depend on the interaction between this variable and the average elevation and slope of the cell.

Before discussing the AMEs, Figure 4 plots the density and cumulative density of the observation-specific estimates of false positive and false negative probabilities from the MC-CRE Logit (panel (a)) and false negative probabilities from the MC-CRE Logit and MC-CRE Scobit (panel (b)). The estimates come from the even-numbered columns in Table 8. Panel (a) shows that the estimated probabilities of a false positive are all very close to zero. The estimated probabilities of a false neg-

Table 7: Results from models ignoring misclassification

		Ad Hoc		Ad Hoc						
	FE-LPM	FE-LPM	CRE Logit	CRE Logit		CRE Scobit			Ad Hoc CRE Scobit	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Beneficiary (0/1)	-0.005*	-0.005	-0.008***	-0.007**	-0.006**	-0.007**	-0.008***	-0.012***	-0.013***	-0.014***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)
Average Elevation (mt)	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Average Slope (degree)	-0.183***	-0.429***	-0.180***	-0.182***	-0.193***	-0.187***	-0.183***	-0.116**	-0.098*	-0.089
	(0.050)	(0.096)	(0.051)	(0.050)	(0.047)	(0.050)	(0.051)	(0.055)	(0.059)	(0.060)
Distance to any road (meters)	-0.005***	-0.005***	-0.007***	-0.007***	-0.006***	-0.006***	-0.007***	-0.008***	-0.009***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Distance to city with > 5,000 people	-0.019	-0.020	-0.016	-0.018	-0.011	-0.014	-0.015	-0.030*	-0.031*	-0.030
	(0.032)	(0.032)	(0.028)	(0.029)	(0.028)	(0.029)	(0.028)	(0.018)	(0.019)	(0.019)
Area of polygon	0.030***	0.030***	0.022***	0.022***	0.026***	0.024***	0.023***	0.027***	0.024***	0.024***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Percent forested, 2000	-0.030**	-0.030**	-0.058***	-0.057***	-0.051***	-0.055***	-0.057***	0.054***	0.055***	0.055***
	(0.013)	(0.013)	(0.014)	(0.014)	(0.013)	(0.013)	(0.014)	(0.017)	(0.017)	(0.017)
Percent of majority indigenous	0.052	0.052	0.041	0.040	0.043	0.042	0.041	0.039***	0.041***	0.042***
	(0.037)	(0.036)	(0.031)	(0.030)	(0.031)	(0.031)	(0.031)	(0.010)	(0.010)	(0.010)
Cloud-free scenes, L7		-0.002		0.002***				0.002***	0.002***	0.002***
		(0.002)		(0.001)				(0.001)	(0.001)	(0.001)
Cloud-free scenes, L7 x Avg Slope		0.027***								
		(0.009)								
Cloud-free scenes, L7 x Mean Elev		-0.000								
		(0.000)								
α					0.250	0.500	0.750	0.250	0.500	0.750

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. Time fixed effects included in all models. * p <.10, ** p< .05, *** p<.01.

Table 8: Results from models allowing misclassification

	MC Logit		MC Scobit					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beneficiary (0/1)	-0.014***	-0.014**	-0.008**	-0.008*	-0.010**	-0.010**	-0.011**	-0.011**
	(0.005)	(0.006)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Average Elevation (mt)	-0.001	0.002	-0.001	0.001	-0.001	0.002	-0.001	0.002
	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Average Slope (degree)	-0.357***	-0.545***	-0.255***	-0.444***	-0.297***	-0.506***	-0.316***	-0.528***
	(0.089)	(0.109)	(0.067)	(0.098)	(0.081)	(0.103)	(0.085)	(0.106)
Distance to any road (meters)	-0.012***	-0.012***	-0.007***	-0.007***	-0.008***	-0.009***	-0.009***	-0.010***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Distance to city with > 5,000 people	-0.038	-0.050	-0.021	-0.030	-0.026	-0.037	-0.029	-0.040
	(0.049)	(0.050)	(0.036)	(0.042)	(0.042)	(0.046)	(0.045)	(0.048)
Area of polygon	0.056***	0.073***	0.039***	0.056***	0.046***	0.063***	0.049***	0.065***
	(0.014)	(0.015)	(0.009)	(0.016)	(0.012)	(0.016)	(0.013)	(0.015)
Percent forested, 2000	-0.106***	-0.124***	-0.066***	-0.082***	-0.080***	-0.096***	-0.086***	-0.102***
	(0.029)	(0.031)	(0.020)	(0.025)	(0.024)	(0.027)	(0.025)	(0.027)
Percent of majority indigenous	0.078	0.086	0.060	0.075	0.070	0.083	0.075	0.086
	(0.055)	(0.058)	(0.041)	(0.046)	(0.048)	(0.052)	(0.051)	(0.055)
G0	0.012	0.014	0.000	0.000	0.000	0.000	0.000	0.000
G1	0.442	0.512	0.235	0.378	0.347	0.452	0.388	0.477
α			0.250	0.250	0.500	0.500	0.750	0.750
logL	-113708.547	-113528.702	-113963.483	-113835.277	-113891.760	-113750.065	-113878.813	-113736.208

Column headers indicate estimator. Standard errors are clustered at the municipality level. Marginal effects evaluated at sample means are displayed. $G0$ and $G1$ are the probability of a false positive and negative, respectively, evaluated at sample means. Column 1 allows the false positive rate to depend on the number of L7 cloud-free scenes. Column 2 allows the false positive rate to depend on the number of L7 cloud-free scenes and its interaction with average elevation and average slope. Time fixed effects included in all models. * p <.10, ** p< .05, *** p<.01.

ative, however, are concentrated between 0.4 and 0.6. The average probability of a false negative is 0.512, as reported in Column 2 in Table 8. Thus, roughly half of all cells actually experiencing deforestation are estimated to be classified incorrectly. Since roughly 50,000 observations in our sample, or 17%, are reported to experience deforestation, this suggests that the true number is about 100,000, or 34%. Taking the false positive rate to be zero, this implies that the unconditional probability of being misclassified is roughly 17%. This is in line with the accuracy studies mentioned above.

Panel (b) compares just the false negative rates across the MC-CRE Logit and MC-CRE Scobit models (since the MC-CRE Scobit models impose a zero false positive rate). The density and cumulative density show essentially a first-order stochastic dominance relationship among the distributions. As α declines from one in the logit model to 0.75, 0.50, and 0.25, the entire distribution shifts to the left, indicating lower overall rates of misclassification. The average probabilities of a false negative are 0.477, 0.452, and 0.378, respectively, as reported in Column 8, 6, and 4 in Table 8. Thus, the shape of the link function is important to the estimated misclassification rate. Nonetheless, all four estimators suggest substantial misclassification as the even MC-CRE Scobit with α equal to 0.25 estimates that the unconditional probability of being misclassified is roughly 10%.

Turning to the point estimates, a few findings emerge. First, the observed proportion of outcomes equal to one is 17%. Combining this proportion with the consistently high estimates of the false negative rate suggests that the true proportion of outcomes equal to one is roughly 30%. From our simulation results, this suggests that the MC-CRE Logit and the MC-CRE Scobit models with α equal to 0.50 or 0.75 are the preferred estimators. Comparing the maximized value of the log likelihood functions indicates that the MC-CRE Logit in column 2 fits the data best.[9] Thus, the MC-CRE Logit, allowing for the misclassification rate to depend on the number of L7 cloud-free scenes and its interaction with topography, is the preferred estimator in this application.

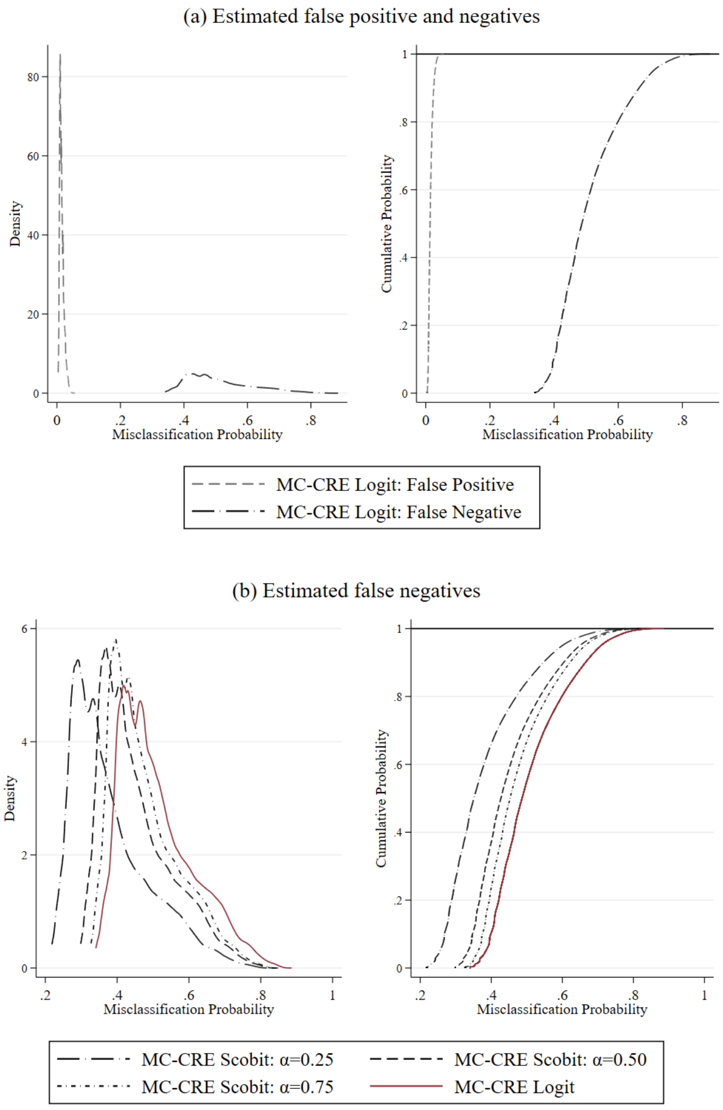
Second, all estimated AMEs of beneficiary status are negative and statistically different from zero, and generally indicate a decrease in the probability of deforestation of 0.01. However, the magnitude does decline monotonically as α declines, with the MC-CRE Logit producing the largest AME (in absolute value) at -0.014. Comparing our preferred estimator, MC-CRE Logit, to the most commonly used estimator in practice, Ad Hoc FE LPM, indicates a substantial effect from addressing misclassification. Specifically, our preferred estimator yields an effect that is nearly three times as large in magnitude, as well as statistically different zero at the $p < 0.01$ level.

Figure 5 provides a detailed comparison of the distribution of the AME of beneficiary status across all estimators excluding the LPMs (since the AME does not vary).[10] Interestingly, not only are the mean and median AME in the MC-CRE Logit larger (in absolute value) than all

[9] A likelihood ratio test easily rejects the MC-CRE Logit in column 1 in favor of the MC-CRE Logit in column 2 at the $p < 0.01$ level.

[10] Figure D1 in Appendix D displays the distributions of AMEs for all covariates.

Figure 4: Distribution of estimated misclassification rates from logit and scobit models

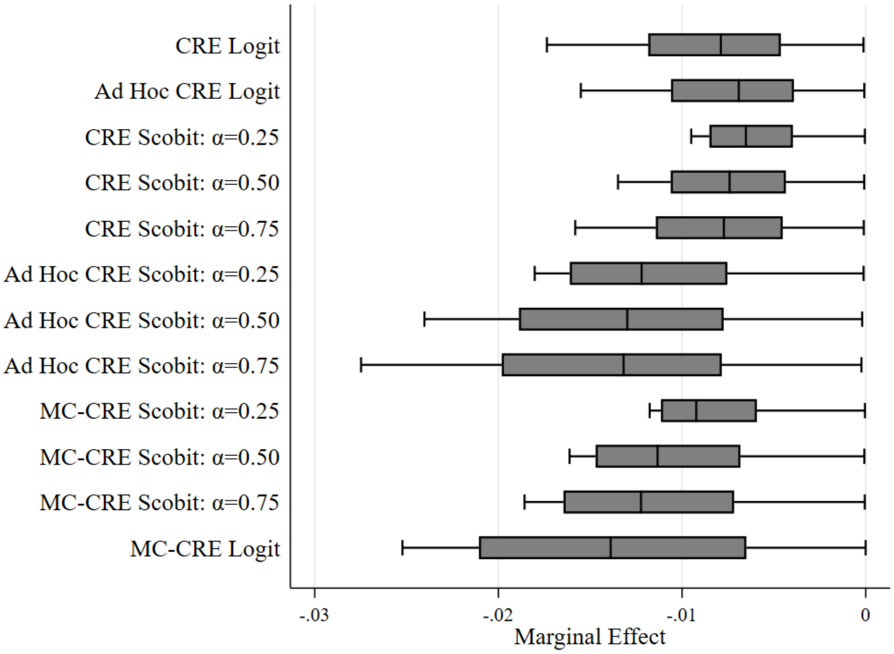


(a) Estimated false positive and negatives

(b) Estimated false negatives

other estimators, the distribution is also quite wide. The 75^{th} percentile of the distribution of the marginal effects is less than -0.02 and maximum is roughly -0.025. This heterogeneity is missed if one instead relies on a LPM.

Figure 5: Distribution of marginal effects of beneficiary status



Notes: Each shaded box spans the interquartile range; mid-line of the box corresponds to the median. Edges of the lines represent the minimum and maximum. Estimates obtained from columns 3-10 in Table 7 and the even-numbered columns in Table 8.

Third, failure to account for measurement error 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 8) to the CRE Logit (column 3 in Table 7), we find the magnitude of the AME is about three times as large for average slope and area of the polygo and twice as large for distance to any road and percent forested in 2000.

Fourth, the AME of average slope varies considerably 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 average elevation). When allowing average slope to affect the misclassification rate, we find a much larger, negative effect of average slope on deforestation. Finally, the AMEs of average slope, distance to any road, the area of the cell, and the percent forested in 2000 follow a similar pattern as beneficiary status. Specifically, the MC-CRE Logit estimates are largest (in absolute value) and decline monotonically as α declines.

In sum, this application confirms that addressing misclassification in the dependent variable can have important impacts on estimation. In this particular case, results from the models usually applied to this type of data are attenuated, both for the treatment variable and for a number of 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 data from research on remote sensing.

6 Conclusion

The opportunities for researchers to exploit remotely sensed data to gain new insights are seemingly infinite. However, to ensure these insights are useful requires researchers to properly understand the nature of this new 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 product. However, 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 in order to understand the construction of the data, and avoid the usage of naïve statistical models that fail to account for the non-classical measurement error that may contaminate the data.

In this paper we have characterized the nature of mismeasurement in commonly used remotely sensed data. Although our focus has been on land use change outcomes, the lessons are generalizable. We have shown how sensor function, ecological attributes, and topographic features can lead to systematic mismeasurement that could be correlated with phenomena whose impacts researchers might be interested in estimating. Moreover, we have shown that binary measures, while perhaps measured with less error, nonetheless guarantee that the errors are correlated with all outcome determinants. Finally, we have demonstrated the feasibility of several estimators when the remotely sensed data is used to create a binary outcome measure such as deforestation. Failure to address misclassification in our analysis of deforestation in Mexico leads to significant attenuation bias. However, our simulation study reveals that attenuation is not guaranteed. With non-classical measurement error, bias in either direction is possible.

While we believe the methods provided here offer a significant advancement in the analysis of remotely sensed data, much work remains to be done. Future work is needed to better understand the nature of the measurement error across different types of remotely sensed data. In particular, investigating and possibly exploiting spatial correlation in measurement error may provide additional avenues through which to minimize the econometric implications of the errors. Future research is also needed to develop useful econometric tools when the outcome is continuous. At the very least, traditional instrumental variable techniques are possible in this case. However, identifying valid exclusion restrictions may be difficult when the measurement error is non-classical. Finally, future work is needed to explore how non-classical measurement error arising from remotely sensed data impacts commonly used program evaluation methods such as difference-in-differences. Our discussion here suggests that measurement error may be serially correlated (e.g., due to reliance on the same algorithm over time or due to time invariant topographic features), but also may contain unconventional trends (e.g., due to deterioration in the quality of a satellite over time). The time series properties of the measurement error may invalidate the parallel trends assumption even if it holds in the absence of measurement error.

References

Abowd, J. M. & Stinson, M. H. (2013), 'Estimating measurement error in annual job earnings: a comparison of survey and administrative data', *Review of Economics and Statistics* **95**(5), 1451–1467.

Alix-Garcia, J. M., Shapiro, E. N. & Sims, K. R. E. (2012), 'Forest Conservation and Slippage: Evidence from Mexico's National Payments for Ecosystem Services Program', *Land Economics* **88**(4), 613–638.
URL: *http://le.uwpress.org/content/88/4/613.short*

Alix-Garcia, J. M., Sims, K. R., Orozco-Olvera, V. H., Costica, L., Medina, J. D. F., Romo-Monroy, S. & Pagiola, S. (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, J. M., Sims, K. R. & Yañez-Pagans, P. (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.

Altonji, J. G., Elder, T. E. & Taber, C. R. (2005), 'Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools', *Journal Political Economy* **113**(1), 151–184.

ARD, Darum, G. & Lund, G. (2002), Mexico - critical analysis of the current deforestation rate estimates, Technical report.

Black, D. A., Berger, M. C. & Scott, F. A. (2000), 'Bounding parameter estimates with nonclassical measurement error', *Journal of the American Statistical Association* **95**(451), 739–748.

Browning, M. & Crossley, T. (2009), 'Are two cheap, noisy measures better than one expensive, accurate one?', *American Economic Review Papers & Proceedings* **99**(2), 99–103.

Burivalova, Z., Bauert, M. R., Hassold, S., Fatroandrianjafinonjasolomiovazo, N. T. & Koh, L. P. (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.

Calabrese, R. & Osmetti, S. A. (2013), 'Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model', *Journal of Applied Statistics* **40**(6), 1172–1188.

Donaldson, D. & Storeygard, A. (2016), 'The view from above: Applications of satellite data in economics', *Journal of Economic Perspectives* **30**(4), 171–98.

Gibson, J. (2020), 'Better Night Lights Data, For Longer', *Oxford Bulletin of Economics and Statistics* .

Gibson, J., Olivia, S., Boe-Gibson, G. & Li, C. (2019), 'Which night lights data should we use in economics, and where?', *Journal of Development Economics* p. 102602.

Goleţ, I. (2014), 'Symmetric and asymmetric binary choice models for corporate bankruptcy', *Procedia - Social and Behavioral Sciences* **124**, 282–291.

Gottschalk, P. & Huynh, M. (2010), 'Are earnings inequality and mobility overstated? The impact of nonclassical measurement error', *Review of Economics and Statistics* **92**(2), 302–315.

Government of Mexico (2011), 'Conjunto de datos vectoriales de uso del suelo y vegetación escala 1:250 000 serie IV Conjunto Nacional', *Instituto Nacional de Geografía e Estadística* .

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* .

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2011), *Survey methodology*, Vol. 561, John Wiley & Sons.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S., Goetz, S. J., Loveland, T. R. et al. (2013), 'High-resolution global maps of 21st-century forest cover change', *Science* **342**(6160), 850–853.

Hausman, J. (2001), 'Mismeasured variables in econometric analysis: problems from the right and problems from the left', *Journal of Economic Perspectives* **15**(4), 57–67.

Hausman, J. A., Abrevaya, J. & Scott-Morton, F. M. (1998), 'Misclassification of the dependent variable in a discrete-response setting', *Journal of Econometrics* **87**(2), 239–269.

Horrace, W. C. & Oaxaca, R. L. (2006), 'Results on the bias and inconsistency of ordinary least squares for the linear probability model', *Economics Letters* **90**(3), 321–327.

Jain, M. (2020), 'The Benefits and Pitfalls of Using Satellite Data for Causal Inference', *Review of Environmental Economics and Policy* **14**(1), 157–169.

Kapteyn, A. & Ypma, J. Y. (2007), 'Measurement error and misclassification: a comparison of survey and administrative data', *Journal of Labor Economics* **25**(1), 513–551.

Kennedy, R. E., Townsend, P. A., Gross, J. E., Cohen, W. B., Bolstad, P., Wang, Y. & Adams, P. (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, G. & Zeng, L. (2001), 'Logistic regression in rare events data', *Political Analysis* **9**(2), 137–163.

Kovalskyy, V. & Roy, D. P. (2013), 'The global availability of landsat 5 tm and landsat 7 etm+ land surface observations and implications for global 30 m landsat data product generation', *Remote Sensing of Environment* **130**, 280–293.

Lancaster, T. (2000), 'The incidental parameter problem since 1948', *Journal of Econometrics* **95**(2), 391–413.

Lewbel, A. (2000), 'Identification of the binary choice model with misclassification', *Econometric Theory* **16**, 603–609.

Li, M., Zang, S., Zhang, B., Li, S. & Wu, C. (2014), 'A review of remote sensing image classification techniques: The role of spatio-contextual information', *European Journal of Remote Sensing* **47**(1), 389–411.

Mankiw, N. G. & Shapiro, M. D. (1986), News or noise? An analysis of GNP revisions, Technical report, National Bureau of Economic Research.

Meyer, B. D., Mok, W. K. & Sullivan, J. X. (2015), 'Household surveys in crisis', *Journal of Economic Perspectives* **29**(4), 199–226.

Mitchard, E., Viergever, K., Morel, V. & Tipper, R. (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'.
URL: *https://ecometrica. com . . .*

Nagler, J. (1994), 'Scobit: an alternative estimator to logit and probit', *American Journal of Political Science* **38**(1), 230–255.

NASA (2021), 'Quotes to note', *Landsat Science website* .
URL: *https://landsat.gsfc.nasa.gov/news/quotes-note?page=17*

Nguimkeu, P., Denteh, A. & Tchernis, R. (2019), 'On the estimation of treatment effects with endogenous misreporting', *Journal of Econometrics* **208**, 487–506.

Papke, L. E. & Wooldridge, J. M. (2008), 'Panel data methods for fractional response variables with an application to test pass rates', *Journal of Econometrics* **145**, 121–133.

Sims, K. R., Alix-Garcia, J., Shapiro-Garza, E., Fine, L. R., Radeloff, V. C., Aronson, G., Castillo, S., Ramirez-Reyes, C. & Yanez-Pagans, P. (2014), 'Improving environmental and social targeting through adaptive management in Mexico's payments for hydrological services program', *Conservation Biology* **28**(5), 1151–1159.

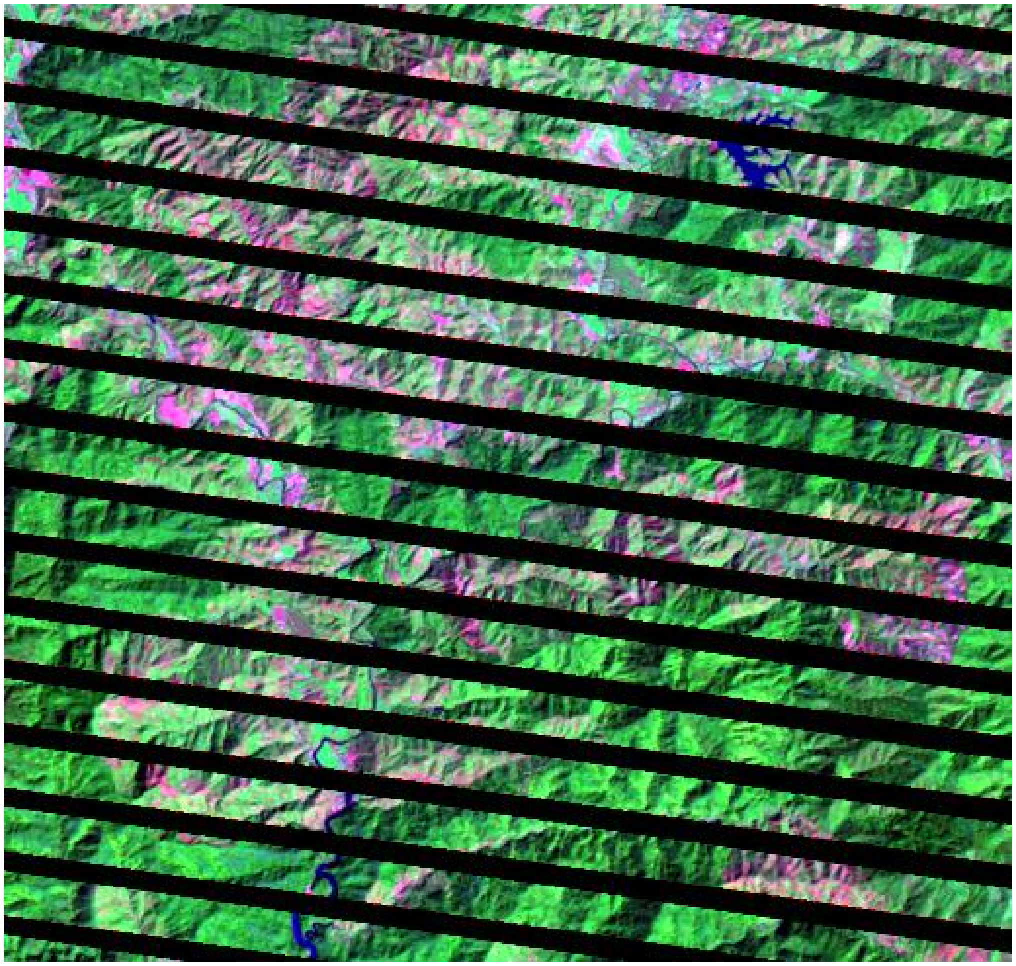
Union of Concerned Scientists (2020), 'Ucs satellite database'.
URL: *https://www.ucsusa.org/resources/satellite-database*

U.S. Geological Service (2021), 'What are the band designations for the landsat satellites?'.
URL: *https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites?*

Young, N. E., Anderson, R. S., Chignell, S. M., Vorster, A. G., Lawrence, R. & Evangelista, P. H. (2017), 'A survival guide to Landsat preprocessing', *Ecology* **98**(4), 920–932.

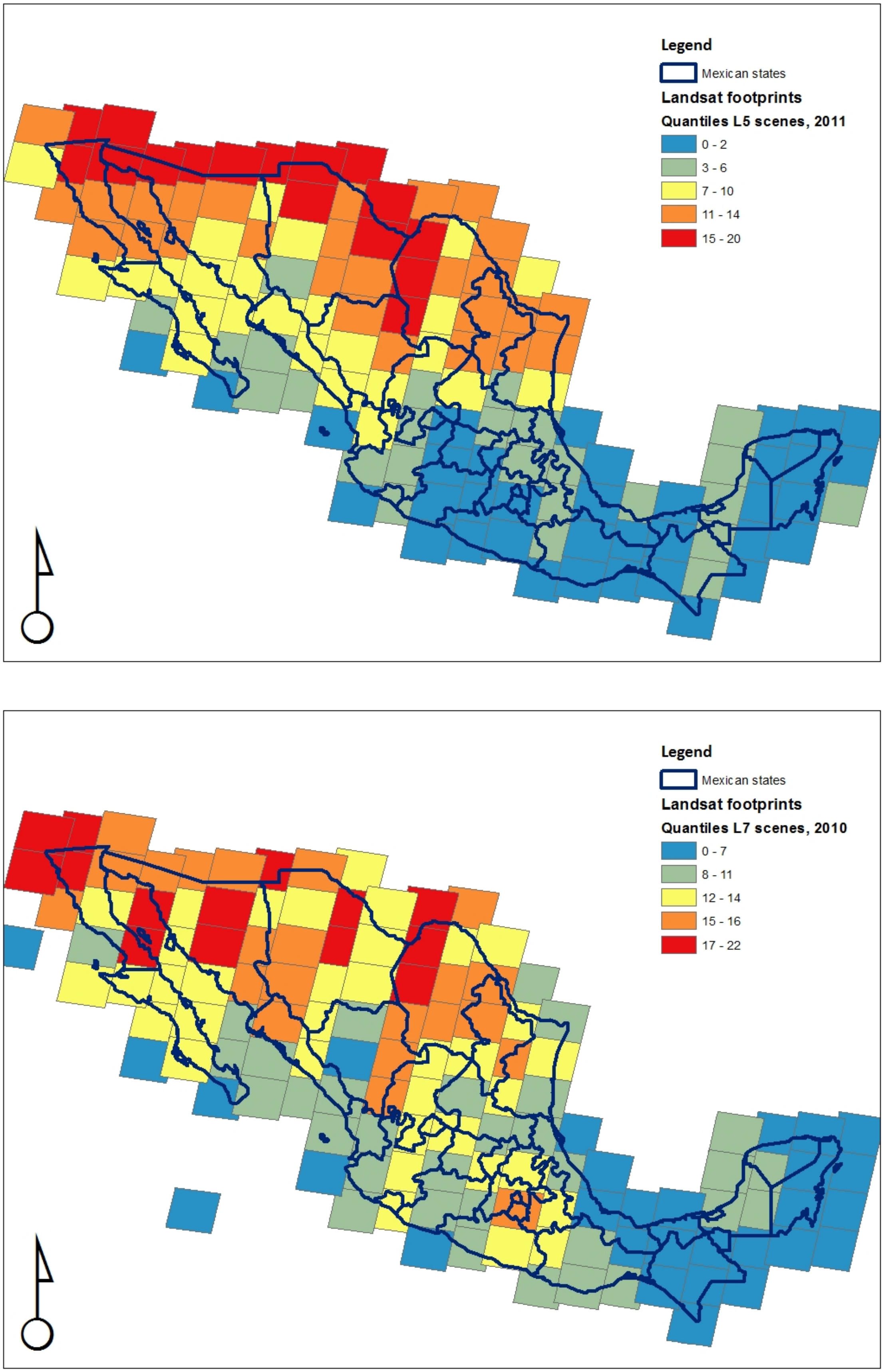
Appendix A Supplemental remote sensing figures

Figure A1: Example of scanline error



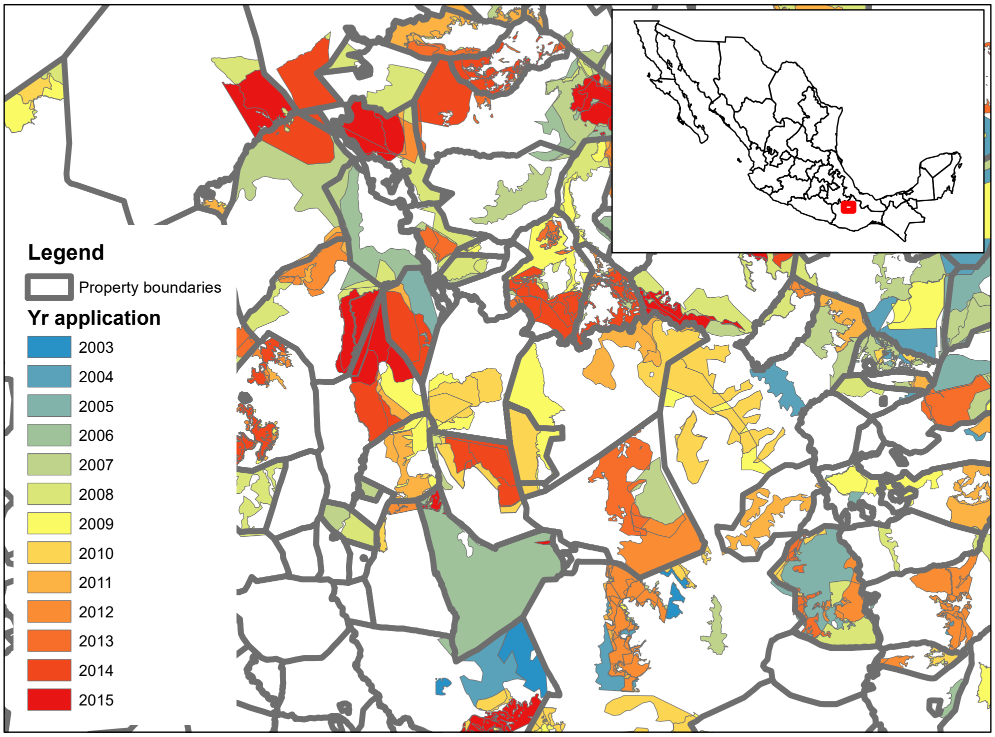
Source: Yale University Center for Earth Observation

Figure A2: Distribution of cloud-free Landsat 5 and 7 scenes across Mexico in 2011 and 2010



Shading of footprints indicates number of scenes available in 2011 (Landsat 5) and 2010 (Landsat 7) with less than 25% cloud cover. Warmer colors indicate more scenes, and shading is by quantiles of the number of available scenes.

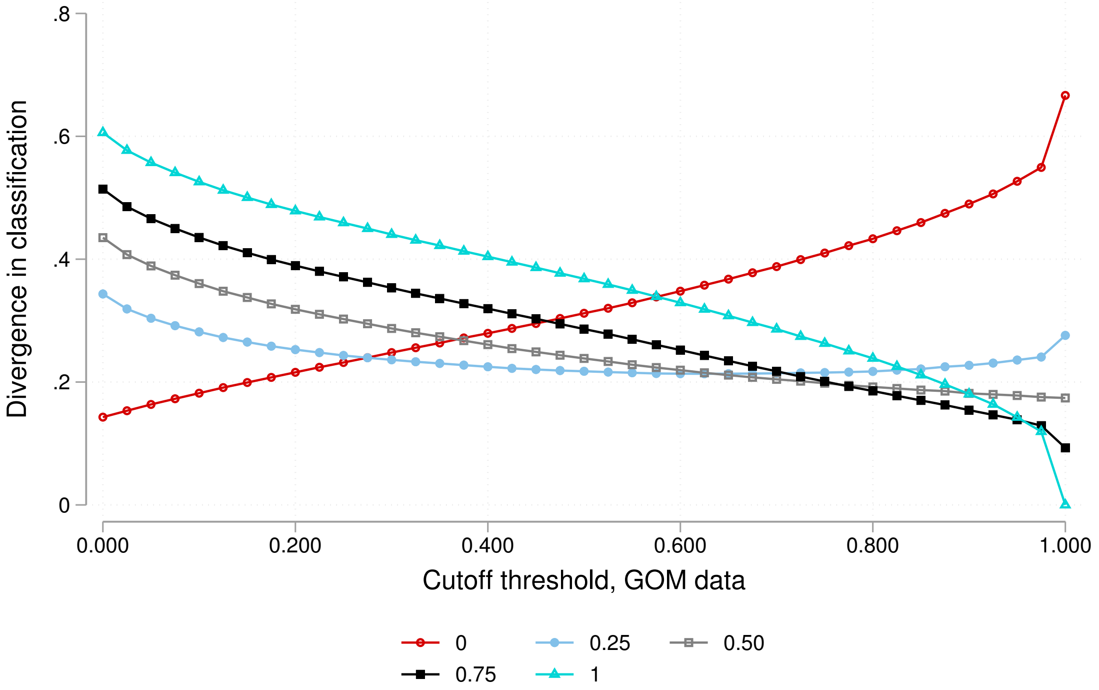
Figure A3: Example of properties with multiple applications



Dark gray boundaries indicate common property borders. Polygons within properties are shaded to indicate their most recent year of application to the PES program. The red rectangle in the inset map indicates the location of the detailed polygons within Mexico.

Appendix B Supplemental results comparing classification with varying cutoff thresholds

Figure B4: Disagreement in classification by cutoff



The y-axis measures the proportion of cells with disagreement in classification across the GOM and Hansen data. The x-axis measures different cutoff thresholds for defining a grid cell as forested in the GOM dataset. The different lines represent different cutoff thresholds for defining a grid cell as forested by the Hansen dataset.

Table B1: Correlations between disagreement in classification and goegraphic covariates

	Cutoff threshold (proportion cell forested)						
	0	0.05	0.10	0.25	0.50	0.75	0.90
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean elevation, 1000s m	-0.061***	0.108***	0.109***	0.087***	0.038***	-0.026***	-0.036***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Mean slope	-0.334***	-0.463***	-0.384***	-0.076***	0.312***	0.497***	0.445***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Dry tropical biome	-0.099***	0.368***	0.439***	0.485***	0.419***	0.253***	0.147***
	(0.006)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.005)
Moist tropical biome	-0.086***	0.250***	0.279***	0.280***	0.236***	0.124***	0.061***
	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)
Pine/oak biome	-0.088***	0.299***	0.366***	0.419***	0.381***	0.293***	0.229***
	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Mangrove biome	-0.046***	0.130***	0.148***	0.156***	0.137***	0.078***	0.035***
	(0.010)	(0.013)	(0.014)	(0.013)	(0.011)	(0.009)	(0.007)
Minimum of Landsat scenes, 2010-2011	-0.246***	-0.137***	-0.088***	-0.044***	-0.000	-0.008	-0.041***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
Minimum of Landsat scenes x slope mean	0.240***	0.360***	0.324***	0.176***	-0.011	-0.055***	-0.003
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mean DV	0.143	0.202	0.222	0.243	0.238	0.201	0.147

Column headers indicate dependent variables. The unit of analysis is 5 x 5 km grid cells across the entire landscape of Mexico. State fixed effects are included. Standard errors are robust, and the estimator is OLS. Beta coefficients are displayed. * p <.10, ** p< .05, *** p<.01.

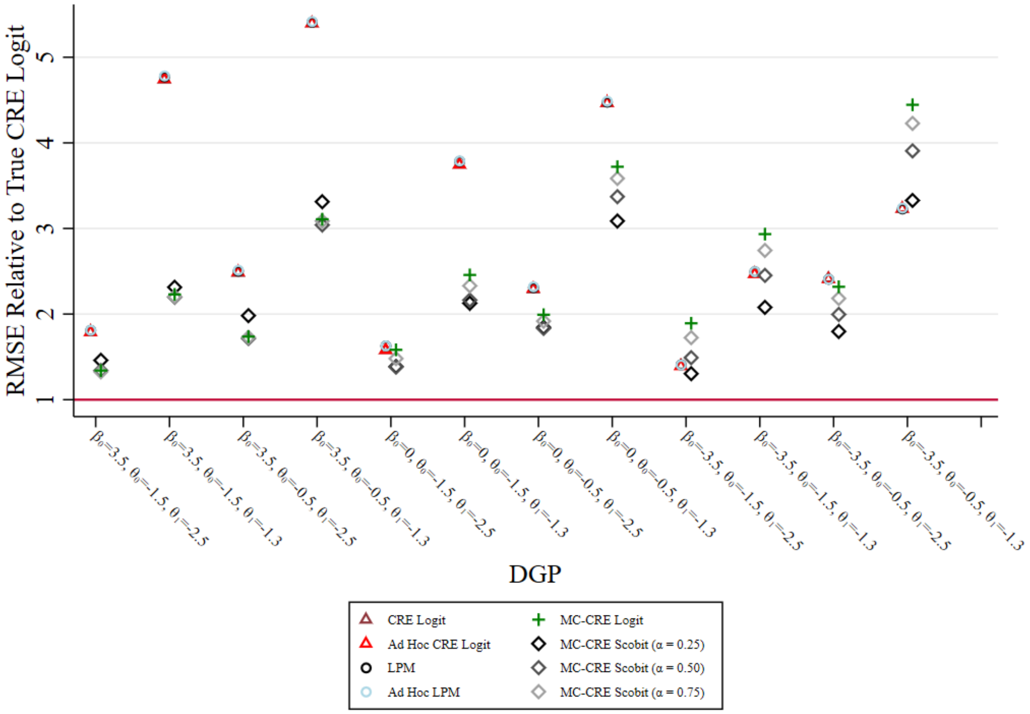
Appendix C Supplemental simulation results

Table C2: Monte carlo results: $\beta_0 = 3.5$

Design	True CRE Logit (1)	CRE Logit (2)	Ad Hoc CRE Logit (3)	LPM (4)	Ad Hoc LPM (5)	MC-CRE Logit (6)	MC-CRE Scobit ($\alpha = 0.25$) (7)	MC-CRE Scobit ($\alpha = 0.50$) (8)	MC-CRE Scobit ($\alpha = 0.75$) (9)
I. Bias									
A. θ_0 = -1.5, and θ_1 = -2.5									
AME(x_1)	-0.012	3.276	3.304	3.390	3.409	-0.041	1.751	0.878	0.349
AME(x_2)	-0.005	0.108	0.107	0.125	0.123	-0.004	0.058	0.028	0.010
AME(d)	-0.191	1.666	1.985	-1.279	-0.864	-0.359	0.599	0.133	-0.170
B. θ_0 = -1.5, and θ_1 = -1.3									
AME(x_1)	-0.011	10.877	10.913	10.970	11.012	0.014	2.756	1.248	0.475
AME(x_2)	-0.005	0.368	0.366	0.380	0.378	0.008	0.098	0.049	0.023
AME(d)	-0.233	6.519	7.019	4.478	5.122	-0.264	1.418	0.458	-0.010
C. θ_0 = -0.5, and θ_1 = -2.5									
AME(x_1)	-0.049	4.682	4.726	4.764	4.811	0.089	1.770	0.430	0.219
AME(x_2)	-0.006	0.170	0.169	0.179	0.178	0.012	0.068	0.022	0.016
AME(d)	-0.216	1.656	2.196	-0.214	0.452	-0.273	0.431	0.029	-0.164
D. θ_0 = -0.5, and θ_1 = -1.3									
AME(x_1)	-0.043	12.460	12.497	12.509	12.555	-0.047	3.673	0.493	0.072
AME(x_2)	-0.005	0.431	0.428	0.436	0.434	0.022	0.144	0.039	0.025
AME(d)	-0.222	6.768	7.515	5.582	6.476	-0.123	1.608	0.324	0.008
II. Root Mean Squared Error									
A. θ_0 = -1.5, and θ_1 = -2.5									
AME(x_1)	2.439	4.361	4.366	4.423	4.425	3.270	3.561	3.268	3.234
AME(x_2)	0.019	0.111	0.109	0.126	0.125	0.026	0.066	0.037	0.027
AME(d)	0.418	1.721	2.032	1.401	1.036	0.638	0.759	0.508	0.540
B. θ_0 = -1.5, and θ_1 = -1.3									
AME(x_1)	2.373	11.249	11.266	11.329	11.354	5.292	5.491	5.213	5.230
AME(x_2)	0.020	0.369	0.367	0.380	0.378	0.043	0.108	0.065	0.049
AME(d)	0.425	6.534	7.033	4.515	5.154	0.891	1.621	0.936	0.840
C. θ_0 = -0.5, and θ_1 = -2.5									
AME(x_1)	2.317	5.752	5.775	5.804	5.828	4.025	4.592	3.974	4.002
AME(x_2)	0.020	0.172	0.170	0.181	0.179	0.035	0.094	0.039	0.036
AME(d)	0.427	1.734	2.253	0.669	0.766	0.683	0.719	0.598	0.637
D. θ_0 = -0.5, and θ_1 = -1.3									
AME(x_1)	2.393	12.911	12.917	12.950	12.968	7.437	7.925	7.275	7.386
AME(x_2)	0.020	0.432	0.429	0.437	0.435	0.067	0.184	0.071	0.066
AME(d)	0.422	6.789	7.535	5.619	6.507	1.226	2.036	1.206	1.189

Results based on 500 repetitions. Bias and Root Mean Squared Error are each multiplied by 100. All models include include d, x_1, and x_2 as covariates. Ad hoc models include z as an additional covariate. True Logit uses correctly measured outcomes; all other models use misclassified outcomes. CRE = Correlated Random Effects. LPM = Linear Probability Model. See text for further details.

Figure C5: Monte carlo results: AME(x_1)



Notes: Markers show the ratio of RMSE of the indicated model to the RMSE of the CRE Logit model using the data free of measurement error. Values of β_0 decrease along the x-axis, resulting in lower presence of ones. Within each value for β_0, θ_0 increases also decreases from left to right, decreasing the rate of false positions. Within values of β_0 and θ_0, values of θ_1 decreases from left to right, lowering the rate of false negatives.

Table C3: Monte carlo results: $\beta_0 = 0$

Design	True CRE Logit	CRE Logit	Ad Hoc CRE Logit	LPM	Ad Hoc LPM	MC-CRE Logit	MC-CRE Scobit ($\alpha = 0.25$)	MC-CRE Scobit ($\alpha = 0.50$)	MC-CRE Scobit ($\alpha = 0.75$)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
I. Bias									
A. θ_0 = -1.5, and θ_1 = -2.5									
AME(x_1)	0.020	1.947	1.964	2.097	2.111	-0.854	1.001	0.162	-0.447
AME(x_2)	-0.002	0.068	0.068	0.087	0.087	-0.025	0.036	0.009	-0.011
AME(d)	-0.036	0.813	0.969	-2.670	-2.457	-0.608	0.511	-0.005	-0.368
B. θ_0 = -1.5, and θ_1 = -1.3									
AME(x_1)	0.047	6.662	6.679	6.752	6.772	-1.095	1.873	0.184	-0.636
AME(x_2)	-0.002	0.228	0.227	0.239	0.238	-0.026	0.071	0.016	-0.011
AME(d)	-0.040	3.596	3.829	1.631	1.939	-0.641	1.134	0.126	-0.366
C. θ_0 = -0.5, and θ_1 = -2.5									
AME(x_1)	-0.009	2.955	2.983	3.037	3.069	-0.669	0.955	0.093	-0.365
AME(x_2)	-0.001	0.112	0.111	0.121	0.120	-0.018	0.036	0.008	-0.008
AME(d)	-0.031	0.140	0.565	-2.044	-1.495	-0.498	0.427	0.002	-0.300
D. θ_0 = -0.5, and θ_1 = -1.3									
AME(x_1)	0.046	7.598	7.632	7.636	7.678	-1.124	1.619	0.027	-0.713
AME(x_2)	-0.002	0.269	0.267	0.272	0.271	-0.022	0.068	0.015	-0.009
AME(d)	-0.036	3.233	3.761	2.239	2.879	-0.543	1.097	0.159	-0.290
II. Root Mean Squared Error									
A. θ_0 = -1.5, and θ_1 = -2.5									
AME(x_1)	1.841	2.909	2.912	3.002	3.004	2.913	2.555	2.543	2.725
AME(x_2)	0.015	0.070	0.070	0.089	0.088	0.035	0.040	0.023	0.026
AME(d)	0.252	0.858	1.007	2.705	2.495	0.771	0.598	0.386	0.577
B. θ_0 = -1.5, and θ_1 = -1.3									
AME(x_1)	1.861	6.970	6.974	7.046	7.056	4.573	3.956	4.027	4.336
AME(x_2)	0.015	0.228	0.228	0.239	0.238	0.048	0.078	0.041	0.041
AME(d)	0.248	3.608	3.841	1.685	1.984	0.945	1.262	0.657	0.770
C. θ_0 = -0.5, and θ_1 = -2.5									
AME(x_1)	1.835	4.201	4.219	4.243	4.260	3.655	3.386	3.363	3.518
AME(x_2)	0.015	0.115	0.113	0.123	0.122	0.036	0.048	0.029	0.031
AME(d)	0.250	0.436	0.701	2.117	1.591	0.762	0.682	0.501	0.625
D. θ_0 = -0.5, and θ_1 = -1.3									
AME(x_1)	1.817	8.111	8.120	8.137	8.156	6.761	5.608	6.125	6.512
AME(x_2)	0.015	0.270	0.268	0.273	0.271	0.066	0.085	0.060	0.061
AME(d)	0.251	3.264	3.789	2.305	2.930	1.162	1.380	0.968	1.043

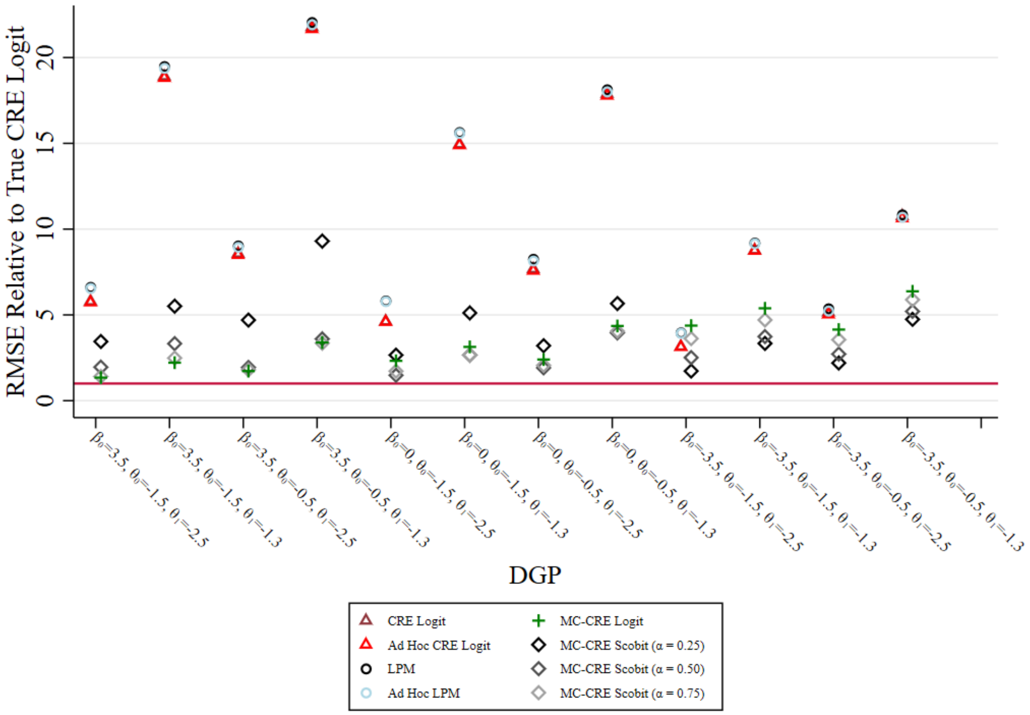
Results based on 500 repetitions. Bias and Root Mean Squared Error are each multiplied by 100. All models include include d, x_1, and x_2 as covariates. Ad hoc models include z as an additional covariate. True Logit uses correctly measured outcomes; all other models use misclassified outcomes. CRE = Correlated Random Effects. LPM = Linear Probability Model. See text for further details.

Table C4: Monte carlo results: β_0 = -3.5

Design	True CRE Logit	CRE Logit	Ad Hoc CRE Logit	LPM	Ad Hoc LPM	MC-CRE Logit	MC-CRE Scobit	MC-CRE Scobit	MC-CRE Scobit
							($\alpha = 0.25$)	($\alpha = 0.50$)	($\alpha = 0.75$)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
I. Bias									
A. θ_0 = -1.5, and θ_1 = -2.5									
AME(x_1)	-0.023	0.731	0.733	0.807	0.813	-0.926	0.271	-0.285	-0.670
AME(x_2)	0.000	0.029	0.029	0.038	0.038	-0.030	0.009	-0.009	-0.022
AME(d)	-0.023	0.147	0.236	-1.524	-1.409	-0.565	0.139	-0.188	-0.414
B. θ_0 = -1.5, and θ_1 = -1.3									
AME(x_1)	-0.027	2.603	2.603	2.650	2.656	-1.013	0.653	-0.256	-0.732
AME(x_2)	0.000	0.090	0.089	0.094	0.094	-0.035	0.020	-0.010	-0.025
AME(d)	-0.023	1.227	1.340	0.335	0.478	-0.625	0.352	-0.178	-0.459
C. θ_0 = -0.5, and θ_1 = -2.5									
AME(x_1)	-0.014	1.134	1.159	1.163	1.193	-0.727	0.272	-0.200	-0.515
AME(x_2)	0.000	0.049	0.048	0.052	0.050	-0.025	0.009	-0.007	-0.018
AME(d)	-0.026	-0.676	-0.272	-1.500	-1.013	-0.456	0.157	-0.135	-0.327
D. θ_0 = -0.5, and θ_1 = -1.3									
AME(x_1)	-0.005	2.923	2.948	2.939	2.972	-0.849	0.651	-0.161	-0.601
AME(x_2)	0.000	0.107	0.106	0.108	0.107	-0.026	0.022	-0.004	-0.018
AME(d)	-0.024	0.689	1.143	0.364	0.878	-0.458	0.391	-0.070	-0.313
II. Root Mean Squared Error									
A. θ_0 = -1.5, and θ_1 = -2.5									
AME(x_1)	1.182	1.651	1.645	1.667	1.664	2.238	1.542	1.763	2.039
AME(x_2)	0.010	0.032	0.031	0.040	0.039	0.044	0.017	0.025	0.036
AME(d)	0.146	0.232	0.298	1.554	1.441	0.762	0.268	0.418	0.623
B. θ_0 = -1.5, and θ_1 = -1.3									
AME(x_1)	1.199	2.963	2.959	2.994	2.995	3.518	2.492	2.940	3.289
AME(x_2)	0.010	0.091	0.090	0.095	0.095	0.056	0.034	0.039	0.049
AME(d)	0.145	1.242	1.354	0.433	0.552	0.951	0.568	0.639	0.823
C. θ_0 = -0.5, and θ_1 = -2.5									
AME(x_1)	1.192	2.879	2.868	2.872	2.872	2.763	2.141	2.379	2.601
AME(x_2)	0.010	0.053	0.051	0.055	0.054	0.042	0.022	0.028	0.036
AME(d)	0.147	0.772	0.467	1.572	1.117	0.710	0.346	0.443	0.598
D. θ_0 = -0.5, and θ_1 = -1.3									
AME(x_1)	1.209	3.903	3.908	3.906	3.919	5.373	4.022	4.722	5.110
AME(x_2)	0.010	0.109	0.108	0.110	0.108	0.065	0.048	0.053	0.060
AME(d)	0.147	0.797	1.213	0.586	0.990	1.036	0.767	0.819	0.943

Results based on 500 repetitions. Bias and Root Mean Squared Error are each multiplied by 100. All models include include d, x_1, and x_2 as covariates. Ad hoc models include z as an additional covariate. True Logit uses correctly measured outcomes; all other models use misclassified outcomes. CRE = Correlated Random Effects. LPM = Linear Probability Model. See text for further details.

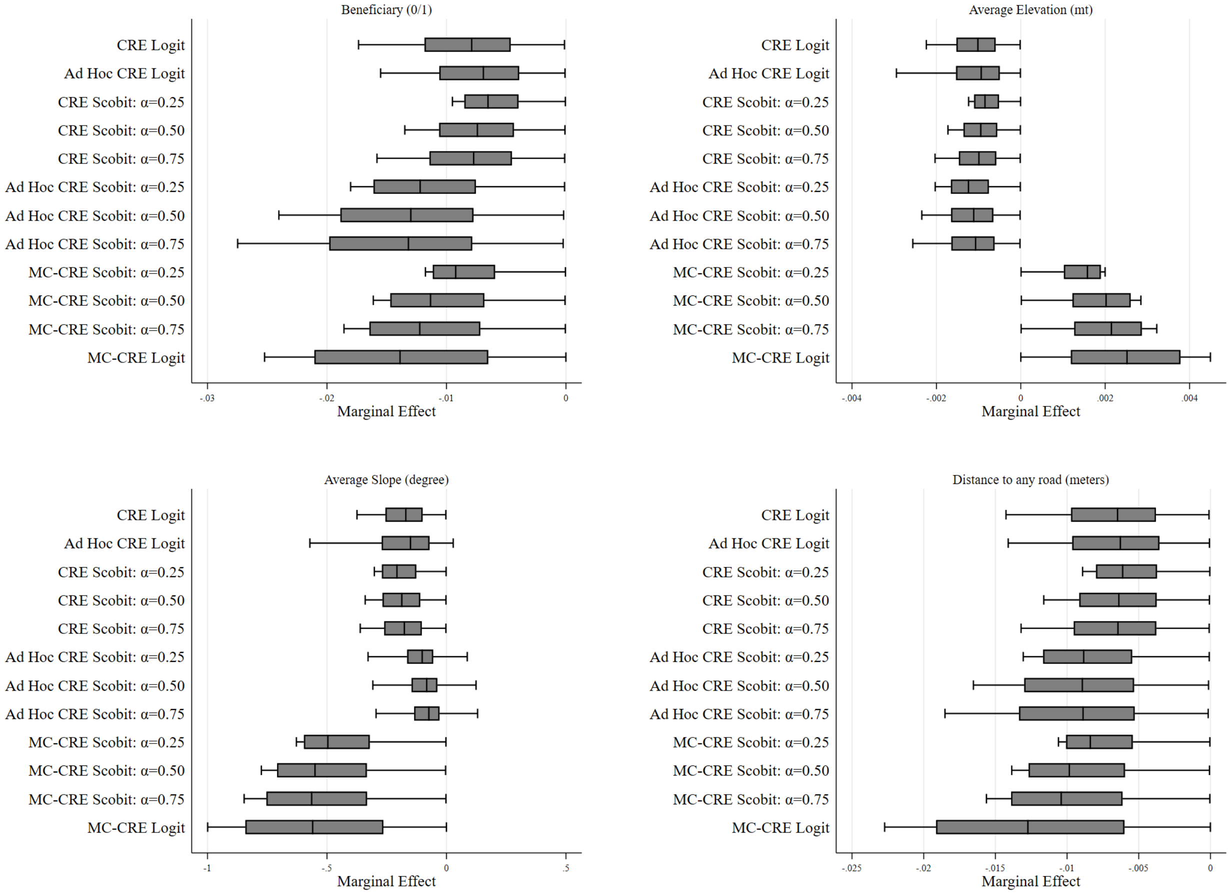
Figure C6: Monte carlo results: AME(x_2)



Notes: Markers show the ratio of RMSE of the indicated model to the RMSE of the CRE Logit model using the data free of measurement error. Values of β_0 decrease along the x-axis, resulting in lower presence of ones. Within each value for β_0, θ_0 increases also decreases from left to right, decreasing the rate of false positions. Within values of β_0 and θ_0, values of θ_1 decreases from left to right, lowering the rate of false negatives.

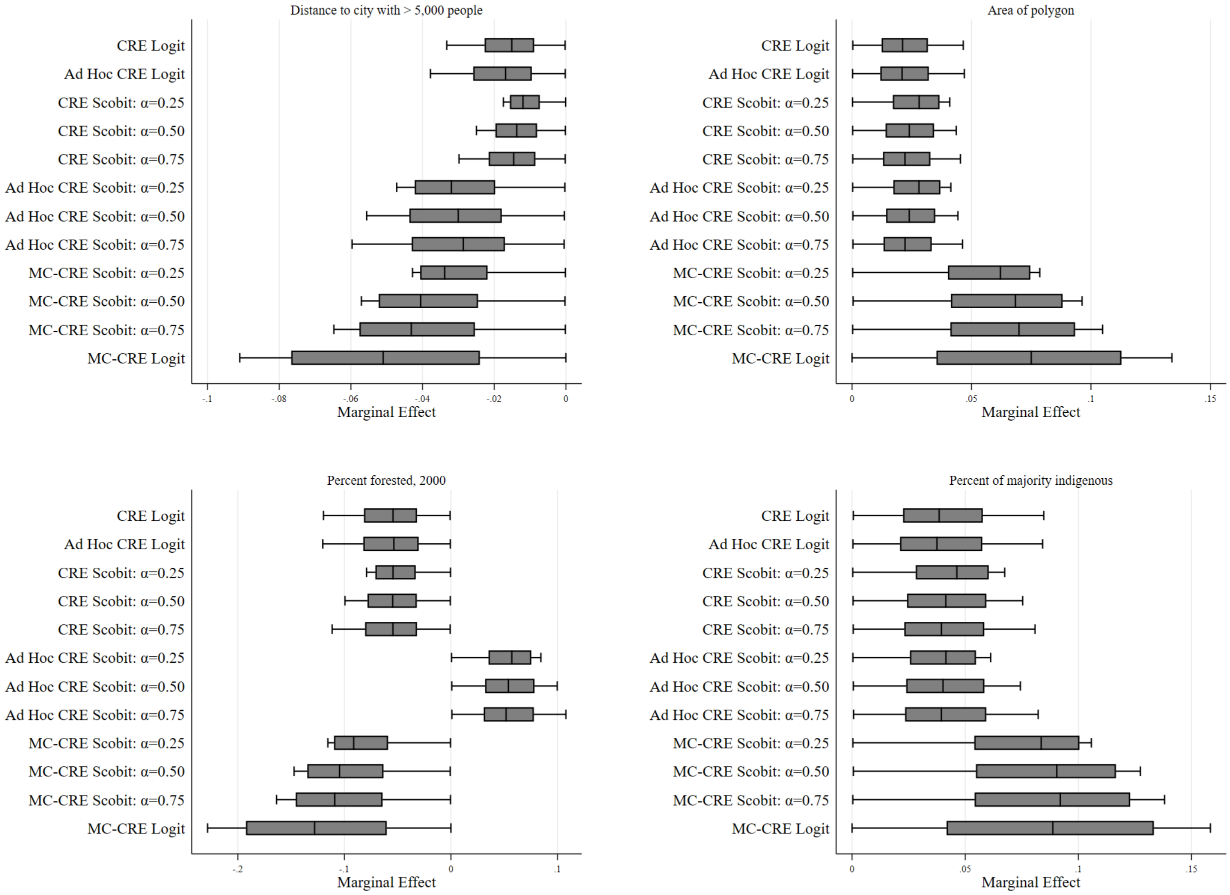
Appendix D Supplemental application results

Figure D1: Distribution of marginal effects of all covariates



Notes: Each shaded box spans the interquartile range; mid-line of the box corresponds to the median. Edges of the lines represent the minimum and maximum. Estimates obtained from columns 3-10 in Table 7 and the even-numbered columns in Table 8.

Figure D1 (cont.): Distribution of marginal effects of all covariates



Notes: Each shaded box spans the interquartile range; mid-line of the box corresponds to the median. Edges of the lines represent the minimum and maximum. Estimates obtained from columns 3-10 in Table 7 and the even-numbered columns in Table 8.