Improving Estimates of Transitions from Satellite Data: A Hidden Markov Model Approach*

Adrian L. Torchiana, Ted Rosenbaum,

Paul T. Scott, and Eduardo Souza-Rodrigues†

July 2022

Abstract

Satellite-based image classification facilitates low-cost measurement of the Earth's surface composition. However, misclassified imagery can lead to misleading conclusions about transition processes (e.g., deforestation, urbanization, and industrialization). We propose a correction for transition rate estimates based on the econometric measurement error literature to extract the signal (truth) from its noisy measurement (satellite-based classifications). No ground-truth data is required to implement the correction. Our proposed correction produces consistent estimates of transition rates, confirmed by Monte Carlo simulations and longitudinal validation data. In contrast, transition rates without correction for misclassification are severely biased. Using our approach, we show how eliminating deforestation in Brazil's Atlantic forest region through 2040 could save $100 billion in CO2 emissions.

*We are extremely grateful to Alexandre Camargo Coutinho and Daniel De Castro Victoria, who generously shared their ground-level land cover data with us, and to Marcos Reis Rosa, who shared the MapBiomas data that we use in our empirical application. We also would like to thank Christian Abizaid, Anne Ruiz-Gazen, and Christine Thomas-Agnan for helpful discussions. Additional comments and suggestions from the Editor and two anonymous referees helped us improve the paper considerably. We also thank Tommaso Alba for outstanding research assistance. Financial support from University of Toronto Mississauga is gratefully acknowledged. All remaining errors are our own. The views expressed in this article are those of the authors. They do not necessarily represent those of Granular Inc., nor the Federal Trade Commission or any of its Commissioners.

†Affiliations: Adrian L. Torchiana, Granular, Inc. (email: adrian.torchiana@gmail.com); Ted Rosenbaum, Federal Trade Commission (email: trosenbaum@ftc.gov); Paul T. Scott, New York University (email: ptscott@stern.nyu.edu); and Eduardo Souza-Rodrigues, University of Toronto (email: e.souzarodrigues@utoronto.ca).

Keywords: Measurement Error, Remote-Sensing Data, Land Cover, Hidden Markov Model
JEL Codes: C13, Q15, R14

1 Introduction

In recent years, publicly available satellite-based data combined with increasingly sophisticated machine learning algorithms have provided unprecedented access to regional and global estimates of Earth's surface composition.[1] Remote sensing data provide relatively low-cost information that is difficult to obtain by other means, with high spatial resolution and wide geographic and temporal coverage. Not surprisingly, they are increasingly used across a number of fields, including economics, geography, biology, ecology, and political science, and in setting policy.[2]

However, image classification techniques, which are used to convert the spectral signature of a pixel into an interpretable category, can lead to non-negligible misclassifications and bias areal estimates (Czaplewski, 1992; Jain, 2020). These classification errors can also affect estimates of the transition processes of outcomes of interest – our focus in this paper. Intuitively, errors in classifications can make transition rates appear excessively high. For example, much of the apparent land cover change in satellite-based data may be the result of misclassification (Abercrombie and Friedl, 2016). When remotely sensed rates of land cover change are used as inputs by decision makers (e.g., regulation in Brazilian Amazonia is based on remotely sensed deforestation rates; see Assunção et al., 2019) biases in transition rates can undermine efficient policy design and enforcement. Similarly, Fowlie et al. (2019) show how errors in satellite-based measures of pollution can lead to misleading information about pollution trends in different geographic areas, leading to over-regulation in "clean" areas and under-regulation in "dirty" areas.

To mitigate these concerns, researchers typically impose a set of heuristic and ad hoc adjustments

[1]There has been 12,012 satellites launched in total since 1957, according to the Index of Objects Launched into Outer Space. Among these, 8,140 were still orbiting our planet by December 2021 (http://www.unoosa.org/oosa/en/spaceobjectregister/index.html).

[2]See Donaldson and Storeygard (2016) for an overview of applications of remotely sensed data in economics. Geographers, biologists, and ecologists have also explored remote-sensing data to investigate land cover and degradation, terrestrial and marine ecosystems, sea level, biodiversity, and carbon emissions and carbon sequestration (Foley et al., 2005; Geller et al., 2017).

to stabilize classifications across years. Yet, Friedl et al. (2010) provide strong evidence that typical heuristic adjustments do not eliminate excessive rates of land cover change. An alternative solution is to correct classification errors using validation data that can be treated as ground truth (Czaplewski, 1992). However, extensive validation data are expensive to obtain and extremely scarce in practice (Goldblatt et al., 2016), and even when validation data exists, there may be few observations available, which limits the accuracy of the estimates.

In this paper we propose a different approach. We present a hidden Markov model (HMM) that corrects for misclassification bias. A HMM is the combination of an unobserved Markov process with observations that depend only on the contemporaneous hidden state (McLachlan and Peel, 2000). For instance, when studying land use change, the ground truth land use is the hidden state and classifications based on remote sensing imagery are the observations.[3] The idea here is to extract the signal (truth) from its noisy measurement (satellite-based classifications). The framework assumes that researchers either have access to panel data (with at least three time periods) of satellite-based classifications, or that they can generate such classifications themselves using remotely sensed data.

Based on Hu's seminal work on non-classical measurement error (summarized in Hu (2017, 2020)), we show how the HMM assumptions allow us to uniquely recover both the true transition probabilities *and* the misclassification probabilities from the observed data. The required assumptions (fully discussed below in Section 4) are not very restrictive in practice and some of them are testable. We discuss two different estimators for the hidden Markov model: a minimum distance (MD) estimator, that builds directly from the constructive identification results; and a maximum likelihood (ML) estimator, which is implemented using the expectation-maximization (EM) algorithm (Dempster et al., 1977; van Handel, 2008). Given estimated transition probabilities, we can construct the most likely trajectory of the states for each pixel in the data.

From the perspective of implementation, there are at least two attractive features to the HMM approach. First, we do not require ground-truth data to implement the correction. Second, the

[3]We use "land use" and "land cover" interchangeably in this paper.

HMM estimator can be implemented using classified data and not raw remote sensing data. These features allow for a division of labor between remote sensing specialists, who can classify the raw data, and applied researchers, who can implement the correction in their application.

We investigate the performance of our strategy in two different settings: (i) a land use change study based on rich longitudinal ground-truth validation data, and (ii) an empirical application focused on the Brazilian Atlantic Forest, one of the world's most threatened biodiversity hotspots.[4] In the first setting, we conduct a validation exercise using ground-truth longitudinal data for the state of Mato Grosso, Brazil, from 2006 to 2010 – a major center of agricultural production with rapid land use change within Brazil's Legal Amazon, a bio-administrative unit covering the Brazilian Amazon biome. The panel data provide a unique opportunity to test the performance of the HMM correction because they allow us to observe true transition rates (typical validation data are composed of a single or repeated cross-sections). Our HMM-based corrections estimate transition rates accurately; in contrast, transition rates computed without correction for misclassification are 3 to 9 times higher than observed in the ground truth data. We also improve the overall accuracy of the original classifications by finding the most likely sequence of land uses for each pixel in the data based on the HMM estimates.

Our second empirical application of the HMM correction concerns the social benefits from eliminating deforestation in Brazil's Atlantic Forest. The Atlantic Forest biome is located in the most developed region of the country and is the most degraded Brazilian biome, containing only roughly 30% of its original forest cover. This highly biodiverse area has been the target of various initiatives aimed at slowing deforestation and the resulting CO2 emissions and was included in Brazil's National Determined Contribution to the Paris Climate Agreement. Using Mapbiomas data, a rich remotely sensed database of Brazilian land cover and the HMM correction, we obtain an estimate of 2.78 million tons of carbon currently present in the forest (equivalent to roughly $774 billion in social value). Using the raw data without correction, we obtain an estimate of just

[4]We also carry out an extensive Monte Carlo study, presented in the Online Appendix. There, we find that the HMM method estimates transition probabilities and misclassification probabilities accurately, and we document important trade-offs between the two estimators: while the MD estimator is substantially faster, the ML estimator performs better in terms of mean-square errors.

2.39 million tons of carbon (equivalent to roughly $666 billion). The significantly lower estimated value based on the raw data is a consequence of predicting excessive land use transitions, which leads to lower predicted ages of the forest, and therefore lower carbon stocks (younger trees store less carbon). Simulating forward to 2040 using the HMM estimates, we predict that eliminating all deforestation from the Atlantic Forest would prevent approximately $100 billion value in carbon emissions. Beyond the current application (and in contexts other than deforestation), the approach we develop can help evaluate accurately the effectiveness, and optimality, of relevant environmental policies.

Related Literature. To the best of our knowledge, the closest papers to ours are by Abercrombie and Friedl (2016), Sandler and Rashford (2018), and Alix-Garcia and Millimet (2021). Like us, Abercrombie and Friedl (2016) consider an HMM-based correction to errors in classifications. They implement the HMM forward-backward algorithm (see van Handel, 2008, Chapter 3) to determine the most likely land cover for each pixel in a given year. In contrast to their work, we link the HMM procedure to formal identification results based on a set of explicit assumptions, bringing transparency to the contexts in which the correction is most appropriate and highlighting which assumptions can be tested in the data directly. Most crucially, we allow for the estimation of time-varying transition probabilities, which is essential in many applied studies, e.g. when estimating how (and explaining why) deforestation processes may change over time.

Sandler and Rashford (2018) and Alix-Garcia and Millimet (2021) focus on accurately modeling land use as a discrete choice problem. Both papers extend (in different directions) the maximum likelihood estimator with misclassified choices proposed by Hausman et al. (1998), and show that their estimators perform well in practice, while standard models (e.g., probit) result in estimated coefficients that are too close to zero, resembling an attenuation bias problem.[5] We focus instead

[5]Sandler and Rashford (2018) extend Hausman et al. (1998) to cover multinomial choice models, apply it to satellite-based agricultural land cover data in the US, and find that biofuel policies can have impacts on land use that are orders of magnitude larger than when misclassifications are ignored (in some cases, corrected effects are 350% larger than the uncorrected effects). Alix-Garcia and Millimet (2021) extend Hausman et al. (1998) to allow for misclassification rates that depend on covariates and incorporate the scobit family of binary choice models, which introduces an additional shape parameter into the link function – nesting the logit model as a special case, and helping

on estimating accurately the land use variable and the transition probabilities – our contributions are therefore highly complementary.

This paper is organized as follows: Section 2 discusses the sources and consequences of misclassifications in remote sensing data. Section 3 lays out the framework and formalizes the misclassification problem. Section 4 presents the HMM, HMM formal identification results, and two estimation methods. Section 5 describes the validation exercise using ground-truth data from Brazil's Embrapa, Section 6 presents our empirical application focused on the Brazilian Atlantic Forest, and Section 7 concludes.[6]

2 Sources and Consequences of Misclassifications

Sources of Misclassifications. In general, classifications are performed following four steps in remote sensing projects: data acquisition, pre-processing, analysis, and evaluation. Each step is associated with different types of potential errors, as we discuss next (Lillesand et al., 2015).

First, data is *acquired* by satellites as raw imagery. At this stage, errors may occur due to the combination of the specific sensor characteristics of the satellite (including sensor noise and response), the angle of the satellite with respect to the sun and the earth's surface, and atmospheric conditions, including cloud cover and haze. Second, *pre-processing* operations are used to correct for these errors (geometric and radiometric corrections), but such corrections are imperfect and may introduce new errors. Third, researchers *analyze* the images by training a classification algorithm on the pre-processed data. Continuous efforts to develop better algorithms, together with increased computer power, may help reduce errors, but are unlikely to eliminate them completely. Finally, the output classifications, extrapolated to the "held-out" or "testing" data set, are compared to

estimation when the outcome is of the rare-events type, as in the deforestation case. Applying their correction to a conservation cash-transfer program in Mexico, they find greater conservation impacts than when misclassifications are ignored.

[6]The Online Appendix presents (i) relevant mathematical derivations for the HMM correction; (ii) the measurement error in observed transition probabilities that is implied by the HMM model, and its consequences for regression analyses; (iii) the details of the EM algorithm; (iv) the Monte Carlo simulation studies; and (v) additional details on the carbon stock empirical application. Code for replicating the Monte Carlo simulations in R is available at https://github.com/atorch/hidden_markov_model.

ground-truth data to evaluate and improve the accuracy of the classification. Here, ground-truth data sampling may lead to further discrepancies due to, e.g., location accuracy (i.e., ground points may not coincide exactly with the pre-processed pixels) and scale misalignment (i.e., the size of ground-truth areas may be different from the units mapped from the imagery).

In general, specialists often consider overall accuracies for land cover classifications (namely, the percentage of correct classifications) to be acceptable when they are greater than 85%, though this threshold may vary depending on the context (Xie et al., 2020).[7] Indeed, classification accuracies can be lower in practice, given the difficulties involved. For instance, the widely-used Cropland Data Layer (CDL), developed by the US Department of Agriculture, National Agricultural Statistics Service (USDA NASS), has an overall accuracy of roughly 90%, in Iowa in 2018; the Mapbiomas data, Collection 5.0, has an overall accuracy of approximately 91.2% for the Brazilian territory, and around 90% for the Atlantic Forest biome, over the years 1985–2018; and the GlobeLand30, a global scale land cover mapping with 30 meters resolution launched by China in 2010, has an overall accuracy of 80% in 2010.[8]

Consequences of Misclassification. Remotely sensed data can be used to document and investigate spatial and temporal trends, as well as in regression analyses and causal inference studies; misclassifications in satellite-based data can affect the results of any of these types of studies.

Transition rates themselves are frequently important objects of interest. In the context of deforestation and afforestation rates, misclassification tends to exaggerate the rate at which land is moving in and out of forest, especially when misclassification rates are large relative to the

[7]For continuous variables, such as pollution, a common measure of overall accuracy (though not the only one) is based on the R^2 obtained from regressing the remotely sensed-based variable on the true measurement (based on ground-level monitors). In practice, the R^2's can vary in the 0.6–0.9 range; see, e.g., van Donkelaar et al. (2019) for a satellite-based measure of air quality in the US and Canada.

[8]The CDL Cropland Data Layer has recall rates (i.e., the percentage of correct predictions given the true land use) of roughly 92% for corn and soy, a recall of roughly 49% for alfalfa, and of 41% for winter wheat, in Iowa in 2018; see `https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/metadata_ia18.htm` for more detail. The overall accuracies of the Mapbiomas data presented above are based on the most aggregated land cover classifications; for the more disaggregated classifications, the accuracies drop slightly to 87% for the national territory, and to 86% for the Atlantic Forest. The recall rate in Brazil (Atlantic Forest) is around 96% (91%) for forest and roughly 79% (84%) for pasture; see `https://mapbiomas.org/en/accuracy-analysis`. For the GlobeLand300, overall accuracies can drop to less than 65% depending on the classifier used; see Chen et al. (2015).

true transition rates (which are typically low given large conversions costs). Even if the net rate of forest cover is accurate, exaggerated gross flows can have misleading implications for carbon dynamics because forest biomass takes many decades to accumulate (yonger forests hold less carbon than older forests). The erroneous rates can therefore distort the effectiveness, and optimality, of environmental policies in practice. The same reasoning applies to understanding the consequences for biodiversity, as they also depend asymmetrically on the estimated rate of habitat destruction and recovery. Estimating transition rates accurately is important more generally, beyond the land cover example; e.g., they are critical inputs for climate change models, which require well-documented evolution of polar ice coverage, sea levels, wildfires, wind patterns, relative humidity, and surface temperature, among other factors (all of which can be measured remotely based on satellite information, given the low spatial coverage of ground monitors); transitions are also important to study investment decisions in, say, housing innovations as in Marx et al. (2019).

In terms of regression analyses and causal inference, we note that measurement error in transition rates is not necessarily classical and can therefore lead to biases in linear and nonlinear models, and when transition rate is either the dependent or the independent variable.[9] In Online Appendix B, we explain formally how measurement error in transition rates leads to biases in linear, logit, and nested logit regression models when the transition rates are the dependent variable. We also show via simulation, in Online Appendix E.5, how this measurement error in a dependent variable leads to biased estimates of a policy designed to reduce deforestation. Researchers using remote sensing data in regression analyses should therefore be cautious when using remotely sensed transition rates. The HMM framework we outline below can help provide unbiased parameter estimates for many of these scenarios.

[9]Our point is different from Hausman et al. (1998) in that we consider continuously measured outcomes (transition rates) rather than binary outcomes. In general, biases are a concern not only with remotely sensed transition rates, but any remotely-sensed variables, such as fire incidents affecting health outcomes (Rangel and Vogl, 2019) and weather shocks impacting civil conflicts (Harari and Ferrara, 2018).

3 Framework

In this section, we illustrate how misclassification of remote sensing data can affect estimates of transition probabilities. Our running example is the land use classification problem, but results can be applied to other classification problems using longitudinal remote-sensing data, such as pollution, fire incidents, and nighttime lights.

Let $S_{it} \in \mathcal{S}$ denote the ground truth land use at location i at time t. In applications, a location is usually a pixel or a spatial point. The set of possible values that S_{it} can take is $\mathcal{S} = \{s_1, ..., s_K\}$, $K < \infty$. We do not restrict the number of elements in $\mathcal{S}$, so the land cover categories may be specific and numerous, or they may be very broad such as forest and non-forest. Extensions to continuously distributed measurements, such as pollution or nighttime light, are possible, at the cost of more burdensome notation and additional technical details.[10] The true land use S_{it} is not observed unless ground-truth data is collected for i at t.

Suppose there exists an observable noisy measurement of S_{it} denoted by $Y_{it} \in \mathcal{Y} = \{y_1, ..., y_K\}$. We assume the sets $\mathcal{Y}$ and $\mathcal{S}$ are equal, but we maintain the distinction in the notation for clarity. In typical applications, Y_{it} is the output of a classification algorithm that relies on machine learning techniques to predict S_{it} given a vector of (pre-processed) remote-sensing variables, R_{it}. For example, R_{it} may be a vector including some vegetation index, and the reflectance patterns of different wavelengths (infrared, red, blue, etc.) for pixel i at time period t. We can take $Y_{it} = f(R_{it})$, for some function f that depends on the data used and the classification algorithm.

We assume the researcher has access to a longitudinal data of land use classifications $\{Y_{it} : i = 1, ..., N; t = 1, ..., T\}$, obtained from remote-sensing data analysis (performed by the researcher herself or by others). In practice, it is common to have a large set of spatial points N and a small number of time periods T. Under standard regularity conditions, longitudinal data on Y_{it} can be used to estimate the transition probabilities $\Pr[Y_{it+1}|Y_{it}]$, as well as the marginal distribution

[10] For variables taking value on the real line, one needs to work in Hilbert spaces, with their corresponding operators (see, e.g., Hu and Schennach, 2008), instead of in Euclidean spaces with transformation matrices, as we do here. In empirical work, one may want to discretize continuously distributed variables in the data before applying our correction – we leave the investigation of optimal discretization for future research.

$\Pr[Y_{it}]$, with high accuracy. We can therefore treat these probabilities as known by the researcher for identification purposes. Importantly, while not explicit in the notation, we consider the analysis conditional on some set of observable covariates. For instance, the data may come from different subregions of a larger region of interest; the analysis can then be performed separately for (i.e., conditioned on) each subregion.[11] Furthermore, we allow the transition probabilities and marginal distributions to vary by year, so the t subscripts on Y_{it} and S_{it} should be understood to index the distribution the random variable is drawn from as well as the year of the observation.

For pixel i at time period t, the probability of observing land use prediction $Y_{it} = y$ is given by

$$\Pr[Y_{it} = y] = \sum_{s \in \mathcal{S}} \Pr[Y_{it} = y | S_{it} = s] \Pr[S_{it} = s],$$

where $\Pr[Y_{it} = y | S_{it} = s]$ is the probability of observing land use y when the ground truth land use is s; this is known as the misclassification probability when $y \neq s$. As mentioned previously, errors in classifications may be the combined result of the specific characteristics of the satellite, together with the pre-processing and classification operations (Lillesand et al., 2015).

In matrix notation, the equation above becomes

$$\mathbf{P}_{Y_t} = \Upsilon \mathbf{P}_{S_t}, \tag{1}$$

where $\mathbf{P}_{Y_t}$ is a $K \times 1$ vector with elements $\Pr[Y_{it} = y_k]$, $k = 1, ..., K$; the $K \times 1$ vector $\mathbf{P}_{S_t}$ has elements $\Pr[S_{it} = s_k]$; and Υ is a $K \times K$ matrix with $\Pr[Y_{it} = y_l | S_{it} = s_k]$, for $l, k = 1, ..., K$. We follow the literature and refer to the elements of Υ as misclassification probabilities, even though it includes the probabilities of correct classifications on the diagonal (also known as the "recall rate"), while the misclassification probabilities are the off-diagonal terms. For now, we consider the case where Υ is time-invariant, but the results can be extended to misclassifications that may change over time (discussed in Remark 1 below).

[11] Incorporating continuously distributed covariates, such as slope and altitude, is more cumbersome, but feasible. One can apply standard kernel smoothing techniques, or parameterize the transition probability functions.

While the vector $\mathbf{P}_{Y_t}$ can be estimated consistently using frequency estimators, it is not possible to recover the true land use distribution $\mathbf{P}_{S_t}$ without additional information. Further, there is no guarantee that the observed (estimated) transition $\Pr[Y_{it+1}|Y_{it}]$ is close to the true transitions $\Pr[S_{it+1}|S_{it}]$.[12]

4 Correction Based on the Hidden Markov Model

We now turn to our proposed solution. Here, we state Hu's (2017) conditions and results using our notation, and we discuss their plausibility and restrictiveness within the context of our satellite-based classification problem.[13]

For each point i, we assume the stochastic process $\{Y_{it}, S_{it} : t = 1, 2, ...\}$ follows a hidden Markov process. Specifically, we assume the ground truth land cover $\{S_{it}\}$ follows a first-order Markovian stochastic process with transition probabilities $\Pr[S_{it+1}|S_{it}]$, while Y_{it+1} is independent of past values $\{Y_{it-h}, S_{it-h}\}$, $h \geq 0$, conditional on S_{it+1}. This conditional independence assumption means that, if we know the true land use S_{it+1}, past variables (Y_{it}, S_{it}) do not contain any additional information about the noisy land-use classification Y_{it+1}. This is a common assumption in the measurement error literature (Bound et al., 2001; Schennach, 2021). Formally,

$$\begin{aligned} & \Pr\left[Y_{it+1}, S_{it+1} | \{Y_{it-h}, S_{it-h}\}_{h\geq 0}\right] \\ = \; & \Pr\left[Y_{it+1}|S_{it+1}, \{Y_{it-h}, S_{it-h}\}_{h\geq 0}\right] \times \Pr\left[S_{it+1} | \{Y_{it-h}, S_{it-h}\}_{h\geq 0}\right] \\ = \; & \Pr[Y_{it+1}|S_{it+1}] \times \Pr[S_{it+1}|S_{it}] . \end{aligned} \tag{2}$$

The HMM assumption is motivated by the fact that land use predictions Y_{it} are typically a function only of contemporaneous remote sensing data, R_{it}. If the process $\{R_{it}, S_{it}\}$ satisfies the

[12] In principle, ground-truth data can be used to estimate Υ, which would allow us to recover the true land use shares $\mathbf{P}_{S_t} = \Upsilon^{-1}\mathbf{P}_{Y_t}$, provided that Υ is invertible (Czaplewski, 1992). However, this approach suffers from the limitations discussed in the Introduction and does not recover true transition probabilities.

[13] The results are based on Section 2 of Hu (2017). See also Hu (2008) for his seminal contribution, with a more complete discussion and proofs, and Hu (2020) for a recent overview of econometric methods and applications to models with latent variables and measurement error.

HMM assumptions, then so must $\{f(R_{it}), S_{it}\}$ for any function f.[14] (Note that the observed process $\{Y_{it}\}$ does *not* necessarily follow a first-order Markov process.)

There are plausible situations in which equation (2) may be violated. One possibility occurs when misclassification probabilities are serially correlated; i.e., lagged values of (Y_{it}, S_{it}) may be useful in predicting current Y_{it} given current land use S_{it}. This may happen, for example, when researchers use past values of remote sensing data R_{it} to classify current land use Y_{it}. Another possibility is when true transitions may also depend on classified land uses, i.e., $\Pr[S_{it+1}|S_{it}, Y_{it}]$ – possible, e.g., when policymakers take actions based on the observed states.[15] Clearly, the appropriate application of the HMM depends on the context.

Equation (2) limits the potential datasets for which an HMM correction is appropriate. In some publicly available land cover datasets, such as Mapbiomas, ad hoc corrections have already been applied based on the time series of classifications, as mentioned in the Introduction. In such cases, it might not be reasonable to assume that misclassification probabilities do not depend on lagged variables. However, even when such ad hoc correction has been applied, it may be possible to obtain the raw data – we did this for our empirical application.

Useful Identities. Given the HMM setting, there are a series of identities that are helpful to obtain the identification results. For any two random variables X, W, define the $K \times K$ matrix $\mathbf{M}_{X,W}$ with elements given by the joint distribution $\Pr[X = s_l, W = s_k]$, with $s_l, s_k \in \mathcal{S}$ and $l, k = 1, ..., K$. Similarly, for any given $y_{t+1} \in \mathcal{Y}$, define the matrix $\mathbf{M}_{y_{t+1},X,W}$, with elements $\Pr[Y_{it+1} = y_{t+1}, X = y_l, W = y_k]$, as well as the diagonal matrix $\mathbf{D}_{y_{t+1}|X}$, with diagonal entries

[14] To see why, note that for any random variable Z, if $R_{it} \perp\!\!\!\perp Z|S_{it}$ (in words, if R_{it} is conditionally independent of Z given S_{it}), it follows that $f(R_{it}) \perp\!\!\!\perp Z|S_{it}$ for any function f. In typical applications, the remotely-sensed data R_{it} are complicated high-dimensional objects. In theory, we could fit an HMM using the process $\{R_{it}, S_{it}\}$. We opted for not doing so because the misclassification probabilities $\Pr[Y_{it}|S_{it}]$ can be represented by a $K \times K$ matrix, which is a much simpler object than a continuous distribution over high-dimensional sensor data. Finally, note that in typical annual classifications, one can make use of within-year time-series variation in remote-sensing data to classify annual land uses; for such cases, we extend our notation allowing the vector R_{it} to incorporate within-year remote sensor covariates.

[15] While we do not investigate the full extent of these possibilities here, we note that identification is possible for some of these cases, involving a more complex Markov process for (Y_{it}, S_{it}) and a more demanding set of identifying assumptions, exploring the ideas in Section 2.5 of Hu (2017).

$\Pr[Y_{it+1} = y_{t+1}|X = s_k]$.[16]

From the joint distribution of (Y_{it}, Y_{it-1}) we obtain

$$\mathbf{M}_{Y_t,Y_{t-1}} = \boldsymbol{\Upsilon}\,\mathbf{M}_{S_t,Y_{t-1}}. \tag{3}$$

Similarly, from the joint distribution of (Y_{it+1}, Y_{it}) we get

$$\mathbf{M}_{Y_{t+1},Y_t} = \boldsymbol{\Upsilon}\,\mathbf{M}_{S_{t+1},S_t}\,\boldsymbol{\Upsilon}^{\intercal}, \tag{4}$$

where the superscript $^{\intercal}$ denotes transpose. And, from the joint distribution of $(Y_{it+1}, Y_{it}, Y_{it-1})$, we have for a given $Y_{it+1} = y_{t+1} \in \mathcal{Y}$,

$$\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}} = \boldsymbol{\Upsilon}\,\mathbf{D}_{y_{t+1}|S_t}\mathbf{M}_{S_t,Y_{t-1}}. \tag{5}$$

Identification and estimation of the HMM is based on (3)–(5). See Online Appendix A for a derivation of these equations.

4.1 Identification of the Hidden Markov Model

Next, we outline the conditions needed to identify the Markov transition process $\Pr[S_{it+1}|S_{it}]$, the marginal distribution $\Pr[S_{it}]$ (including the initial distribution), and the misclassification probabilities $\boldsymbol{\Upsilon}$ using at least three periods of data on Y_{it}.

The first two conditions were discussed above and we state them here for completeness.

Condition 1. *The joint process $\{Y_{it}, S_{it}\}$ follows a hidden first-order Markov process, satisfying equation* (2).

Condition 2. *Y_{it} and S_{it} have the same support, i.e., $\mathcal{Y} = \mathcal{S}$.*[17]

[16] As we allow the distribution of Y_{it} to vary by year, note that the time subscripts on $y_{t+1} \in \mathcal{Y}$ serve to define the distribution used for Y_{it+1}.

[17] It is possible to extend the identification results to when the support of Y_{it} has more points than the support of S_{it}; i.e., when $card(\mathcal{Y}) \geq card(\mathcal{S})$. To see how, note that we can first combine some values in the support of $\mathcal{Y}$ to obtain

Next, we impose a mild restriction on observed classifications Y_{it}:

Condition 3. *The matrix* $\mathbf{M}_{Y_t,Y_{t-1}}$ *has full rank, i.e.,* $\text{rank}\left(\mathbf{M}_{Y_t,Y_{t-1}}\right) = K$.

This condition is testable. If the land use classifications Y_{it} are sufficiently persistent, $\mathbf{M}_{Y_t,Y_{t-1}}$ may be strictly diagonally dominant (note that for persistent processes, diagonal elements of this matrix will be larger than off-diagonal elements), and therefore full rank. This is plausible in land use applications because converting land is typically costly, which induces persistence in the data. Importantly, this condition implies that both matrices $\mathbf{\Upsilon}$ and $\mathbf{M}_{S_t,Y_{t-1}}$ are invertible, too (a fact that we use below); see equation (3). As a word of caution, note that, in practice, the larger the set of land uses considered (i.e., the larger the K), the less likely $\mathbf{M}_{Y_t,Y_{t-1}}$ will have full rank. That is because with more types of land uses in a given data (some of which could be rare), the higher the chances that some of them will not be observed in a time period, leading to a zero column (or row) in $\mathbf{M}_{Y_t,Y_{t-1}}$. Therefore, researchers using the HMM approach need to be careful when selecting the set of land uses in practice.

Combining (3) and (5), we get

$$\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}\mathbf{M}^{-1}_{Y_t,Y_{t-1}} = \mathbf{\Upsilon}\,\mathbf{D}_{y_{t+1}|S_t}\,\mathbf{\Upsilon}^{-1}. \tag{6}$$

This is an eigenvalue-eigenvector decomposition of a matrix constructed entirely from the data, i.e., from $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}\mathbf{M}^{-1}_{Y_t,Y_{t-1}}$. The columns of $\mathbf{\Upsilon}$ are the eigenvectors. Because each column of $\mathbf{\Upsilon}$ must sum to one, the scale of the eigenvectors is fixed. The diagonal elements of $\mathbf{D}_{y_{t+1}|S_t}$ are the eigenvalues. The next two assumptions guarantee a unique eigenvalue-eigenvector decomposition. The uniqueness of the decomposition means we can uniquely recover the misclassification probabilities $\mathbf{\Upsilon}$ and the diagonal matrix $\mathbf{D}_{y_{t+1}|S_t}$ from the joint distribution of the observed classifications $(Y_{it+1}, Y_{it}, Y_{it-1})$.

a transformed $\widetilde{Y}_{it}$ with support $\widetilde{\mathcal{Y}} = \mathcal{S}$, then apply the identifying assumptions to the process $(\widetilde{Y}_{it}, S_{it})$ and identify the parameters of the transformed process, and then "undo" the transformation and recover the entire process for the original (Y_{it}, S_{it}) – see Corollary 2.4.1 in Hu (2017). Though feasible, we do not exploit this possibility as Condition 2 seems reasonable for most applications.

Condition 4. $\Pr[Y_{it+1} = y|S_{it} = s] \neq \Pr[Y_{it+1} = y|S_{it} = s']$ *for at least one* $y \in \mathcal{Y}$ *whenever* $s \neq s'$, *and* $s, s' \in \mathcal{S}$.

Condition 4 assumes the eigenvalues are all distinct. This is testable: we only need to perform the eigenvalue-eigenvector decomposition of $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}\mathbf{M}^{-1}_{Y_t,Y_{t-1}}$ and check it.[18]

To interpret this condition, consider an example in which there are three land uses: forest, pasture, and crops. Take the observed y as forest. Suppose it is very likely to observe a forest classification tomorrow (i.e. Y_{it+1} = forest) when today's true land use is forest; moreover, suppose it is very unlikely that we would observe forest tomorrow when today's land use is pasture, and even less likely to see forest tomorrow when today's land use is crops (i.e., pasture and crops are both persistent, but pasture is abandoned more often than cropland). In this case,

$$\Pr[Y_{it+1} = y|S_{it} = \text{forest}] > \Pr[Y_{it+1} = y|S_{it} = \text{pasture}] > \Pr[Y_{it+1} = y|S_{it} = \text{crops}],$$

for $y = forest$, and so Condition 4 is satisfied here.

In case Condition 4 is violated for some y, we can use another land-use classification $y' \neq y$ for which the condition is valid. If we find no such y, then identification is not guaranteed. When Condition 4 holds for more than one value y, the model becomes overidentified.

Next we turn to the eigenvectors:

Condition 5. $\Pr[Y_{it} = s^*|S_{it} = s^*] > \Pr[Y_{it} = s|S_{it} = s^*]$ *for any* $s \neq s^*$, *and* $s, s^* \in \mathcal{S}$.

Condition 5 fixes the order of the eigenvectors. It implies s^* is the mode of the distribution $\Pr[Y_{it}|S_{it} = s^*]$. In words, given that the true land use is s^*, the probability that the noisy measure Y_{it} equals s^* is greater than the probability that Y_{it} equals any other land use $s \neq s^*$. This condition is satisfied when Υ is strictly diagonally dominant – this is reasonable as accurate land use classifiers (which is common in practice) should generate correct classifications well in excess of incorrect

[18]Condition 4 corresponds to Assumption 3 in Hu (2017). We take the function $\omega(y)$ defined in his assumption to be the Dirac function.

classifications.[19] (Note however that this condition is less likely to be satisfied when K is large: the greater the number of possible land uses, the less likely the correct classification s^* will be the modal outcome.)

Next, given identification of the misclassification probabilities Υ from the eigenvalue-eigenvector decomposition (6) under Conditions 1–5, we identify the joint distribution $\Pr[S_{it+1}, S_{it}]$ under the assumption that Υ is time-invariant. For completeness, we impose this condition explicitly (and discuss its relaxation in Remark 1 below):

Condition 6. $\Pr[Y_{it+1}|S_{it+1}] = \Pr[Y_{it}|S_{it}]$, *for all* t.

Given Condition 6 and equation (4), we obtain

$$\mathbf{M}_{S_{t+1},S_t} = \Upsilon^{-1}\mathbf{M}_{Y_{t+1},Y_t}\left(\Upsilon^{\intercal}\right)^{-1}, \tag{7}$$

which implies identification of $\mathbf{M}_{S_{t+1},S_t}$, and hence of both $\Pr[S_{it+1}|S_{it}]$ and $\Pr[S_{it}]$. The argument above leads to the following proposition:

Proposition 1. *(Theorem 1, Hu (2017)). Suppose Conditions 1–6 hold. Then, the joint distribution of the observed classifications* $(Y_{it+1}, Y_{it}, Y_{it-1})$ *uniquely identifies* $\Pr[Y_{it}|S_{it}]$, $\Pr[S_{it+1}|S_{it}]$, *and* $\Pr[S_{it}]$.

Remark 1. (Time-varying misclassifications.) It is plausible to consider misclassification probabilities that do not vary over time when there are no common shocks (due to, say, meteorological conditions) affecting atmospheric noise, nor technical difficulties requiring satellite maintainance that may affect the quality of the raw data, nor substantive technical changes to the pre-processing and classification operations generating Y_{it} during the sampling period. However, when some of these conditions fail, misclassification probabilities may change over time – a possibility that we can accommodate in our framework.

[19] Condition 5 corresponds to Assumption 4.2 in Hu (2017). Alternatively, one could instead assume the misclassification probabilities $\Pr[Y_{it} = y|S_{it} = s]$ are decreasing in s for some y; this would correspond to Assumption 4.1 in Hu (2017) and it also pins down the ordering of the eigenvectors.

Formally, denote the time-varying matrix by Υ_t. Equation (6) then becomes $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}\mathbf{M}^{-1}_{Y_t,Y_{t-1}} = \Upsilon_t \mathbf{D}_{y_{t+1}|S_t} \Upsilon_t^{-1}$, which means that we still need just three time periods to identify the misclassification probabilities. We also need three time periods to identify the marginal distribution $\Pr[S_{it}]$ – see equation (1). However, equation (7) becomes $\mathbf{M}_{S_{t+1},S_t} = \Upsilon^{-1}_{t+1}\mathbf{M}_{Y_{t+1},Y_t}\left(\Upsilon_t^{\intercal}\right)^{-1}$, implying that we now need $T \geq 4$ periods of data to identify $\mathbf{M}_{S_{t+1},S_t}$ (we need $t+1$, t, and $t-1$ to identify Υ_t and t, $t+1$, and $t+2$ to identify Υ_{t+1}). Clearly, Υ_1 and Υ_T are not identified, and neither are the transition probabilities in the first and last time periods.

4.2 Estimators for the HMM Correction

We consider two estimators for the HMM correction: a minimum distance (MD) estimator and a maximum likelihood (ML) estimator. The estimators offer complementary advantages: we find the ML estimator is more precise while the MD estimator is computationally faster.

As mentioned previously, we assume the researcher has access to a panel data of classifications $\{Y_{it} : i = 1, ..., N; t = 1, ..., T\}$. Following Conley (1999), we allow (Y_{it}, S_{it}) to be a random field – i.e., we let an observation be a realization of a random process at a point in a Euclidean space. In the temporal dimension, we assume each point follows the hidden first-order Markov process discussed in the previous section; in the cross-sectional dimension, we assume weak dependence – that is, as the spatial distance between pixels increases, the outcomes (Y_{it}, S_{it}) and (Y_{jt}, S_{jt}), for $i \neq j$, become essentially independent.[20]

[20]Specifically, we assume the sample consists of realizations of the random variables at a collection of locations inside a sample region. As in Conley (1999), we assume the sample region grows in area as the sample size increases to ensure that the vector indexing cross-sectional dependence is not superfluous. (That is in contrast to an "infill" asymptotics, in which case observations get increasingly dense in a fixed region, violating weak dependence.) We assume (Y_{it}, S_{it}) satisfies the mixing condition stated in Section 3.1.3 in Conley (1999) – see also his assumptions A1, A3, and B1–B3. These assumptions allow one to make use of Law of Large Numbers applied to weakly dependent data to obtain consistency of the estimators. Similarly, inference can be based on central limit theorems for stationary and mixing random fields on regular lattices, as developed, e.g., by Bolthausen (1982).

4.2.1 Minimum Distance Estimator

In principle, we can estimate the misclassification probabilities and the joint distribution of S_{it} using a plug-in estimator based on equations (6)–(7). However, the eigenvalue-eigenvector decomposition may result in estimated probabilities that are negative or greater than one in some data sets. In our experience, this is more likely to happen when the sample size is small and the true parameters are close to one (e.g. transition probabilities of 0.99). For this reason, it is better to implement a constrained minimum distance estimator (as suggested by Hu, 2017, in his Section 2.6).

For convenience, we denote $\mathbf{M}_{S_{t+1},S_t} = \mathbf{M}_t$ for all t, and collect all matrices into $\mathbf{M} = \{\mathbf{M}_t : t = 1, ...T-1\}$, where $T \geq 3$. Define the following functions, for some $y \in \mathcal{Y}$,

$$\begin{aligned} g_{1yt}(\mathbf{M}, \mathbf{\Upsilon}) &= \left\| \mathbf{M}_{y_{t+2},Y_{t+1},Y_t} \mathbf{M}^{-1}_{Y_{t+1},Y_t} \mathbf{\Upsilon} - \mathbf{\Upsilon} \mathbf{D}_{y_{t+2}|S_{t+1}} \right\|, \\ g_{2t}(\mathbf{M}, \mathbf{\Upsilon}) &= \left\| \mathbf{M}_{Y_{t+1},Y_t} - \mathbf{\Upsilon} \mathbf{M}_t \mathbf{\Upsilon}^\intercal \right\|, \end{aligned} \tag{8}$$

where $\|\cdot\|$ is a matrix norm. Notice that g_{1yt} is analogous to equation (6) with slight rearrangement, while g_{2t} is analogous to equation (7). So, under the true joint distributions and misclassification probabilities, $\mathbf{M}$ and $\mathbf{\Upsilon}$, respectively, we have that $g_{1yt} = g_{2t} = 0$. (We omit $\mathbf{D}_{y_{t+2}|S_{t+1}}$ as an argument of g_{1yt} because it is a function of $\mathbf{\Upsilon}$ and $\mathbf{M}_{S_{t+2},S_{t+1}}$.)

Let g_1 be a vector that stacks g_{1yt} for all $t \in \{1, \ldots, T-2\}$, and let g_2 be a vector that stacks g_{2t} for all $t \in \{1, \ldots, T-1\}$. Define the vector $g = (g_1^\intercal, g_2^\intercal)^\intercal$, and consider the population criterion function $Q(\mathbf{M}, \mathbf{\Upsilon}) = g(\mathbf{M}, \mathbf{\Upsilon})^\intercal \mathbf{W} g(\mathbf{M}, \mathbf{\Upsilon})$, where $\mathbf{W}$ is a symmetric positive-definite weighting matrix. By construction, $Q(\mathbf{M}, \mathbf{\Upsilon}) \geq 0$, and the true matrices $(\mathbf{M}, \mathbf{\Upsilon})$ are the unique solution to the following minimization problem:

$$\min_{\mathbf{M},\mathbf{\Upsilon}} \; g(\mathbf{M}, \mathbf{\Upsilon})^\intercal \; \mathbf{W} \, g(\mathbf{M}, \mathbf{\Upsilon}), \tag{9}$$

subject to each matrix entry being in $[0, 1]$ and probabilities summing up to one.

The minimum distance estimator is the sample analog of (9):

$$(\widehat{\mathbf{M}}, \widehat{\boldsymbol{\Upsilon}}) = \arg\min_{\mathbf{M},\boldsymbol{\Upsilon}} \ \widehat{g}\left(\mathbf{M}, \boldsymbol{\Upsilon}\right)^{\intercal} \ \widehat{\mathbf{W}}\, \widehat{g}\left(\mathbf{M}, \boldsymbol{\Upsilon}\right), \tag{10}$$

subject to the same constraints as above, where $\widehat{g}$ is a vector with elements defined in the same way as in (8), but replacing $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}$, $\mathbf{M}_{Y_t,Y_{t-1}}$, and $\mathbf{M}_{Y_{t+1},Y_t}$ by their respective frequency estimators $\widehat{\mathbf{M}}_{y_{t+1},Y_t,Y_{t-1}}$, $\widehat{\mathbf{M}}_{Y_t,Y_{t-1}}$, and $\widehat{\mathbf{M}}_{Y_{t+1},Y_t}$, and $\widehat{\mathbf{W}}$ is a data-dependent symmetric positive-definite weighting matrix that converges in probability to $\mathbf{W}$.[21] This is a standard minimum distance estimator defined over a finite-dimensional parameter space: under i.i.d. data and standard regularity conditions, the estimator is consistent and asymptotically normal (Newey and McFadden, 1994). We expect that the same asymptotic properties can be established using arguments similar to Conley (1999). As usual, inference must be adjusted when parameters are at or near the boundary (Politis and Romano, 1994; Andrews, 1999, 2000).

In general, if we estimate a model with K hidden states from T years of data, we have to optimize over $K(1 + KT)$ parameters subject to $TK + 1$ equality constraints and boundary conditions for every parameter ensuring it is in [0,1]. For instance, when there are $K = 2$ states and $T = 3$ time periods, we have 7 parameters to estimate in total.[22]

4.2.2 Maximum Likelihood Estimator

Next, we consider a maximum likelihood estimator. Let $\Pr[Y_i]$ be the joint distribution of $Y_i = (Y_{i1}, ..., Y_{iT})$ for a given point i. The pseudo-log likelihood function is

$$L = \sum_{i=1}^{N} \ln \Pr[Y_i], \tag{11}$$

[21] When Condition 4 is satisfied for more than one value of $y \in \mathcal{Y}$, the vector g_{1t} may be augmented accordingly. When that happens, or when $T \geq 4$, the model becomes overidentified. When the panel data is unbalanced (assuming that the data is missing-at-random), we only use the observations for which we have (i) two consecutive periods to estimate $\mathbf{M}_{Y_t,Y_{t-1}}$ and $\mathbf{M}_{Y_{t+1},Y_t}$, and (ii) three consecutive periods to estimate $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}$.

[22] The total number of parameters in T years, before accounting for constraints, is K (corresponding to the initial distribution), plus $(T-1)K^2$ (corresponding to $(T-1)$ transition matrices), and K^2 (the misclassification matrix). The number of equality constraints is 1 (for the initial distribution), plus $(T-1)K$ (for the $T-1$ transition probability matrices), and K (for the time-invariant misclassification probabilities).

where the likelihood function for observation i integrates-out the hidden states:

$$\Pr[Y_i] = \sum_{s_1 \in \mathcal{S}} \cdots \sum_{s_T \in \mathcal{S}} \Pr[S_{i1} = s_1] \Pr[Y_{i1}|S_{i1} = s_1] \prod_{t=2}^{T} \Pr[S_{it} = s_t|S_{it-1} = s_{t-1}] \Pr[Y_{it}|S_{it} = s_t]. \tag{12}$$

The ML estimator chooses the initial distribution $\Pr[S_{i1}]$, the transition probabilities for S_{it}, and the misclassification probabilities that maximizes the function L. As is well-known, the ML estimator is consistent, asymptotically normal, and asymptotically efficient (Newey and McFadden, 1994). As before, we expect the asymptotic properties to extend to spatially weakly dependent data following the arguments developed by Conley (1999). Because maximizing L directly is difficult in practice, we follow the literature and use the expectation-maximization (EM) algorithm (Dempster et al., 1977; van Handel, 2008) – see Online Appendix C for details.

While the ML estimator enforces Conditions 1 and 2 by construction (as well as Condition 6, when we assume misclassification probabilities are time-invariant), it does not impose the remaining assumptions (i.e., Conditions 3–5) in the estimation routine. That is in contrast to the MD estimator, which imposes all identification conditions explicitly. Although feasible, imposing these conditions is not necessary in the ML estimation procedure because, when they hold in the data generating process (and so can be treated as regularity conditions), the ML estimator will converge in probability to the true transition and misclassification probabilities satisfying these conditions.

Remark 2. (Monte Carlo Simulations.) In Online Appendix E, we present our Monte Carlo study. Here, we highlight the main take-aways. First, we find that the HMM approach estimates transition probabilities and misclassification probabilities accurately, including cases where the transition probabilities are time-varying. Second, both MD and ML estimators outperform a standard frequency estimator that ignores misclassifications, which tends to overestimate transition rates when they are persistent. Third, we find important trade-offs between the two estimators. While the MD estimator is substantially faster, the ML estimator performs better in terms of mean-square errors and is less likely to result in estimates of transition probabilities that are at the edge

of the unit interval $[0, 1]$. Fourth, when using the (fast) MD estimator as the initial value for the (more accurate) ML estimator, the combined approach is about 10 times faster on average than the ML estimator alone with random initialization. ML initiatlized with the MD estimate also results in mean-square errors similar to the ML estimator, so this approach has much to recommend it. Fifth, when we allow misclassification probabilities to depend on past values of (Y_{it}, S_{it}), violating equation (2), our MD and ML estimators are biased (as expected), but they are substantially less biased than the frequency estimator ignoring errors in classifications. Finally, we find that when the ML estimates of transition rates are used as the dependent variable in a simple linear treatment effects regression analysis, one obtains unbiased estimates of the treatment effects. In contrast, using a frequency estimator leads to biased estimates of the impact of the treatment.

5 Validation Exercise Using Land Cover Data

We now investigate the performance of the HMM approach using unique validation data from the Brazilian Agricultural Research Corporation (Embrapa).[23] We outline the data, the implementation of our methodology, and the validation results.

5.1 Ground-Truth and Remote-Sensing Data

The ground-truth data contain information on land use at 409 spatial points observed annually from 2006 to 2010, in the state of Mato Grosso, Brazil. The state of Mato Grosso has attracted considerable interest from researchers and policy makers both because it is a major center of agricultural production within Brazil's Legal Amazon (a bio-administrative unit covering the Brazilian Amazon biome) and because of the rapid land use change there due to agricultural development. The field data were collected from private farms within 14 municipalities in the most intensely cropped region of central Mato Grosso; see the study area and the sample points in Figure F1 of

[23] We are grateful to Alexandre Camargo Coutinho and Daniel De Castro Victoria, who generously shared their data with us.

the Online Appendix.[24] The data is unprecedented in spatial and temporal coverage for the state – and arguably in general – and provide a unique opportunity to test the performance of the HMM correction in practice because they allow us to observe true land use transition probabilities and compare them to our estimates. Further, they also allow us to compare our HMM estimates of misclassification probabilities with a direct estimate of misclassifications $\Pr[Y_{it}|S_{it}]$.

Embrapa's land cover data include various land use categories, but the vast majority of points are either in crops or pasture. We therefore consider two land uses, $\mathcal{S} = \{\text{crops}, \text{pasture}\}$. A small number of points do not fit into either of these categories (e.g. points classified as natural vegetation); we drop these observations. Ultimately, we have 403 unique spatial points, each point observed for one to five years in 2006–2010 (resulting in an unbalanced panel).[25]

We merge the ground truth land use data with remote-sensing data. Specifically, we use measurements from the sixteen-day composite Terra MODIS 250m.[26] MODIS data provide measurements over time of five variables that we observe for each pixel i: the reflectance of (i) near infrared (NIR), (ii) middle infrared (MIR), (iii) red, and (iv) blue, as well as (v) the enhanced vegetation index (EVI). Given that MODIS collects information for each pixel every sixteen days, each variable is recorded 23 times per year. In total, we have 115 MODIS covariates per year – these correspond to the vector of variables R_{it} discussed previously in Section 3. We merge the MODIS data with the Embrapa ground-truth data considering the September-to-August harvest years for consistency. In this way, the 2006 ground-truth data, for instance, are merged with sensor data from September 2005 to August 2006.

[24]The data were collected via farmer or farm manager interviews. The cropping practices were recorded for each individual sites and integrated into a Geographic Information System (GIS) to be combined with the MODIS remote-sensing data (see more below). The area covered extends from the coordinate $(59^o25'14''W, 14^o2'39''S)$ [lower left] to the point $(54^o25'19''W, 11^o42'16''S)$ [upper right]. A total of 40 farmers or farm managers were interviewed as research participants. For more details, see Coutinho et al. (2011) and Brown et al. (2013).

[25]Of the 403 spatial points, 63 are missing ground truth land use data in one or more years. Overall, we observe ground truth land use for 93.5% of point-years. In the estimation procedure, we assume the missing data is missing at random.

[26]More precisely, the MOD13Q1 (Collection 5), with spatial resolution of 250 meters and 16-day composite interval, obtained from the United States Geological Survey's Land Processes Distributed Active Archive Center (LP DAAC). We used one MODIS tile (h12v10), which covers the entire field study area. This is consistent with the analysis in Brown et al. (2013).

5.2 Generating Land Use Classifications

After merging the ground truth and the MODIS datasets, we construct the satellite-based classifications, $\{Y_{it}\}$. To that end, we randomly split the panel data into two disjoint sets. The first ("training set") is used to train a machine learning classifier. We use a gradient-boosted ensemble of classification trees, commonly referred to as a GBM (Hastie et al., 2009, Chapter 10), to predict the land cover using the MODIS covariates.[27] With the second set of data ("test set"), we obtain the out-of-sample predictions $Y_{it} = f(R_{it})$ based on the GBM classifier. The out-of-sample predictions in the test set constitute our land use classification panel data, $\{Y_{it} : i = 1, ..., N, t = 1, ..., T\}$.

The training set contains 60 cross-sectional points (286 point-years) and the test set contains 343 points (1715 point-years).[28] We opt for a larger fraction of the Embrapa data set to be part of the test set in order to reflect the typical scenario faced by applied researchers using satellite-based data: they will typically have access to large panels of machine-learning-based classifications.

5.3 Implementing the HMM Correction

We consider two hidden Markov model specifications: a restricted model with time-invariant transition probabilities, and another in which the transitions are allowed to vary over time. In both cases, we hold the misclassification probabilities to be the same over time. We estimate the HMM parameters using both the MD and the ML estimators.[29] The frequency estimator (ignoring misclassifications) is computed directly from the data, based on sample frequencies. Confidence intervals are calculated based on subsampling, as suggested by Politis and Romano (1994) and Andrews (1999, 2000) when parameters are at or near the boundary.[30]

[27]The purpose of boosting is to apply "weak" classification algorithms sequentially to produce a "strong" classifier. The GBM uses a sequence of decision trees in which each individual tree tries to recover the loss (i.e., the difference between actual and predicted values) obtained by the previous ones in the sequence. The loss function is minimized using a gradient descent algorithm. To select the optimal number of trees, we follow standard practice and use cross-validation. See Chapter 10 of Hastie et al. (2009) for recommendations on tuning GBMs.

[28]Our results are similar when we select different sizes for the training set (e.g., 48 or 80 cross-sectional points).

[29]For the MD estimator, we use the identity matrix as the weighting matrix $\mathbf{W}$, and we use both $y_{t+1} = crops$ and $y_{t+1} = pasture$, as they both satisfy Condition 4.

[30]We implement 200 replications of a standard i.i.d. subsampling, resampling 250 spatial points over the sample time period. (We acknowledge, though, that this might not be completely accurate in the presence of spatial dependence.) The 95% confidence intervals are calculated as $[\widehat{\theta} - \delta_{0.025}, \widehat{\theta} + \delta_{0.975}]$, where $\widehat{\theta}$ denotes the parameter estimate, and

5.4 Validation Results

The out-of-sample performance of our GBM classifications is shown in Table 1. This table presents the so-called "confusion matrix," which tabulates the test points according to their ground truth class and predicted class. It also allows us to estimate the misclassification probabilities Υ directly.

The GBM's land use predictions are fairly accurate given the size of the training set and the difficulty of the classification problem. Overall, it correctly predicts land use for 92% of the test points, which is in the range of acceptable accuracies (see Section 2). For crops, the fraction of correctly predicted (or the recall rate) is 92.6% (i.e., $\Pr[Y_{it} = \text{crops} \mid S_{it} = \text{crops}] = 0.926$), while the recall for pasture is 79.5% (i.e., $\Pr[Y_{it} = \text{pasture} \mid S_{it} = \text{pasture}] = 0.795$). This implies that Υ is diagonally dominant, as required for identification of the HMM approach (see Condition 5 in Section 4.1). This increases our confidence that the HMM is identified when applied to the Embrapa test data.[31]

Figure 1 shows the estimated results for the frequency estimator and the restricted HMM (i.e., imposing time-invariant transition probabilities). The ground-truth data indicate that the probability of switching from cropland to pasture in the following year equals 0.7%, while the probability of maintaining cropland is 99.3%. The ground-truth probability of switching from pasture to crops is 13.8%, and the probability of maintaining pasture land is 86.2%. So, both land uses are persistent over time, and cropland is more persistent than pasture.

The frequency estimator estimates transitions from crops to pasture as 6.2%, and transitions in the opposite direction, from pasture to crops, as 48.2%. These transitions are substantially biased: the first one is roughly 9 times higher than the truth, while the second is 3 times higher than the correct transition. In contrast, the HMM estimates for the transitions probabilities (using both MD and ML estimates) are approximately 1.2% for cropland to pasture and 6.5% from pasture to cropland, which are substantially closer to the true ground-truth transition probabilities than the

δ_q is the quantile q of the subsampling distribution. We do not implement bootstrap procedures because they are inconsistent when parameters are at or close to the boundary (Andrews, 2000). We treat the GBM parameters as fixed and subsample only on the test data, noting that researchers may not have access to the training data in practice.

[31] While not presented here, we find that the (testable) Conditions 3 and 4 are also satisfied in the data.

frequency estimates. The confidence intervals in the figure indicate that these results hold even after accounting for sampling uncertainty. This is consistent with the Monte Carlo results discussed in the Online Appendix: the frequency estimator tends to overestimate switching rates when land use is persistent.

Figure 2 is analogous to Figure 1, but shows results for the *unrestricted* model, i.e. allowing for time-varying transition probabilities. The results are similar to the time-invariant case: the frequency estimator provides excessive land use changes, while the HMM corrections result in point estimates that are closer to the true transitions. That is the case even when true transitions are exactly zero, as in the first year of the data, 2006–2007. We also find some evidence that transition rates can vary over time (though not substantially in this data set).

Next, we turn to the HMM estimates of the misclassification probabilities; Figure 3 shows the results computed using MD and ML for both the restricted and unrestricted models. The point estimates are all reasonably close to the true misclassification probabilities obtained from the out-of-sample confusion matrix for the GBM predictions (see the last column in Table 1.) This is notable since the HMM is estimated using only panel data of observed classifications $\{Y_{it}\}$, with no information on true land uses $\{S_{it}\}$.

Finally, we generate the most likely sequence of land uses for each point in the test set based on the HMM maximum likelihood estimates (using the so-called "Viterbi" algorithm; see Online Appendix C) and reclassified the data points accordingly. Figure 4 shows that correcting classifications in this way increases the overall accuracy of the land use classifications to 96%, improving on our original classifier's accuracy of 92%.

6 Empirical Exercise: Carbon Stocks in the Atlantic Forest

We now investigate an application of the HMM approach by estimating the value of the carbon stocks in the Brazilian Atlantic Forest and how they might change over time if deforestation were curtailed. These carbon stocks represent part of the social costs of deforesting an area, and so

quantifying them provides a crucial input to environmental policy analysis. Given that the carbon stock varies with the age of the forest, it is critical to obtain an accurate measure of the forest age, which in turn can be estimated based on transition rates between forest and non-forest; the HMM approach allows the researcher to obtain accurate measurements of these important transition rates. We also use the HMM approach to compute the value of the carbon stock by 2040 if deforestation is completely eliminated between 2020 and 2040.

The Atlantic Forest is a region of approximately 1.4 million square kilometers, it occupies approximately 15% of the Brazilian territory (stretching from the northeastern to the southern regions), and it accounts for about 70% of the country's population and about 80% of the national gross domestic product. It has suffered centuries of exploitation and, as a result, it now contains only an estimated 30% of the original native forest cover (Rosa et al., 2021). It also hosts one of the world's most diverse and threatened tropical forest on the planet.[32] Given that the Atlantic Forest is a priority hot spot for biodiversity conservation, and is a tropical forest storing large amounts of carbon on the ground, it has been the focus of many conservation and restoration policy initiatives (Brancalion et al., 2016). Indeed, conserving this area is an important component of the country's National Determined Contribution to the Paris Climate Agreement.[33]

6.1 Data and Implementation

We use data from the Mapbiomas (Collection 5.0), an initiative that has produced annual land cover time series based on 30 meters resolution Landsat satellite data.[34] It covers the whole Brazilian territory from 1985 through the present and provides an unprecedented tool for understanding forest dynamics – consistent monitoring of forest dynamics in the Atlantic Forest was not possible until the creation of the MapBiomas project in 2015.

The data for the Atlantic Forest includes a total 7.9 billion pixels observed from 1985 to 2020;

[32]The region harbors roughly 20,000 plant species, more than 1400 species of terrestrial vertebrates, and thousands of invertebrate species, many of which are endemic – and many of which are endangered (Laurance, 2009).

[33]For more details, see `http://www.mma.gov.br/images/arquivos/florestas/planaveg_plano_nacional_recuperacao_vegetacao_nativa.pdf`.

[34]Marcos Reis Rosa generously shared with us the MapBiomas data, for which we are extremely thankful.

Figure F2 in the Online Appendix presents a map with the data points. Mapbiomas considers several land cover classifications that are generated in each year using the random forest machine learning classification algorithm available in Google Earth Engine (Gorelick et al., 2017). We aggregate these land covers into three states: forest, deforested, and others.[35] In the notation of the paper, we treat these aggregated MapBiomas classifications for a given year and pixel as Y_{it}.

We split the biome into "tiles," each containing 1000 x 1000 pixels and covering an area of roughly 900 square kilometers. We exclude tiles where over 90% of the pixels are missing or over 50% of the land is water or sand. The HMM parameters are estimated separately for each tile; to ease computational burden, we randomly select 1% of the pixels within each of these tiles for estimation. (For the tiles where only two of the three aggregated classes are present in the observed data, we estimate a two state model.) Following our Monte Carlo study, we first estimate the HMM parameters using the (fast) MD estimator and then use these parameter estimates as starting values for the (more accurate) ML estimator. We exclude from the final results the tiles for which we obtain estimates for the misclassification probability matrix Υ that are not diagonally dominant, since that violates the assumptions underlying our procedure. In total, we run estimates for 1,174 tiles, roughly 75% of the Atlantic Forest. (For our counterfactual simulations, we assume that the forest age and transitions for this 75% is equal to the other 25%.)

To calculate the carbon stocks, we first compute the most likely land use path for each pixel, given the estimated HMM parameters for each tile. We do so in order to distinguish between older and newer forests. Specifically, we define the forest age as the number of years that the pixel has been classified as forest since the last time it was classified as deforested; for pixels that were never deforested, the age is greater than or equal to 32, which is the length of our panel data.[36] Then we make use of the cross-sectional carbon map developed by Englund et al. (2017) for the year 2017,

[35]We combine savanna, grassland and forest from the raw data into a "forest" classification; we define the agriculture/pasture raw classifications as a "deforested" classification; and we combine wetlands, sand, rocky outcrop, and other non-forest classifications into an "other" classification. We allow for transitions between all of these states. We use Mapbiomas' random forest classifications generated year by year and that are prior to the application of the post-classification filters and map integration (after which the final classifications are obtained and that are publicly available online). Post-classification filters apply spatial and temporal filters. Map integration apply a set of specific hierarchical prevalence rules to solve for potential conflicting classifications. For details, see Souza et al. (2020).

[36]Using the path implied by the HMM parameters, in 2017, 62% of the forest was at least 32 years old.

and regress the carbon stock on the forest age and a forest indicator to generate an expected carbon stock for each pixel of a given forest age or non-forest.[37] We also translate the amount of the carbon stock into dollar values, using estimates of the social cost of carbon from the Interagency Working Group on Social Cost of Greenhouse Gases, United States Government (2021). We compare the estimated transition processes and carbon values described above with the coresponding results based on the raw classifications. For more details see Section F.2 of the Online Appendix.

Finally, for our counterfactual exercise, we treat the HMM-based classifications in 2020 as "truth" and simulate the Markov process forward from 2021 to 2040 under two scenarios. Our baseline scenario assumes that, within each tile, deforestation will remain constant at its 2020 level until 2040. Our "no-deforestation" scenario assumes that there will be no deforestation from 2020 to 2040.

6.2 Results

Figure 5 shows the evolution of the deforestation rate, the reforestation rate, and the fraction of land that is forested from 1986 to 2020, using the Mapbiomas raw data and the estimates from the HMM approach.[38] As expected, the raw data exhibits excessive transitions: the raw deforestation and reforestation rates are typically roughly three times the rates obtained under the HMM. The HMM correction therefore suggests that there should be less carbon emissions from deforestation but less carbon sequestration from regrowth than the raw data indicate. The substantial difference in the estimated rates is apparent despite the fact that the fractions of forested land are similar between the HMM and raw estimates, oscillating between 29% and 31.5% over time. (The HMM forest shares are around 0.5 percentage points above the raw data shares.)

These results illustrate that, even though uncorrected land use classifiers can generate reasonable estimates of the level of a given land cover in a given year, the corrected and uncorrected approaches

[37]This carbon map was specifically designed to provide accurate measures of above ground carbon for Brazil by combining information from other carbon maps and a detailed map of land use changes.

[38]For each year, we take the average over all of the tiles, weighting the deforestation rates by the share of land that is forest and the reforestation rate by the fraction of land that is not.

will yield very different results for applications where land use transitions and/or the age distribution is important. In Online Appendix Figure F16, we show the difference in the age distribution of the forest between the raw data and the HMM-based classifcations. Consistent with Figure 5, we find that the raw data generate forests that are excessively young in light of the high deforestation and reforestation rates.

Next, we focus on the differences in the carbon stocks. Table 2, Panel A, presents the overall amount and value of the carbon stocks in 2020, the last year for which we have data. The HMM approach estimates approximately 2.8 billion tons of carbon on the ground, corresponding to a total social value of $774 billion (assuming a social cost of carbon of $76 per ton of CO2).[39] If instead we use the forest age implied from the raw data, we obtain an estimate of just 2.33 billion tons of carbon, which translates into $666 billion. Therefore, there is a $110 billion difference in the value of the carbon stock that results from the differences in the implied age distribution of the forest in the raw and HMM-based classifiers.

Finally, we simulate forward up to 2040 the deforestation and reforestation processes under the baseline and the "no-deforestation" scenarios. Both scenarios are based on the corrected age distribution for 2020 derived from the HMM approach. Table 2, Panel B, shows the results. We find that eliminating all deforestation would preserve 270 million tons of carbon on the ground, which is equivalent to a social benefit of approximately $100 billion dollars.[40]

7 Conclusion

Remotely sensed data have proven useful in the study of a variety of important phenomena, including the pollution incidence, urbanization, land use change, and the evolution of biodiversity. In this paper, we show how econometric tools can be used to improve the measurement of remotely sensed transitions, such as rates of land use change. Relying on a set of assumptions that can be analyzed

[39]The estimated social cost of carbon for 2020, based on the 2.5% discount rate, is $76 per ton of CO2, according to the Interagency Working Group on Social Cost of Greenhouse Gases, United States Government (2021).

[40]For 2040, we use the 2.5% discount rate of the Interagency Working Group estimate, corresponding to a social cost of carbon of $103 per ton of CO2.

on a case-by-case basis, the method avoids the need for ground truth data. In the context of Brazilian land use change, we find the HMM correction performs well and makes an important difference in measured rates of land use change.

References

Abercrombie, S. P. and M. A. Friedl (2016). Improving the consistency of multitemporal land cover maps using a hidden markov model. *IEEE Transactions on Geoscience and Remote Sensing 54*, 703–713.

Alix-Garcia, J. and D. L. Millimet (2021). Remotely incorrect? accounting for nonclassical measurement in satellitedata on deforestation. Technical report, Oregon State University.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica 67*(6), 1341–1383.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica 68*(2), 399–405.

Assunção, J., R. McMillan, J. Murphy, and E. Souza-Rodrigues (2019). Optimal environmental targeting in the Amazon rainforest. Technical report, NBER Working Paper 2536.

Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *The Annals of Probability 10*, 1047–1050.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement error in survey data. *Handbook of econometrics*, 3705–3833.

Brancalion, P. H., L. C. Garcia, R. Loyola, R. R. Rodrigues, V. D. Pillar, and T. M. Lewinsohn (2016). A critical analysis of the Native Vegetation Protection Law of Brazil (2012): updates and ongoing initiatives. *Natureza & Conservacao 14*, 1–15.

Brown, J. C., J. H. Kastens, A. C. Coutinho, D. d. C. Victoria, and C. H. Bishop (2013). Classifying multiyear agricultural land use data from mato grosso using time-series modis vegetation index data. *Remote Sensing of Environment 130*, 39 – 50.

Chen, J., J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu, W. Zhang, X. Tong, and J. Mills (2015). Global land cover mapping at 30m resolution: A pok-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing 103*, 7–27.

Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics 92*(1), 1 – 45.

Coutinho, A., D. d. C. Victoria, A. da Paz, J. Brown, and J. Kastens (2011). Dynamics of agriculture in the soy production pole of the state of mato grosso. In *Proceedings of the Brazilian Symposium of Remote Sensing, Curitiba, Brasil, 30 abril – 5 maio, 2011, INPE (2011)*, pp. 6128–6135.

Czaplewski, R. L. (1992). Misclassification bias in areal estimates. *Photogrammetric Engineering & Remote Sensing 58*(2), 189–192.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), pp. 1–38.

Donaldson, D. and A. Storeygard (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives 30*, 171–198.

Englund, O., G. Sparovek, G. Berndes, F. Freitas, J. P. Ometto, P. V. D. C. E. Oliveira, C. Costa, and D. Lapola (2017, July). A new high-resolution nationwide aboveground carbon map for Brazil. *Geo: Geography and Environment 4*(2), e00045.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik,

C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder (2005). Global consequences of land use. *Science 309*(5734), 570–574.

Fowlie, M., E. Rubin, and R. Walker (2019, May). Bringing satellite-based air quality estimates down to earth. *AEA Papers and Proceedings 109*, 283–88.

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang (2010). Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment 114*(1), 168 – 182.

Geller, G. N., P. N. Halpin, B. Helmuth, E. L. Hestir, A. Skidmore, M. J. Abrams, N. Aguirre, M. Blair, E. Botha, M. Colloff, T. Dawson, J. Franklin, N. Horning, C. James, W. Magnusson, M. J. Santos, S. R. Schill, and K. Williams (2017). *Remote Sensing for Biodiversity*. Springer, Cham.

Goldblatt, R., W. You, G. Hanson, and A. Khandelwal (2016). Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine. *Remote Sensing 8*(8), 634.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment 202*, 18–27.

Harari, M. and E. L. Ferrara (2018). Conflict, Climate, and Cells: A Disaggregated Analysis. *The Review of Economics and Statistics 100*(4), 594–608.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: Chapman & Hall/CRCPress.

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

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics 144*(1), 27–61.

Hu, Y. (2017). The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics. *Journal of Econometrics 200*(2), 154–168.

Hu, Y. (2020). *The Econometrics of Unobservables – Latent Variable and Measurement Error Models and Their Applications in Empirical Industrial Organization and Labor Economics*. Manuscript.

Hu, Y. and S. M. Schennach (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica 76*(1), 195–216.

Interagency Working Group on Social Cost of Greenhouse Gases, United States Government (2021, February). Technical Support Document: Social Cost of Carbon, Methane,. Technical report.

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

Laurance, W. F. (2009). Conserving the hottest of the hotspots. *Biological Conservation 142*(6), 1137–1137.

Lillesand, T., R. W. Kiefer, and J. W. Chipman (2015). *Remote Sensing and Image Interpretation* (7 ed.). John Wiley & Sons.

Marx, B., T. M. Stoker, and T. Suri (2019). There is no free house: Ethnic patronage in a Kenyan slum. *American Economic Journal: Applied Economic* (forthcoming).

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. John Wiley & Sons.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics IV*, 2113–2241.

Politis, D. and J. P. Romano (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics 22*, 2031–2050.

Rangel, M. A. and T. S. Vogl (2019). Agricultural Fires and Health at Birth. *The Review of Economics and Statistics 101*(4), 616–630.

Rosa, M. R., P. H. S. Brancalion, R. Crouzeilles, L. R. Tambosi, P. R. Piffer, F. E. B. Lenti, M. Hirota, E. Santiami, and J. P. Metzger (2021). Hidden destruction of older forests threatens Brazil's Atlantic Forest and challenges restoration programs. *Science Advances 7*(4), eabc4547.

Sandler, A. M. and B. S. Rashford (2018). Misclassification error in satellite imagery data: Implications for empirical land-use models. *Land Use Policy 75*, 530–537.

Schennach, S. M. (2021). Measurement systems. *Journal of Economic Literature* (Forthcoming).

Souza, C. M., J. Z. Shimbo, M. R. Rosa, L. L. Parente, A. A. Alencar, B. F. T. Rudorff, H. Hasenack, M. Matsumoto, L. G. Ferreira, P. W. M. Souza-Filho, S. W. de Oliveira, W. F. Rocha, A. V. Fonseca, C. B. Marques, C. G. Diniz, D. Costa, D. Monteiro, E. R. Rosa, E. Vélez-Martin, E. J. Weber, F. E. B. Lenti, F. F. Paternost, F. G. C. Pareyn, J. V. Siqueira, J. L. Viera, L. C. F. Neto, M. M. Saraiva, M. H. Sales, M. P. G. Salgado, R. Vasconcelos, S. Galano, V. V. Mesquita, and T. Azevedo (2020). Reconstructing three decades of land use and land cover changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sensing 12*(17).

van Donkelaar, A., R. V. Martin, C. Li, and R. T. Burnett (2019). Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology 53*(5), 2595–2611. PMID: 30698001.

van Handel, R. (2008). Hidden Markov Models: Lecture notes. https://www.princeton.edu/~rvan/orf557/hmm080728.pdf. [Online; accessed 2017-06-06].

Xie, Z., R. G. Pontius Jr, J. Huang, and V. Nitivattananon (2020). Enhanced intensity analysis to quantify categorical change and to identify suspicious land transitions: A case study of nanchang, china. *Remote Sensing 12*(20).

	GBM Classification (Y_{it})			*Fraction Correctly Predicted (Recall)*
Embrapa Data (S_{it})	Crops	Pasture	**Total**	
Crops	1409	112	1521	0.926
Pasture	15	58	73	0.795
Total	1424	170	1594	

Table 1: Confusion Matrix based on Embrapa Validation Data

Panel A: Carbon Stock and Social Value of Forest in 2020		
Measurement	Carbon Stock (billion tons)	Social Value (billion dollars)
HMM-Viterbi	2.78	774.35
Raw Data	2.39	666.50
Panel B: Carbon Stock and Social Value in 2040		
Scenario	Carbon Stock (billion tons)	Social Value (billion dollars)
Baseline	2.87	1085.62
No Deforestation	3.14	1185.71

Table 2: Carbon Stock and Social Value of Forest

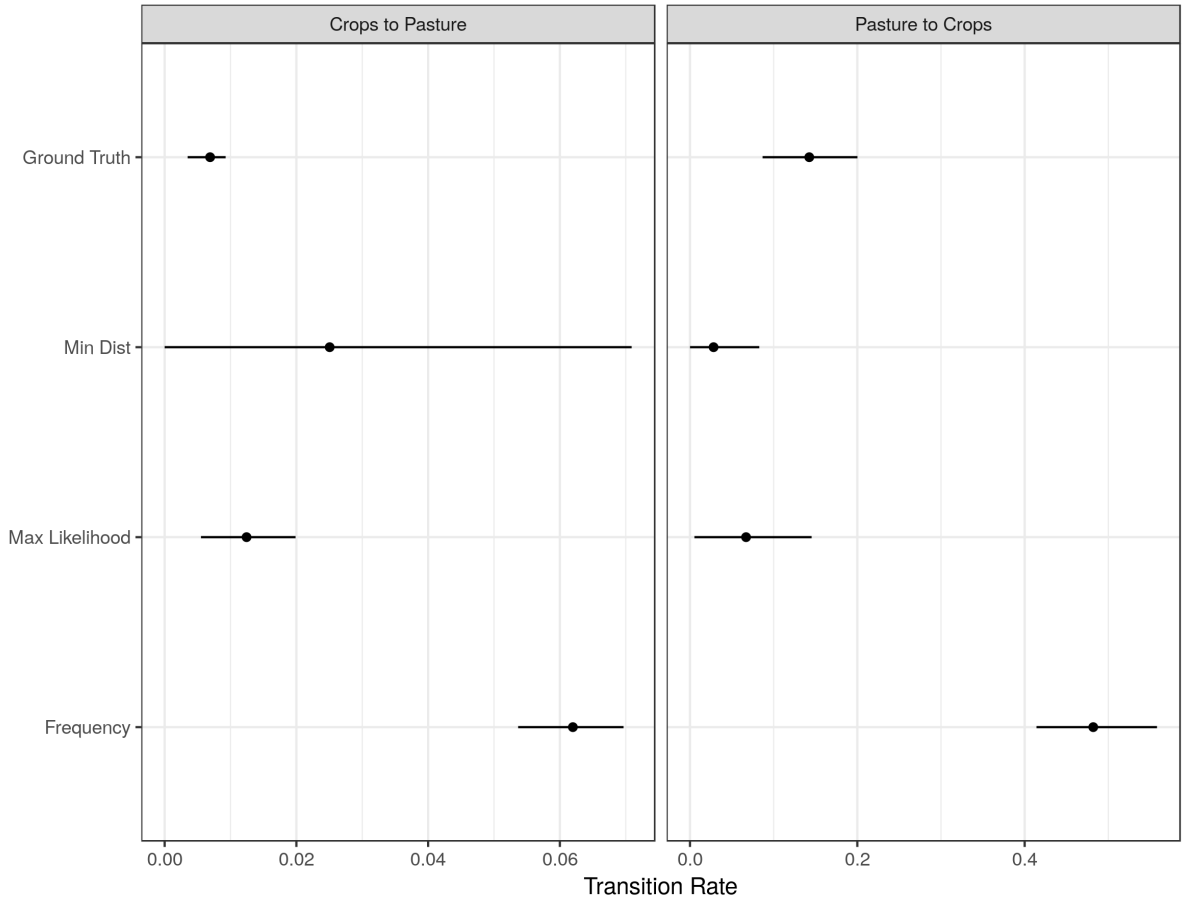


Figure 1: Time-invariant Transition Probabilities – Embrapa Validation Data

Note: *Ground Truth* data are the observed transition probabilities in the Embrapa test set, the *Frequency* estimator uses the GBM based land use classifications to estimate transitions, while *Min Dist* and *Max Likelihood* are the minimum distance and maximum likelihood HMM estimators for the transition rates. Error bars represent 95% confidence intervals based on subsampling. The results shown in this figure combine all years in the Embrapa test set, i.e. they assume time-invariant transition probabilities.

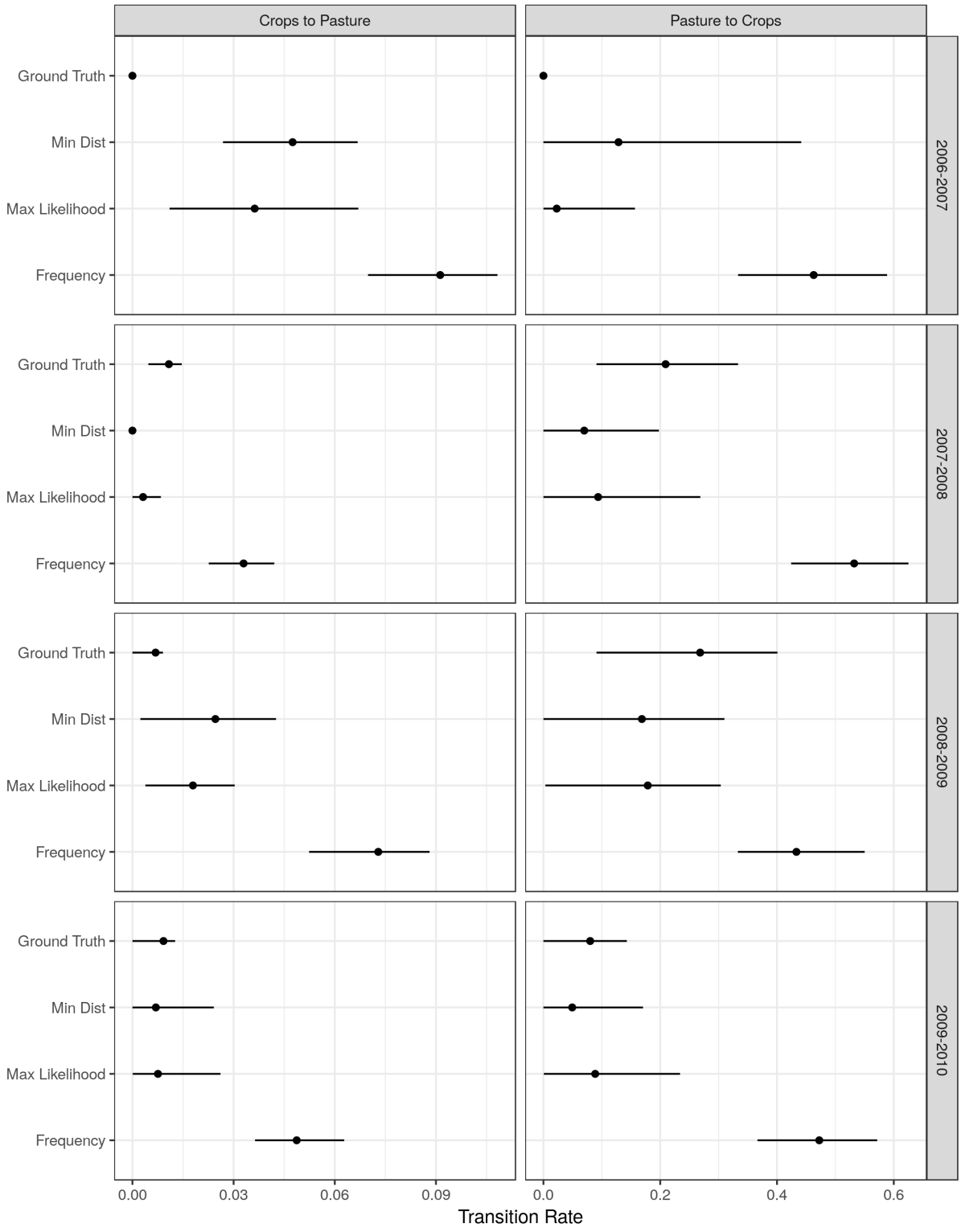


Figure 2: Time-varying Transition Probabilities – Embrapa Validation Data

Note: *Ground Truth* data are the observed transition probabilities in the Embrapa test set, the *Frequency* estimator uses the GBM based land use classifications to estimate transitions, while *Min Dist* and *Max Likelihood* are the minimum distance and maximum likelihood HMM estimators for the transition rates. Error bars represent 95% confidence intervals based on subsampling. The results shown in this figure combine assume time-varying transition probabilities.

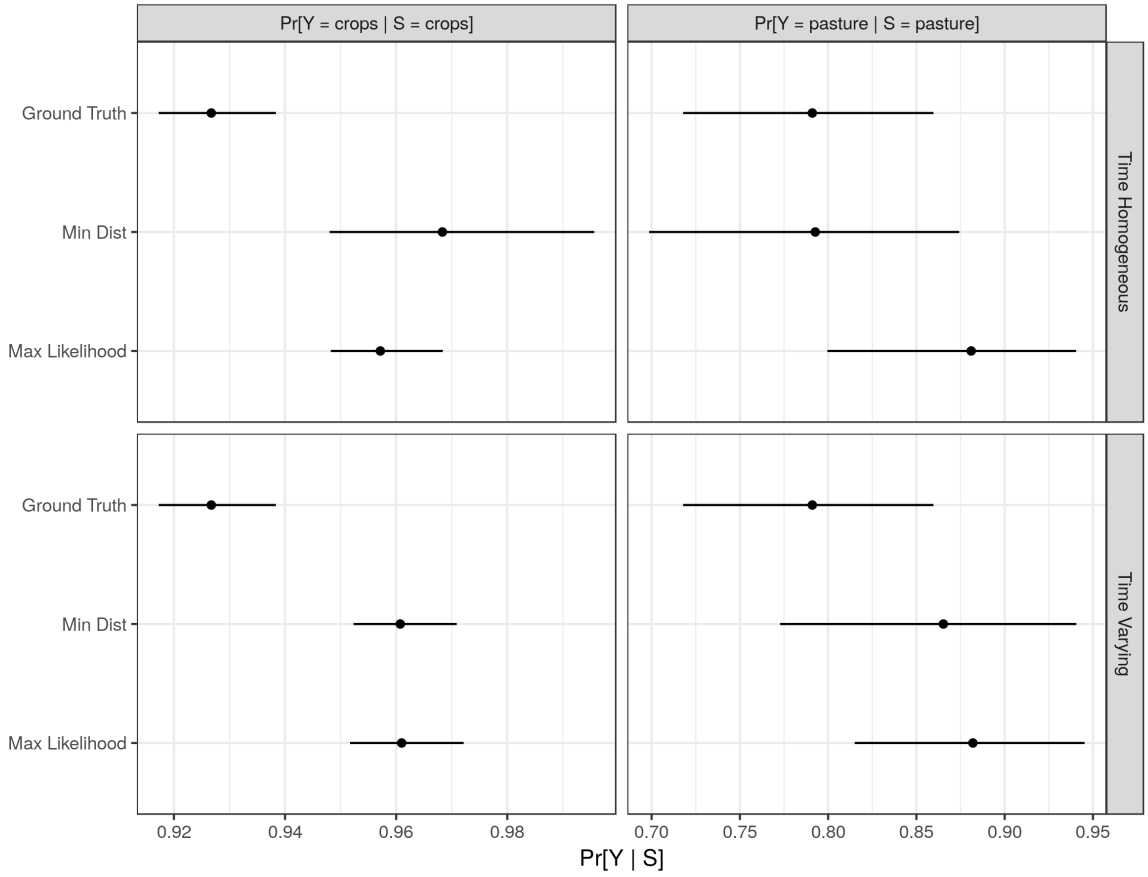


Figure 3: Misclassification Probabilities – Embrapa Validation Data

Note: *Ground Truth* corresponds to the misclassification probabilities from the "confusion matrix" comparing the Embrapa test set points and the GBM predictors. The *Min Dist* and *Max Likelihood* correspond to the minimum distance and maximum likelihood HMM estimates of the misclassification probabilties. Error bars represent 95% confidence intervals based on subsampling. The top panel presents the results based on the restricted model with time-invariant transition probabilities; and the bottom figure, the misclassifications based on the model with time-varying transition probabilities.

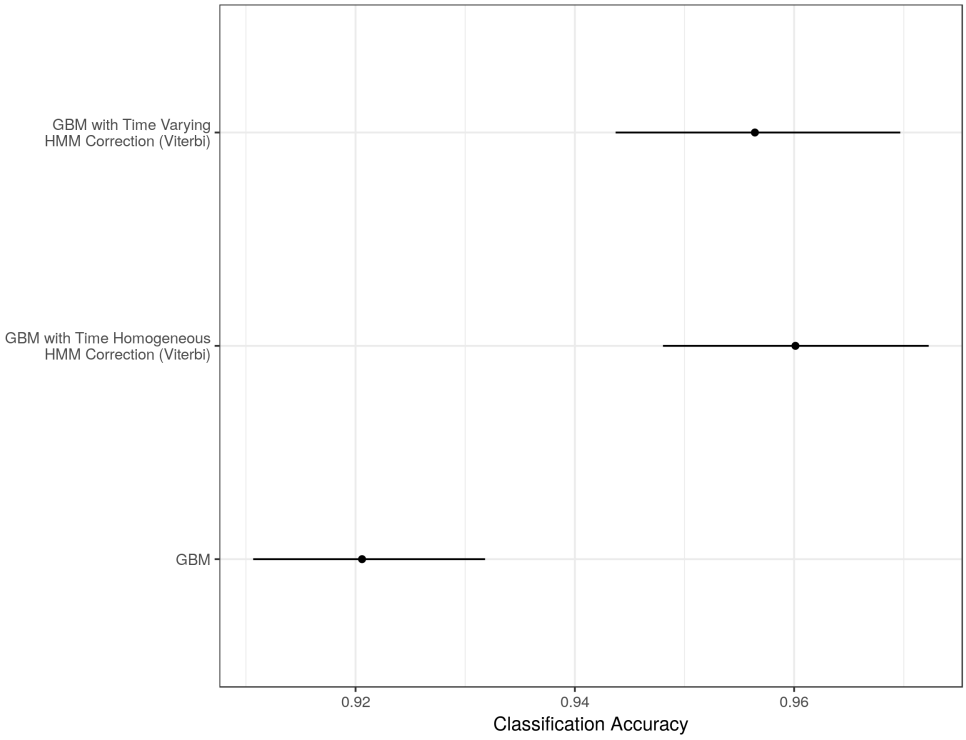
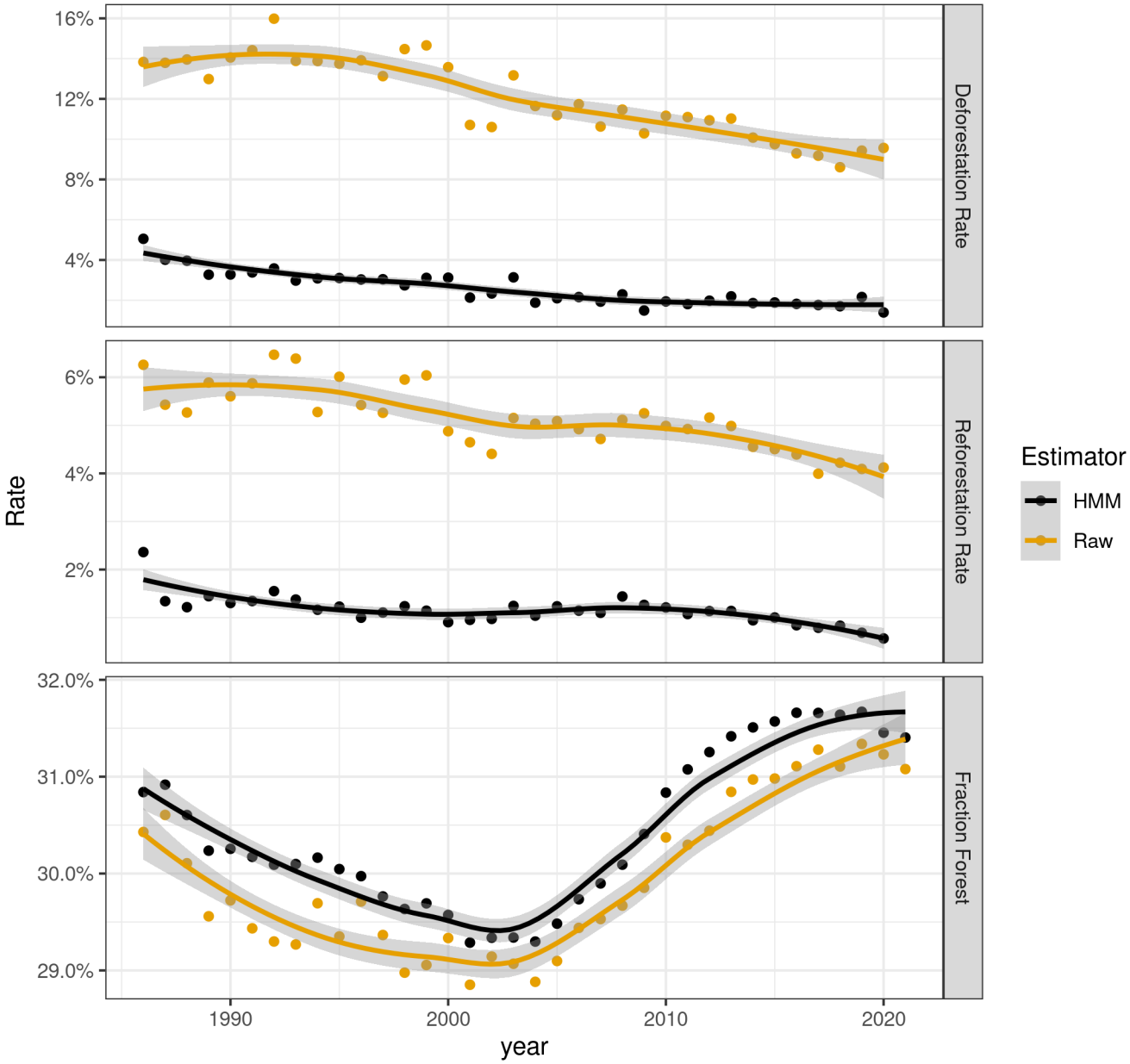


Figure 4: Classification Accuracy of GBM and HMM-Viterbi methods in the Embrapa Validation Data

Note: *GBM* corresponds to the accuracy (i.e, the fraction of correctly predicted points) in the test set of the GBM classifier. The *GBM with Time Homogeneous HMM Correction (Viterbi)* and the *GBM with Time Varying HMM Correction (Viterbi)* correspond to the accuracy of the classifications in the test set based on the Viterbi method, after applying the HMM (maximum likelihood estimator) correction assuming time-homogeneous and time-varying transitions, respectively.



Points reflect the HMM parameters aggregated over all of the tiles
(where the tile-specific deforestation rates are weighted by the fraction forest and the reforestation rates are weighted by the fraction not-forest.) Lines reflect a Loess trend.

Figure 5: Atlantic Forest Trends Over Time

Online Appendix

A Mathematical Derivation of Useful Identities

Under the HMM assumptions, and by the law of total probability, the joint distribution of (Y_{it}, Y_{it-1}) satisfies

$$\Pr\left[Y_{it}, Y_{it-1}\right] = \sum_{s \in \mathcal{S}} \Pr\left[Y_{it} | S_{it} = s\right] \Pr\left[S_{it} = s, Y_{it-1}\right]. \tag{A1}$$

Similarly, the joint distribution of (Y_{it+1}, Y_{it}) is such that

$$\begin{aligned} \Pr\left[Y_{it+1}, Y_{it}\right] &= \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it+1} = s'\right] \Pr\left[S_{it+1} = s' | S_{it} = s\right] \\ &\quad \times \Pr\left[Y_{it} | S_{it} = s\right] \Pr\left[S_{it} = s\right] \\ &= \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it+1} = s'\right] \Pr\left[S_{it+1} = s', S_{it} = s\right] \Pr\left[Y_{it} | S_{it} = s\right], \end{aligned} \tag{A2}$$

where the first equality follows from the law of total probability and the HMM assumption (i.e., equation (2) in the main text); and the second equality uses the fact that $\Pr\left[S_{it+1} = s' | S_{it} = s\right] \Pr\left[S_{it} = s\right] = \Pr\left[S_{it+1} = s', S_{it} = s\right]$.

Finally, the joint distribution of $(Y_{it+1}, Y_{it}, Y_{it-1})$ satisfies

$$\Pr\left[Y_{it+1}, Y_{it}, Y_{it-1}\right] = \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it} = s\right] \Pr\left[Y_{it} | S_{it} = s\right] \Pr\left[Y_{it-1}, S_{it} = s\right], \tag{A3}$$

because

$$
\begin{aligned}
& \Pr\left[Y_{it+1}, Y_{it}, Y_{it-1}\right] \\
= & \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1}, Y_{it}, Y_{it-1}, S_{it} = s', S_{it-1} = s\right] \\
= & \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1} | Y_{it}, S_{it} = s'\right] \Pr\left[Y_{it}, S_{it} = s' | Y_{it-1}, S_{it-1} = s\right] \Pr\left[Y_{it-1}, S_{it-1} = s\right] \\
= & \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it} = s'\right] \Pr\left[Y_{it} | S_{it} = s'\right] \Pr\left[S_{it} = s' | S_{it-1} = s\right] \Pr\left[Y_{it-1}, S_{it-1} = s\right] \\
= & \sum_{s' \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it} = s'\right] \Pr\left[Y_{it} | S_{it} = s'\right] \left(\sum_{s \in \mathcal{S}} \Pr\left[S_{it} = s' | S_{it-1} = s\right] \Pr\left[Y_{it-1}, S_{it-1} = s\right] \right) \\
= & \sum_{s' \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it} = s'\right] \Pr\left[Y_{it} | S_{it} = s'\right] \Pr\left[Y_{it-1}, S_{it} = s'\right],
\end{aligned}
$$

where the first equality follows from the law of total probability; the second equality decomposes the joint distribution in terms of the corresponding conditional distributions; the third equality makes use of the HMM assumption (equation (2)); the fourth equality rearranges the terms in the summations; and the fifth equality follows from the law of total probability.

In matrix notation, equations (A1)–(A3) are equivalent to the equations (3)–(5) presented in the main text.

B Measurement Error under the HMM Assumptions

In this section, we investigate the measurement error in observed transition probabilities under the HMM assumptions. While the implications of (nonclassical) measurement error in discrete variables are well understood (see, e.g., the survey by Schennach, 2021), mismeasured transition probabilities have been less studied in regression analyses. Here, we focus first on deriving the relationship between (a) the observed transitions, (b) the true latent transitions, and (c) the measurement error term, in order to shed light on the type of errors (e.g., classical vs nonclassical) that arises as a consequence of the HMM assumptions. Then, we use this relationship to investigate how it may affect the estimation of regression model parameters.

We assume the researcher is interested in measuring transitions at some regional level, and has access to a panel data with many regions m, where each region is composed of several pixels i. For instance, she may be interested in deforestation rates or pollution trends at the municipality level. Take a pixel i in a region m at time t. (For expositional ease, we omit the subscripts i and m below.) The transition probability of the observed state from y at t to y' at $t+1$ can be written as

$$\begin{aligned}
\Pr\left[Y_{t+1} = y' | Y_t = y\right] &= \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \Pr\left[Y_{t+1} = y', S_{t+1} = s', S_t = s | Y_t = y\right] \\
&= \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \Pr\left[Y_{t+1} = y' | S_{t+1} = s', S_t = s, Y_t = y\right] \Pr\left[S_{t+1} = s', S_t = s | Y_t = y\right] \\
&= \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \Pr\left[Y_{t+1} = y' | S_{t+1} = s'\right] \Pr\left[S_{t+1} = s' | S_t = s\right] \Pr\left[S_t = s | Y_t = y\right],
\end{aligned} \tag{B4}$$

where the third equality follows from the main HMM assumption – equation (2).

To simplify, suppose we have just two states. For concreteness, and following our running example, suppose $\mathcal{S} = \{d, f\}$, where d = deforested and f = forested. Here we focus on the measurement error in the transition probability from forest to deforested (i.e., the deforestation rate), but the reasoning applies to transitions involving any two states. Specifically, take $y' = d$ and $y = f$. Then, (B4) becomes

$$\begin{aligned}
\Pr\left[Y_{t+1} = d | Y_t = f\right] \;=\; & \Pr\left[Y_{t+1} = d | S_{t+1} = d\right] \Pr\left[S_{t+1} = d | S_t = f\right] \Pr\left[S_t = f | Y_t = f\right] \\
& + \Pr\left[Y_{t+1} = d | S_{t+1} = d\right] \Pr\left[S_{t+1} = d | S_t = d\right] \Pr\left[S_t = d | Y_t = f\right] \\
& + \Pr\left[Y_{t+1} = d | S_{t+1} = f\right] \Pr\left[S_{t+1} = f | S_t = f\right] \Pr\left[S_t = f | Y_t = f\right] \\
& + \Pr\left[Y_{t+1} = d | S_{t+1} = f\right] \Pr\left[S_{t+1} = f | S_t = d\right] \Pr\left[S_t = d | Y_t = f\right].
\end{aligned}$$

By rearranging the equation above, and noting that probabilties add up to one, we obtain

$$\begin{aligned}\Pr\left[Y_{t+1}=d|Y_t=f\right] &= \left(\Pr\left[Y_{t+1}=d|S_{t+1}=d\right]-\Pr\left[Y_{t+1}=d|S_{t+1}=f\right]\right)\\ &\quad\times\Pr\left[S_t=f|Y_t=f\right]\Pr\left[S_{t+1}=d|S_t=f\right]\\ &\quad+\Pr\left[Y_{t+1}=d|S_{t+1}=f\right]\\ &\quad+\left(\Pr\left[Y_{t+1}=d|S_{t+1}=d\right]-\Pr\left[Y_{t+1}=d|S_{t+1}=f\right]\right)\\ &\quad\times\Pr\left[S_t=d|Y_t=f\right]\Pr\left[S_{t+1}=d|S_t=d\right]. \qquad \text{(B5)}\end{aligned}$$

Next, denote the *observed* deforestation rate in region m at $t+1$ by $D_{mt+1}=\Pr\left[Y_{t+1}=d|Y_t=f\right]$ (i.e., the share of forested pixels in region m at t that become deforested at $t+1$) and the corresponding *true* deforestation rate by $D^*_{mt+1}=\Pr\left[S_{t+1}=d|S_t=f\right]$. (Note that we allow these rates to vary over the regions m and over time t.) Define also the term multiplying the true deforestation rate in (B5):

$$\begin{aligned}\beta_{mt+1} &= \left(\Pr\left[Y_{t+1}=d|S_{t+1}=d\right]-\Pr\left[Y_{t+1}=d|S_{t+1}=f\right]\right)\times\Pr\left[S_t=f|Y_t=f\right]\\ &= \left(\Pr\left[Y_{t+1}=d|S_{t+1}=d\right]-\Pr\left[Y_{t+1}=d|S_{t+1}=f\right]\right)\\ &\quad\times\Pr\left[Y_t=f|S_t=f\right]\frac{\Pr\left[S_t=f\right]}{\Pr\left[Y_t=f\right]}, \qquad \text{(B6)}\end{aligned}$$

where the second equality holds by the Bayes rule. (Note again that we incorporate the subscripts m and $t+1$ explicitly in the definition of β_{mt+1}, as this term may vary over regions and time periods.) Assuming the classifier is accurate (in the sense that correct classification is more likely than misclassification), the first term on the right-hand side of (B6) is positive, which implies that β_{mt+1} is between zero and one. Clearly, β_{mt+1} depends on the misclassification probabilities and on the ratio of the shares of true and observed forested areas in region m in the previous period t. In general, the greater the percentage of correct classifications, and the higher the share of true forested areas relative to the share of oberved forested areas, the greater the β_{mt+1}.

Next, define the term composed of those in (B5) that are not multiplying the true deforestation

rate:

$$
\begin{aligned}
U_{mt+1} &= \Pr\left[Y_{t+1} = d | S_{t+1} = f\right] + \left(\Pr\left[Y_{t+1} = d | S_{t+1} = d\right] - \Pr\left[Y_{t+1} = d | S_{t+1} = f\right]\right) \\
&\quad \times \Pr\left[S_t = d | Y_t = f\right] \Pr\left[S_{t+1} = d | S_t = d\right] \\
&= \Pr\left[Y_{t+1} = d | S_{t+1} = f\right] + \left(\Pr\left[Y_{t+1} = d | S_{t+1} = d\right] - \Pr\left[Y_{t+1} = d | S_{t+1} = f\right]\right) \\
&\quad \times \left(1 - \Pr\left[Y_t = d | S_t = d\right]\right) \left(\frac{1 - \Pr\left[S_t = f\right]}{\Pr\left[Y_t = f\right]}\right) \Pr\left[S_{t+1} = d | S_t = d\right], \qquad \text{(B7)}
\end{aligned}
$$

where the second equality holds by the Bayes rule. Assuming again that the classifier is accurate, we have that U_{mt+1} is positive. Similar to β_{mt+1}, U_{mt+1} depends on the misclassification probabilities and on the shares of true and observed forested areas in the previous period, but, in contrast to β_{mt+1}, it also depends on the true persistence of the deforested areas ($\Pr\left[S_{t+1} = d | S_t = d\right]$).

Substituting the definitions of D_{mt+1}, D^*_{mt+1}, β_{mt+1}, and U_{mt+1} into equation (B5), and taking the lagged expression, we obtain the following (random-coefficients) model:

$$D_{mt} = \beta_{mt} D^*_{mt} + U_{mt}.$$

Adding and subtracting the average of β_{mt} and U_{mt} (over m and t), denoted by β and α, respectively, we get

$$D_{mt} = \alpha + \beta D^*_{mt} + V_{mt}, \qquad \text{(B8)}$$

where

$$V_{mt} = \left(\beta_{mt} - \beta\right) D^*_{mt} + \left(U_{mt} - \alpha\right). \qquad \text{(B9)}$$

The slope β of the regression equation (B8) – also known as the "factor loading" – is between zero and one (given that β_{mt} is between zero and one for all m and t), provided that the land use classifier is accurate everywhere. This contrasts with standard measurement error models, in which loadings are typically equal to one. The measurement error, V_{mt}, depends on both (a) the interaction between D^*_{mt} and the (mean-zero) random coefficients β_{mt} (which in turn depends on misclassification probabilities and on shares of true and observed forested areas), and (b) the (mean-

zero) term U_{mt} (which also depends on misclassification probabilities and on shares of true and observed forested areas, in addition to on the true persistence of deforested areas). In the absence of ground-truth data and of additional (identifying) assumptions, the residual V_{mt} is unobservable.

While it is non-trivial to derive the exact dependence between D^*_{mt} and V_{mt}, given (B6)–(B9), their correlation seems unlikely to be zero. That is because, on the one hand, transition probabilities are between zero and one, so when $D^*_{mt} = 0$, we must have $D_{mt} \geq 0$, and when $D^*_{mt} = 1$, we must have $D_{mt} \leq 1$, suggesting a negative correlation between D^*_{mt} and V_{mt}. On the other hand, regions in which deforestation is persistent (i.e., places with high levels of $\Pr[S_t = d | S_{t-1} = d]$, and so with high levels of U_{mt}, all else constant) are likely good regions for agricultural production, leading to high deforestation rates, which in turn induces a positive correlation between D^*_{mt} and V_{mt}. These observations suggest that D^*_{mt} and V_{mt} likely correlate, though the direction of the correlation is unclear ex-ante; their probabilistic relationship may even be nonlinear. Either way, the derivation presented here suggests the presence of nonclassical measurement error in transition rates under the HMM assumptions.

B.1 Consequences of HMM Measurement Error for Regression Models

Next, we investigate the consequences of measurement error in transition probabilities for regression models. We focus on mismeasured dependent variables (when the researcher may be interested, say, in the determinants of deforestation or in some causal effect of a policy intervention), but a similar reasoning applies to mismeasured covariates (as when researchers are interested in the health impacts of changes in pollution levels or of changes in fires incidents). We start with the standard linear regression model, then we study the widely used logit model, and extend the investigation to the nested logit model.

Linear Model. Suppose we want to estimate the following simple regression model

$$D^*_{mt} = \gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt},$$

where X_{mt} is a potential determinant of interest, e.g., the price of a commodity like beef or palm oil, or a policy intervention indicator. For simplicity, we assume X_{mt} is uncorrelated with ε_{mt}. Suppose we only observe D_{mt} satisfying the HMM assumptions, and so satisfying equations (B8)–(B9). Let the (possibly nonlinear) dependence between V_{mt} and D^*_{mt} (defined in the previous section) be specified as

$$V_{mt} = h\left(D^*_{mt}\right) + \epsilon_{mt},$$

for some unknown (possibly nonmonotonic) function $h(.)$. Then

$$\begin{aligned} D_{mt} &= \alpha + \beta\, D^*_{mt} + V_{mt} \\ &= \alpha + \beta\, D^*_{mt} + h\left(D^*_{mt}\right) + \epsilon_{mt} \\ &= \alpha + \beta\, \left(\gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt}\right) + h\left(\gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt}\right) + \epsilon_{mt} \\ &= \delta_0 + \delta_1\, X_{mt} + \xi_{mt}, \end{aligned}$$

where $\delta_0 = (\alpha + \beta\gamma_0)$, $\delta_1 = \beta\gamma_1$, and

$$\xi_{mt} = h\left(\gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt}\right) + \epsilon_{mt} + \beta\varepsilon_{mt}.$$

If we regress the mismeasured D_{mt} on X_{mt} using OLS to estimate γ_1, we obtain biased results for two reasons. First, if X_{mt} and ξ_{mt} were uncorrelated, OLS would be unbiased for $\delta_1 = \beta\gamma_1 \neq \gamma_1$, and so it would be biased for γ_1 given that β is between zero and one when the land use classifier is accurate. That leads to an attenuation bias. (When the land use classifier is not accurate, β could be negative, reversing the sign of the estimates.) Second, because X_{mt} may correlate with the unobservable ξ_{mt}, through the term $h(.)$. As discussed previously, $h(.)$ reflects the nonclassical nature of the measurement error in transition probabilities (i.e., it reflects the fact that D^*_{mt} and V_{mt} likely correlate). This correlation can be positive or negative; when the correlation is positive and sufficiently large, it can bias OLS upward. In sum, when transition probability is the dependent variable of interest in a linear regression model, its measurement error most likely biases (without

necessarily attenuating) the OLS estimator.

Logit Model. Consider the following logit model of deforestation rates:

$$\ln\left(\frac{D^*_{mt}}{1-D^*_{mt}}\right) = \gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt},$$

where X_{mt} is independent of ε_{mt}. The right-hand side of this regression equation corresponds to a mean value of a latent variable, and can be interpreted as the mean utility received by agents in region m at t from deforestation. Given that we observe D_{mt} instead of D^*_{mt}, we have

$$\ln\left(\frac{D_{mt}}{1-D_{mt}}\right) = \gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt} + \left[\ln\left(\frac{D_{mt}}{1-D_{mt}}\right) - \ln\left(\frac{D^*_{mt}}{1-D^*_{mt}}\right)\right]. \tag{B10}$$

To have a sense of the effect that the last term on the right-hand-side of (B10) can have on estimation, apply a second-order Taylor approximation to $\ln\left(\frac{D_{mt}}{1-D_{mt}}\right)$ about $\ln\left(\frac{D^*_{mt}}{1-D^*_{mt}}\right)$ to obtain an approximation to equation (B10):

$$\begin{aligned}\ln\left(\tfrac{D_{mt}}{1-D_{mt}}\right) \approx \quad & \gamma_0 + \gamma_1 X_{mt} + \varepsilon_{mt} - \left(\tfrac{1}{D^*_{mt}\left(1-D^*_{mt}\right)}\right)\left(D^*_{mt} - D_{mt}\right) \\ & - \tfrac{1-2D^*_{mt}}{\left(D^*_{mt}\left(1-D^*_{mt}\right)\right)^2}\left(D^*_{mt} - D_{mt}\right)^2.\end{aligned}$$

If the measurement error in deforestation rate were classic, the expectation of the first-order term above would be zero, i.e.,

$$E\left[\left(\frac{1}{D^*_{mt}\left(1-D^*_{mt}\right)}\right)\left(D^*_{mt} - D_{mt}\right)|X_{mt}\right] = 0.$$

However, given that $D^*_{mt} - D_{mt} = -\alpha + (1-\beta)D^*_{mt} - V_{mt}$, where $(1-\beta) > 0$, and that D^*_{mt} likely correlates with V_{mt} (see equation (B8)), the expectation of the first-order term may be different from zero.

Similarly, the second-order term is not mean-zero either. This is easier to see if the measurement error were classic, with $Var(D^*_{mt} - D_{mt}|X_{mt}) = \sigma^2$. In this case, the expectation of the second-order

term above would be

$$E\left[\frac{1-2D^*_{mt}}{\left(D^*_{mt}\left(1-D^*_{mt}\right)\right)^2}\left(D^*_{mt}-D_{mt}\right)^2|X_{mt}\right]=\sigma^2 E\left[\frac{\left(1-2D^*_{mt}\right)}{\left(D^*_{mt}\left(1-D^*_{mt}\right)\right)^2}|X_{mt}\right], \tag{B11}$$

which is not zero. More importantly, not only are the expectations of the first- and second-order terms of the approximation not zero, they also vary with X_{mt}, since D^*_{mt} depends directly on X_{mt}. This means that these terms can create bias in both the estimates of γ_0 and γ_1.

Nested Logit Model. Consider a nested logit model with three elements in the choice set: crops, pasture, and forest. Furthermore, assume that crops and pasture are in the same nest, and forest is in a separate nest. A regression equation for such a model has the following form:

$$\ln\left(\frac{DC_{mt}}{1-DP_{mt}-DC_{mt}}\right)=\gamma_0+\gamma_1 X_{mt}+\lambda\ln\left(\frac{DC_{mt}}{DC_{mt}+DP_{mt}}\right)+\varepsilon_{mt},$$

where DC_{mt} is the observed rate at which forest is converted to cropland and DP_{mt} is the observed rate at which forest is converted to pasture. The new parameter not appearing in the basic logit model above is λ, which controls the degree to which the shocks to latent variables are correlated within the nest. If $\lambda = 0$, we just have a multinomial logit model with no correlation in shocks. As $\lambda \to 1$, the shocks within the nest become highly correlated.

In this nested logit regression equation, measurement error in the transition rates implies both a left-hand-side and a right-hand-side measurement error problems. The first case was discussed previously, in the context of a binary logit model; the second case may induce an attenuation bias in the estimate of λ.

C The EM and the Viterbi Algorithms

We now briefly explain the EM and the Viterbi algorithms.

C.1 The EM Algorithm

To simplify notation, let θ represent the collection of HMM parameters, i.e. θ is a list containing $\Pr[S_{i1}]$, $\Pr[S_{it+1}|S_{it}]$ for $t = 1, ..., T-1$, and $\Pr[Y_{it}|S_{it}]$, for all $t = 1, ..., T$. Let y denote the entire panel of observations $\{y_{it}\}$; similarly, let s denote values of the hidden state for the entire panel. Define the log likelihood

$$l(\theta) \equiv \ln \Pr[Y = y; \theta] \tag{C12}$$

and let

$$J(\theta, \theta') \equiv \sum_s \Pr[S = s|Y = y; \theta'] \ln \left\{ \frac{\Pr[Y = y, S = s; \theta]}{\Pr[Y = y, S = s; \theta']} \right\}. \tag{C13}$$

The EM algorithm begins with an initial guess $\theta^{(1)}$ then alternates between steps 1 and 2 below for iterations $j = 1, 2, \ldots$ until convergence:

1. The expectation (E) step: compute the posteriors $\Pr\left[S|Y = y; \theta^{(j)}\right]$
2. The maximization (M) step: set $\theta^{(j+1)}$ to $\arg\max_\theta J\left(\theta, \theta^{(j)}\right)$

The EM algorithm produces a sequence of parameter estimates for which the log likelihood $l\left(\theta^{(j)}\right)$ is monotonically increasing. In problems where the likelihood function is non-concave, this means the algorithm could converge to a local maximum.

A key aspect of the E-step of the EM algorithm is the Baum-Welch algorithm. It efficiently calculates probabilities of the form

$$\Pr[S_{it}|Y_{i1}, Y_{i2}, \ldots, Y_{iT}],$$

where $t \leq T$. In words, the model allows us to condition on a long sequence of noisy land use classifications at a given spatial point, and make probabilistic statements about the point's true land use at any period in that history. This is valuable if we are interested in land cover at a specific point: the fact that we condition on the entire sequence $Y_{i1}, Y_{i2}, \ldots Y_{iT}$ can potentially improve predictions when compared to classifiers that use only contemporaneous data to predict land use.

For instance, suppose we have 15 years of data at a particular spatial point, and that the land use set is $\mathcal{S} = \{\text{forest}, \text{deforested}\}$. Imagine that our land use prediction model outputs $Y_{it} = \text{forest}$ for the first 10 years, followed by deforestation for a single year, followed by four years of forest.

Intuitively, if our classifier is reasonably accurate but imperfect, we would guess that the isolated deforestation prediction is erroneous and that the true land use was forest for the entire 15 years. This is conceivable given that it takes far longer than a year to regrow forest on newly deforested land, and given the implausibility of all the classifications other than the eleventh being wrong (or at least several of them). Thus, we might in principle simply relabel the eleventh year as "forest". By implementing such ad-hoc reclassifications, one can effectively smooth out implausible transitions in the data. However, while heuristic-based adjustments such as this simple solution improve estimations of transition rates by making use of time-series information, rather than just cross-sectional information (as typically done in annual land cover classifications), such adjustments are at the whim of the researcher and so may be highly arbitrary. Further, they can be incomplete as there may be cases requiring corrections that are not considered by the researcher. Indeed, typical heuristic adjustments do not eliminate excessive transitions in land use applications, as documented by Friedl et al. (2010). In contrast, the HMM approach naturally accomplishes this sort of smoothing by explicitly modeling the probability of errors in predicted land use, along with the transition probabilities in the true underlying state – and with no heuristics nor ad hoc adjustments involved. The amount of smoothing depends on the estimated parameters – in the edge cases where the off-diagonals of Υ are zero, for example, we do not need any smoothing. Identifying the parameters from observed data is therefore crucial in applications, and the Baum-Welch algorithm allows us to smooth out implausible transitions efficiently.

In our application, the M step of the EM algorithm has a closed-form solution. Denote the posterior probabilities by $\pi_{it}[k] \equiv \Pr[S_{it} = k | Y = y; \theta^{(j)}]$ and $\pi_{it}[k, l] \equiv \Pr[S_{it} = k, S_{it+1} = l | Y = y; \theta^{(j)}]$; these can be computed in an efficient forward-backward pass over time using the Baum-Welch algorithm (i.e., the E step), and the calculations can be done in parallel across spatial points given that we are not modeling spatial dependence (i.e., not conditioning on other pixels'

land uses). The updated values of θ are

$$\begin{aligned}
\Pr\left[S_{i1} = k\right]^{(j+1)} &= \frac{\sum_i \pi_{i1}[k]}{\sum_{i,s} \pi_{i1}[s]}, \\
\Pr\left[S_{it+1} = l | S_{it} = k\right]^{(j+1)} &= \frac{\sum_i \pi_{it}[k,l]}{\sum_i \pi_{it}[k]}, \\
\Pr\left[Y_{it} = y | S_{it} = k\right]^{(j+1)} &= \frac{\sum_{i,t:Y_{it}=y} \pi_{it}[k]}{\sum_{i,t} \pi_{it}[k]}.
\end{aligned} \tag{C14}$$

See van Handel (2008) for a reference on the EM algorithm applied to discrete HMMs. Extending the EM algorithm to deal with cases where Y_{it} is missing at random (e.g. due to cloud cover) is straightforward: in the M step update to Υ, the sums in both the numerator and denominator are restricted to cases where Y_{it} is non-missing. Modifying the Baum-Welch algorithm (i.e. the E step) to deal with missingness-at-random in Y_{it} is equally simple, as we only need to compute $\Pr\left[S_{it}|Y_{i1}, Y_{i2}, \ldots, Y_{iT}\right]$ conditioned on the available information for each pixel i. (For instance, if Y_{i2} is missing, we compute $\Pr\left[S_{it}|Y_{i1}, Y_{i3}, \ldots, Y_{iT}\right]$ for all $t \leq T$.)

D The Viterbi Algorithm

The HMM correction is not a classifier *per se*, but it can be used to generate the most likely trajectory of the states for each pixel in the data using the Viterbi algorithm (van Handel, 2008, Chapter 3). The Viterbi algorithm is a dynamic programming algorithm that generates these predictions given the estimated HMM parameters and the history of observations $\{Y_1, Y_2, \ldots, Y_T\}$. Formally, it chooses the sequence $\{s_1, s_2, \ldots, s_T\}$ that maximizes the conditional probability path estimate $\Pr\left[S_1, S_2, \ldots, S_T|Y_1, Y_2, \ldots, Y_T\right]$ for any given pixel.

Briefly, the probability path estimate $\Pr\left[S_1, S_2, \ldots, S_T|Y_1, Y_2, \ldots, Y_T\right]$ can be expressed in terms of initial, transition and misclassification distributions by exploiting the HMM structure and the Bayes formula. Based on such expression, the maximization problem can be solved recursively, as the Bellman equation in dynamic optimization problems, solving for one variable only in each step (see Section 3.3 in van Handel, 2008). Notice that finding the most likely path is different from the

problem of finding the most likely state in a given period $\Pr[S_t|Y_1, Y_2, \ldots, Y_T]$, which is calculated efficiently by the Baum-Welch algorithm, as noted previously.

As a word of caution, while the Viterbi algorithm computes the classification *for a given point*, it may not yield unbiased estimates of land use shares or transition rates for an area (but these can be recovered directly from the HMM, so the Viterbi should not be needed in this circumstance). That is not surprising given that there is an information loss when we go from the knowledge of the full probability distribution to knowing just the most likely outcome. It is worth noting too that the Viterbi algorithm is more likely to be useful in longer time series, when there is more information from the HMM parameters on the likelihood of different paths.

E Monte Carlo Studies

In this section, we present several Monte Carlo experiments to investigate the finite-sample performance of the MD and ML estimators. First, we fix the parameters of the model (the initial distribution, the transition probabilities, and the misclassification probabilities) and vary the sample size (i.e., the number of grid points). Second, we fix the number of grid points and evaluate how the estimators perform at different true transition probabilities, misclassification probabilities, and with different numbers of time periods. Third, we incorporate spatial dependence in our design. Fourth, we investigate the performance of our correction when the HMM model is misspecified; specifically, we allow for serial correlation in misclassification probabilities, violating therefore equation (2) in the main text. Finally, we analyze a simple treatment effects regression where transition rates are the dependent variable and test whether HMM estimates can yield unbiased estimates.

E.1 Basic Setup

We consider two land uses, $\mathcal{S} = \{1, 2\}$, observed in $T = 4$ time periods. The initial distribution over hidden states is

$$\mathbf{P}_{S_1} = (0.9, 0.1)^\intercal ,$$

where the initial share of land cover $s = 1$ is 0.9. The transition matrices are

$$\mathbf{P}_1 \equiv \mathbf{P}_{S_2|S_1} = \begin{pmatrix} 0.96 & 0.04 \\ 0.02 & 0.98 \end{pmatrix}, \quad \mathbf{P}_2 \equiv \mathbf{P}_{S_3|S_2} = \begin{pmatrix} 0.9 & 0.1 \\ 0.02 & 0.98 \end{pmatrix}, \quad \mathbf{P}_3 \equiv \mathbf{P}_{S_4|S_3} = \begin{pmatrix} 0.8 & 0.2 \\ 0.02 & 0.98 \end{pmatrix}.$$

So the probability that a pixel i with land cover $s = 1$ in period $t = 1$ stays with the same land cover in the next time period, $t = 2$, is $\Pr[S_{i2} = 1|S_{i1} = 1] = 0.96$. The transition probability decreases to $\Pr[S_{i3} = 1|S_{i2} = 1] = 0.9$ in the next period $t = 3$, and decreases further to $\Pr[S_{i4} = 1|S_{i3} = 1] = 0.8$ in the last period $t = 4$. To simplify, we keep the transitions conditioned on state $s = 2$ the same over time: $\Pr[S_{it+1} = 2|S_{it} = 2] = .98$ for all t.

The misclassification probabilities are time-invariant and given by

$$\Upsilon = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix}.$$

Recall that the elements of Υ are $\Pr[Y_{it} = y|S_{it} = s]$ (with Y_{it} along the rows and S_{it} along the columns). This means that the probability of classifying land use $y = 1$ when the true land cover is actually $s = 2$ is just $\Pr[Y_{it} = 1|S_{it} = 2] = 0.2$. Correct classification probabilities are 0.9 (for $s = 1$) and 0.8 (for $s = 2$), which are within the range of accuracies observed in practice in typical land cover classifications.

The HMM generates the observed transitions for Y_{it}:

$$\mathbf{P}_{Y_2|Y_1} = \begin{pmatrix} 0.815 & 0.185 \\ 0.363 & 0.637 \end{pmatrix}, \quad \mathbf{P}_{Y_3|Y_2} = \begin{pmatrix} 0.775 & 0.225 \\ 0.37 & 0.63 \end{pmatrix}, \quad \mathbf{P}_{Y_4|Y_3} = \begin{pmatrix} 0.72 & 0.28 \\ 0.472 & 0.528 \end{pmatrix}.$$

These transitions put much greater probabilities on the off-diagonals than the true transitions. (E.g., $\Pr[Y_{i2} = 2|Y_{i1} = 1] = 0.185$ while $\Pr[S_{i2} = 2|S_{i1} = 1] = 0.04$.) This implies excessive land cover switching. Frequency estimators of the transition probabilities for Y_{it} are consistent for $\mathbf{P}_{Y_{t+1}|Y_t}$, and are therefore inconsistent for the true transitions $\mathbf{P}_{S_{t+1}|S_t}$.

To evaluate the performance of the proposed HMM corrections, based on the MD and ML estimators, we generated samples with $N = 100$, $N = 500$, $N = 1,000$, $N = 10,000$ spatial grid points, observed for $T = 4$ time periods. For each sample size, we generate 100 Monte Carlo replications. In each replication, we estimate the observed transitions for Y_{it} using frequency estimators, and run both MD and ML estimator starting from six randomly chosen initial values. The initial values for the diagonals of the true $\mathbf{P}_{S_{t+1}|S_t}$ and Υ matrices are i.i.d. uniform on $[0.6, 0.98]$. The initial values for the first element of the initial distribution P_{S_1} are drawn i.i.d. uniform on $[.85, .95]$. For the MD estimator we take the identity matrix as the weighting matrix, $\mathbf{W} = \mathbf{I}$, we used both classifications, $y_{t+1} = 1$ and $y_{t+1} = 2$, as they both satisfy Condition 4.

E.2 Baseline Results

Table F1 presents the average bias, the standard deviation, and the mean-squared error across the Monte Carlo replications (on the rows). For each parameter, we show results for the frequency estimator, the MD, and the ML estimators (on the columns).

As expected, the performances of the MD and ML estimators in terms of the average bias and mean-square errors are substantially better than the performance of the frequency estimator for both the initial distribution of land cover and the transition rates. Naturally, both corrections improve with the sample size, while the frequency estimator does not. The HMM corrections also estimate the misclassification probabilities accurately.

As the table shows, the ML often dominates the MD estimator by having smaller biases. Also, especially for smaller sample sizes, the ML has much smaller standard deviations than the MD estimator. This is not surprising given that the maximum likelihood estimator is efficient. This can be seen graphically in Figure F3, where we show the distribution across replications of the estimated

transition probabilities $\Pr[S_{it+1} = 2|S_{it} = 1]$, and misclassification probabilities $\Pr[Y_{it} = 2|S_{it} = 1]$, using box and whisker plots. The true parameter values are marked by dotted lines. The variability of the MD estimator suggests some caution when using it in small samples. (These graphs slightly understate the observed variability of the MD estimator, since the graph is truncated at .5 and some estimated values go above that.) Indeed, in our experience, the greater standard deviation of the MD estimator (compared to the ML) implies a higher frequency of estimated transition probabilities that are too close to, or exactly at, the boundary of the parameter space. That happens more frequently when true transition probabilities are near zero or one.

While not shown in the table, the ML takes longer to converge than the MD estimator. That is because the EM algorithm loops over the entire panel in its E and M steps; by contrast, the minimum distance estimator loops over the entire panel only once to compute frequency estimators of the joint distribution of Y_{it}, and can then evaluate its objective function quickly by looping only over time, as opposed to the entire panel. These considerations suggest combining the MD and ML in practice, whenever possible, taking into account their strengths. Indeed, when using the (fast) MD estimator as the initial value for the (asymptotically more efficient) ML estimator in the simulations, we find that the "MD followed by ML" approach takes longer to converge than the MD alone, but it is substantially faster than the ML alone (as expected). Specifically, in our baseline setting, MD takes around 0.5 second on average to converge; the ML using MD as initial values takes about 19.5 seconds on average (and runs for 3 iterations); and the ML alone with random initialization takes around 188 seconds on average (and runs for 30 iterations). So, MD followed by ML is about 10 times faster than ML alone. And their performances are similar in terms of bias, variance, and mean-square error, as might be expected.

We also verify the performance of the estimator with $T = 5$ and $T = 6$. Relative to our $T = 4$ period baseline, we fix the transition probabilities for the first and last period and set the transitions for the middle periods equal to each other.[41] While the additional time periods require the estimation

[41] Specifically, we set

$$\mathbf{P}_{S_2|S_1} = \begin{pmatrix} 0.96 & 0.04 \\ 0.02 & 0.98 \end{pmatrix}, \quad \mathbf{P}_{S_t|S_{t-1}} = \begin{pmatrix} 0.9 & 0.1 \\ 0.02 & 0.98 \end{pmatrix}, \forall 1 < t < T, \text{ and } \quad \mathbf{P}_{S_T|S_{T-1}} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.98 \end{pmatrix}.$$

of additional parameters, the larger number of time periods could help improve the precision of the misclassification probability estimates. In Figure F4, we replicate the results from Figure F3 with $N = 1,000$ observations and $T = 4, 5$, and 6 time periods. As these graphs show, the results are similar across the different number of time periods.

E.3 Varying Parameter Configurations

We now fix the sample size at $N = 1,000$ and $T = 4$, and investigate the performance of the HMM corrections for several different parameter configurations. In particular, we hold fixed the transition probabilities of land use at the levels described before and vary the misclassification probabilities for the hidden state $s = 1$. Then we hold fixed the misclassification probabilities and vary the transition probability for state $s = 1$ in the last period.

Figure F5 presents the results for when we vary the misclassification probability for state 1, $\Pr[Y_{it} = 2|S_{it} = 1]$ (i.e., $\Upsilon(2, 1)$), between 5 and 25 percent, while holding other parameters fixed. The top panel shows the behavior of the estimates of the transition probabilities $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$, and the bottom panel shows the behavior of the estimates of the misclassification probability $\Pr[Y_{it} = 2|S_{it} = 1]$. The lines are non-parametric loess regression lines with a shaded 95% confidence interval, where the data is fit from the different Monte Carlo simulations.

Intuitively, as the true misclassification probability increases, the frequency estimates of the transitions increase for every period, even though the actual transition rate is constant. In other words, the frequency estimator predicts many more transitions than actually occur. In contrast, the MD and the ML estimators predict a flatter transition rate. Also, the MD performance degrades for the transition probabilities as the misclassification rate increases. While it is unclear why that happens, these results suggest that the ML estimator might be preferred in practice when the main object of interest is the transition probability, $\Pr[S_{it+1}|S_{it}]$. In constrast, when we look at the estimates of the misclassification rate, $\Pr[Y_{it}|S_{it}]$, the estimates are more similar for the MD and ML approaches, but the ML is more biased as the true misclassification rate increases.

Figure F6 presents the results for when we vary the transition probability for hidden state $s = 1$ in the last period, $\Pr[S_{i4} = 2|S_{i3} = 1]$ (i.e., $\mathbf{P}_3(1,2)$), between 5 and 40 percent. The format of these graphs is similar to those in Figure F5. These graphs show that both MD and ML estimators continue to perform well at estimating transitions and misclassification rates with no notable differences between them (aside from those discussed above).

E.4 Spatial Dependence and Serial Correlation

We now incorporate spatial dependence and serial correlation in our Monte Carlo exercises. For each specification, we run 100 simulations.[42]

Set up. To allow for spatial correlation in the true transition process, $\Pr[S_{it+1}|S_{it}]$, we first arrange all pixels in a two dimensional square lattice of dimension 100-by-100 – i.e. we observe 10,000 pixels per time period. The lattice is partitioned into 100 square "fields" of 100 pixels each (so that each field is composed of 10-by-10 pixels). We assume the true land use follows a first-order Markov process *at the field level*, meaning that we always have $S_{it} = S_{jt}$ when pixels i and j belong to the same field, and that S_{it} and S_{jt} are fully independent otherwise. Intuitively, one can think of the fields as being parcels of land managed by the same person, and that different fields are managed by different (independent) farmers. This is plausible in empirical applications and it satisfies the spatial weak dependence assumption (Conley, 1999). (Note that when fields contain just one pixel, there is no spatial dependence in S_{it}, and the model presented here coincides with the one covered in our previous Monte Carlo exercises.) The initial distribution over the hidden

[42]We have also incorporated missing data that are missing at random, reflecting the common practical issue of (random) clouds preventing full land use classifications. Specifically, we randomly select 10% of the pixels in every period to be unobserved (i.e., not classified as either $s = 1$ nor $s = 2$), and ran the same set of specifications described below. As expected, observations that are missing-at-random do not bias our estimators, but increase their variances. In the interests of space, we do not present these simulated results here.

states is $\mathbf{P}_{S_1} = (0.7, 0.3)^\intercal$, and the transition matrices are

$$\mathbf{P}_1 \equiv \mathbf{P}_{S_2|S_1} = \begin{pmatrix} 0.96 & 0.04 \\ 0.02 & 0.98 \end{pmatrix}, \quad \mathbf{P}_2 \equiv \mathbf{P}_{S_3|S_2} = \begin{pmatrix} 0.9 & 0.1 \\ 0.07 & 0.93 \end{pmatrix}, \quad \mathbf{P}_3 \equiv \mathbf{P}_{S_4|S_3} = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}.$$

We also allow for spatial dependence (and serial correlation) in the misclassification probabilities, $\Pr[Y_{it}|S_{it}]$, in a parsimonious way. To that end, we introduce the variable $Z_{it} \in \{-1, 1\}$, which captures the difficulty is classifying the land cover correctly: when $Z_{it} = 1$, the probability that the machine learning classifier makes a mistake is higher than when $Z_{it} = -1$. We then adjust equation (2) by conditioning it on Z_{it}, so that $\Pr[Y_{it+1}, S_{it+1} | \{Y_{it-h}, S_{it-h}\}_{h\geq 0}, Z_{it+1}] = \Pr[Y_{it+1}|S_{it+1}, Z_{it+1}] \times \Pr[S_{it+1}|S_{it}]$. Abusing notation slightly, we set the misclassification probabilities to be:

$$\Pr[Y_{it} \mid S_{it}, Z_{it} = -1] = \begin{bmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{bmatrix}, \text{ and}$$

$$\Pr[Y_{it} \mid S_{it}, Z_{it} = 1] = \begin{bmatrix} 0.81 & 0.19 \\ 0.39 & 0.61 \end{bmatrix}.$$

In our simulations, half of the pixels are difficult to classify (i.e. $\Pr[Z_{it} = -1] = \Pr[Z_{it} = 1] = 1/2$), implying an overall misclassification probabilities of

$$\Upsilon = 0.5 \cdot \Pr[Y_{it} \mid S_{it}, Z_{it} = 1] + 0.5 \cdot \Pr[Y_{it} \mid S_{it}, Z_{it} = -1] = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}, \tag{E15}$$

which equals the misclassification probabilities in the basic setup, presented in Section E.1.

We do not include Z_{it} in the data set, so when we estimate the model, we *do not* condition on Z_{it}. In this way, any spatial dependence and serial correlation in Z_{it} will be translated into spatial and serial correlation in misclassification. Note that both MD and ML estimators will estimate the (unconditional) Υ given by (E15). (Note also that had we conditioned on Z_{it} when estimating the parameters, we would return to the standard HMM model presented in the main text – but now we

would be able to estimate the (conditional on Z) misclassification probabilities separately.)

We model spatial dependence in Z_{it} in the following way. First, we assume Z_{it} is generated according to the Ising model, which specifies a join distribution of binary random variables over the lattice in a given time period t (see, e.g., Hastie et al., 2015, Chapter 9).[43] The degree of spatial correlation is controlled by the Ising "temperature" parameter, denoted here by β. When $\beta = 0$, there is no spatial correlation; when $\beta > 0$ there is positive spatial correlation, and the higher the value that β takes, the stronger the spatial correlation. Misclassifications are spatially (but not temporally) correlated when $\beta > 0$ and Z_{it} is i.i.d. over time. In our simulations, we set either $\beta = 0$ or $\beta = 2$. (The qualitative results are similar when we consider other values for β.)

To incorporate serial correlation in misclassifications, we allow Z_{it} to correlate over time; to simplify, we assume perfect correlation (i.e., Z_{it} is time-invariant). In this way, we force misclassification probabilities to depend on past values of (Y_{it}, S_{it}).[44] Importantly, when that happens, the main HMM assumption – equation (2) – is violated and our correction is not guaranteed to work.

Results. We start discussing the results for when there is spatial correlation in the true transition process, $\Pr[S_{it+1}|S_{it}]$, but no spatial, nor serial correlation in misclassification. Figure F7 presents an example of the initial distribution, and the evolution of both the true and the observed land uses in the lattice, in panels (a) and (b), respectively. It is clear from the figure that the pixels' outcomes are spatially correlated, and that misclassifications have no spatial nor temporal dependence (given that the same probability distribution $\Pr[Y_{it}|S_{it}]$ holds everywhere and in all time periods).

[43] Specifically, take the vector $Z = (Z_1, \ldots, Z_N)$, with $Z_i \in \{-1, 1\}$ for all pixels i. (We omit the time subscript to simplify.) The Ising model sets the joint probability distribution for Z to be $\Pr[Z = z] = \frac{\exp\{H(z)\}}{A}$, where $H(z) = h\sum_i z_i + \beta \sum_{i,j} z_i z_j$, which is called the Hamiltonian function; and $A = \sum_z \exp H(z)$, which is the normalization constant. The parameter h indicates whether $Z_i = 1$ is more likely (when $h > 0$) or whether $Z_i = -1$ is more likely (when $h < 0$); we set $h = 0$ in our simulations to retain symmetry. The temperature parameter is β, indicating positive correlation across pixels in the lattice (when $\beta > 0$), or negative correlation (when $\beta < 0$), or no correlation (when $\beta = 0$); we set β to be equal to 0 or 2 in the simulations.

[44] That is because Z_{it} is not part of the data set, as mentioned previously: had we conditioned on Z_{it}, the serial correlation in misclassification would disappear. This is similar to a linear panel data model with fixed effects. To see the connection, suppose we have $Y_{it} = Z_i + \varepsilon_{it}$, where Z_i is a fixed effect and ε_{it} is i.i.d. shocks. Then, conditional on Z_i, there is no serial correlation in Y_{it}; but *there exists* serial correlation in Y_{it} when we do not condition on Z_i (since $Y_{it} = Y_{it-1} + \varepsilon_{it} - \varepsilon_{it-1}$).

Figure F8 presents the estimated results across the simulations. As expected, both MD and ML estimators are unbiased for both the transition and misclassification probabilities, given that the HMM is correctly specified at the pixel level, while the frequency estimator is severely biased. In addition, since neither the MD nor the ML estimators make use of all the possible information available (namely, the pixels' spatial correlation), their variances are larger here when compared to the i.i.d. case presented in Section E.2.

Next, we add spatial dependence in misclassification (but still no serial correlation). As mentioned previously, we incorporate the variable Z_{it} in the data generating process (but not in the data), assuming it is i.i.d. over time and fixing the "temperature" parameter to be $\beta = 2$. Figure F9 presents an example of the evolution of the true land uses (in panel (a)), the distribution of Z_{it} (in panel (b)), and the evolution of the observed land uses (in panel (c)). The evolution of true land use is similar to the previous example; the distribution of Z_{it} is highly correlated in the lattice as well, leading to spatially correlated classification errors in any given time period, as can be seen by contrasting the panels (a) and (c) of the figure.

Figure F10 presents the estimated results for this scenario. As before, both MD and ML estimators are unbiased. But now they have even higher variances as neither spatial correlation in land uses nor in misclassification are incorporated explicitly in the estimation strategy. The frequency estimator continues to be highly biased for transitions.

In the next simulations, we drop the spatial correlation in Z_{it} and make this variable constant over time. This translates into misclassifications that are serially (but not spatially) correlated. Importantly, this renders the HMM model misspecified. Figure F11 shows the evolution of the true land uses (in panel (a)), the distribution of Z_{it} (in panel (b)), and the evolution of the observed land uses (in panel (c)) for one simulated example. We observe the same patterns as in the previous case with two differences: there is no spatial correlation in classification errors, but the errors tend to persist over time. In terms of the estimated results, presented in Figure F12, the MD and ML estimators are now biased for transitions (though not substantially) and for misclassification probabilities (particularly so for $\Pr[Y_{it} = 2|S_{it} = 2]$). Yet, the frequency estimator is significantly

more biased than the HMM corrections.

Finally, we incorporate spatially dependent and time-invariant Z_{it}, imposing therefore both spatially and serially correlated misclassifications. Figure F13 shows one simulated example, and Figure F14 presents the estimated results. As expected, the MD and ML estimators are biased, given that the HMM model is misspecified, but not substantially so for the transition probabilities (though it is more biased for the misclassification probabilities). Once again, the frequency estimator shows significant biases for the estimated transition process.

E.5 Regression Analysis

We extend our baseline Monte Carlo simulations to illustrate an application of the HMM in a regression context. We simulate a policy that reduces the deforestation rate in the regions where it is implemented and compare the HMM and raw data approaches to estimating the treatment effect.

For this application, we use the baseline HMM from Section E.1 with two land uses observed in four periods and the transition and misclassification matrices as described above. We assume that this model describes land use transitions in 100 regions and that within each region we observe 1000 pixels. The only change from the baseline set up is that in the transition from $T = 3$ to $T = 4$ the probability of transitioning from state 1 to state 2 is 0.1 instead of 0.2 in 20 of the regions (the "treated" regions). For concreteness, we consider transition rates from state 1 to state 2 as the "deforestation rate."

The researcher is interested in estimating the difference in this transition probability between the treated and untreated regions and uses a simple cross-sectional regression framework,

$$D_m = \alpha + \beta T_m + \epsilon_m,$$

where D_m is the deforestation rate in region m, T_m is a binary variable reflecting whether the region was treated, and ϵ_m is the error term, with $E[\epsilon_m T_m] = 0$. In this setup, the researcher would use the period four deforestation rates from these 100 regions in the estimation. This framework could

easily be extended to difference-in-differences type regression analysis, but for simplicity we do not here.

In Figure F15, we show the distribution of the estimated parameters α (the baseline) and β (the "treatment effect") from 100 Monte Carlo simulations. We consider three different scenarios:

Ground Truth. The estimated deforestation rate D_r is based on ground truth. In the context of the model, this is $\Pr[S_4 = 2|S_3 = 1]$.

HMM-ML. The estimated deforestation rate D_r is based on the HMM maximum likelihood estimate for $\Pr[S_4 = 2|S_3 = 1]$.

Observations. The estimated deforestation rate D_r is based upon the raw classifier, without applying the HMM correction. In the context of the model, this corresponds to $\Pr[Y_4 = 2|Y_3 = 1]$.

We find that the ground truth data yields a precise and unbiased estimate of the true treatment effect of $\beta = -0.1$ and of the baseline $\alpha = 0.2$. The HMM-ML approach gives a precise estimate of the baseline deforestation rate and a close to unbiased measure of the treatment effect. The raw classifier yields biased estimates of the baseline deforestation rate and of the treatment effect, with an estimated effect that is closer to zero than the truth. This further illustrates the point from Section B.1 of this appendix that misclassifications can lead to biased parameter estimates in regressions, even when the measurement error is in the dependent variable.

F Additional Details on the Carbon Stock Application

F.1 Distribution of Forest Age

In Figure F16, we plot the cumulative distributions of the forest age for both the raw and the HMM-based approaches. The graph illustrates that the forest age predicted by the raw data is significantly younger than that predicted by the HMM-based estimates. That is a direct result of the high deforestation and reforestation rates obtained from the raw data: a pixel is more likely to

be deforested and then reforested, leading to a young forest, while the HMM estimates suggest that a pixel is less likely to be disturbed, resulting in older forests.

F.2 Relationship Between Carbon and Forest Age

We estimate the carbon stock for a given forest age using data on the 2017 carbon stock from Englund et al. (2017) and our HMM-based estimates of forest age for each pixel. First, we show informally the relationship between the carbon stock and the age of the forest for 2017 in Figure F17. As expected, the graph shows an increase in the carbon stock as the forest age.[45]

Next, we estimate the following regression model:

$$cs_i = \alpha + \beta\, forest_i + \gamma\, a_i\, I(a_i < \bar{a}_{max})\, forest_i + \delta\, I(a_i > \bar{a}_{max})\, forest_i + u_i,$$

where cs_i is the carbon stock of pixel i in 2017; $forest_i$ is an indicator variable for whether the pixel is classified as forest in 2017 in the HMM-Viterbi sequence; a_i is the forest age of pixel i in 2017 given from the HMM-Viterbi sequence; $\bar{a}_{max}$ is the maximum age we can detect given our data (which corresponds to 32 years old); $I(.)$ is the indicator function; u_i is an idiosyncratic shock; and $(\alpha, \beta, \gamma, \delta)$ are the regression parameters.

In this regression, (a) we allow for forest to have a different baseline level of carbon from non-forest, captured by the coefficient β; (b) we model a linear relationship between the age of the forest (between 1-32 years old) and the carbon stock, captured by γ; and (c) we allow for a different mean level of carbon for forest that is over 32 years old (i.e., pixels that were classified as forest for all years of our sample), captured by δ.

The results are presented in Table F2. We find that the average baseline level of carbon in forests is approximately 4 tons greater than in non-forest pixels. Every additional year the pixel remains forested adds approximately 0.6 tons of carbon, on average. For forests that are over 32 years old, the average amount of carbon in a pixel is approximately 51.5 tons ($= \alpha + \beta + \delta$).

[45] We do not include in this graph any points that were classified as forest for the entirety of our sample, since we do not know their age.

		N=100			N=500			N=1000			N=10000		
		Freq	MD	ML	Freq	MD	ML	Freq	MD	ML	Freq	MD	ML
	Bias	-0.076	-0.022	-0.031	-0.072	-0.013	-0.014	-0.070	-0.005	-0.008	-0.071	-0.002	-0.007
$P_{S_1} = .9$	s.d.	0.039	0.081	0.057	0.018	0.038	0.028	0.012	0.024	0.020	0.004	0.007	0.008
	RMSE	0.085	0.083	0.064	0.074	0.040	0.031	0.071	0.024	0.022	0.071	0.008	0.011
	Bias		0.008	-0.018		-0.004	-0.008		-0.003	-0.006		-0.001	-0.004
$\Upsilon(2,1) = .1$	s.d.		0.053	0.040		0.018	0.014		0.011	0.010		0.004	0.004
	RMSE		0.053	0.043		0.018	0.016		0.012	0.011		0.004	0.006
	Bias		0.098	-0.058		-0.007	-0.025		-0.008	-0.020		-0.002	-0.006
$\Upsilon(1,2) = .2$	s.d.		0.198	0.108		0.096	0.059		0.051	0.044		0.017	0.017
	RMSE		0.220	0.122		0.096	0.064		0.051	0.048		0.017	0.018
	Bias	0.113	0.028	0.027	0.106	0.010	0.008	0.104	0.007	0.006	0.104	0.002	0.004
$P_1(1,2) = .04$	s.d.	0.039	0.065	0.055	0.016	0.027	0.021	0.011	0.017	0.014	0.004	0.006	0.005
	RMSE	0.120	0.070	0.061	0.107	0.028	0.022	0.105	0.018	0.015	0.104	0.006	0.006
	Bias	0.528	0.149	0.189	0.540	0.081	0.115	0.539	0.056	0.090	0.543	0.023	0.069
$P_1(2,1) = .02$	s.d.	0.136	0.227	0.219	0.058	0.135	0.123	0.040	0.090	0.081	0.012	0.049	0.041
	RMSE	0.545	0.271	0.288	0.543	0.157	0.168	0.541	0.106	0.121	0.544	0.054	0.080
	Bias	0.093	0.025	0.012	0.089	0.001	0.004	0.089	0.001	0.002	0.090	0.000	0.003
$P_2(1,2) = .1$	s.d.	0.043	0.103	0.062	0.019	0.031	0.028	0.012	0.019	0.018	0.005	0.007	0.007
	RMSE	0.103	0.105	0.063	0.091	0.031	0.028	0.090	0.018	0.018	0.090	0.007	0.007
	Bias	0.479	0.113	0.104	0.474	0.045	0.049	0.468	0.032	0.036	0.468	0.006	0.018
$P_2(2,1) = .02$	s.d.	0.120	0.192	0.154	0.052	0.082	0.070	0.037	0.067	0.047	0.011	0.026	0.016
	RMSE	0.493	0.222	0.185	0.476	0.094	0.085	0.469	0.074	0.059	0.468	0.026	0.024
	Bias	0.078	0.054	0.020	0.069	0.002	0.003	0.072	0.002	0.005	0.072	0.001	0.004
$P_3(1,2) = .2$	s.d.	0.057	0.152	0.083	0.022	0.049	0.036	0.017	0.029	0.026	0.005	0.009	0.009
	RMSE	0.096	0.160	0.085	0.072	0.049	0.036	0.074	0.029	0.026	0.072	0.010	0.010
	Bias	0.352	0.062	0.089	0.359	0.044	0.046	0.363	0.025	0.040	0.363	0.006	0.021
$P_3(2,1) = .02$	s.d.	0.097	0.141	0.127	0.040	0.089	0.072	0.029	0.057	0.053	0.009	0.024	0.019
	RMSE	0.365	0.154	0.154	0.361	0.098	0.085	0.364	0.062	0.066	0.364	0.025	0.028

Table F1: Baseline Monte Carlo Simulation Results

	Carbon Stock
α	8.941***
	(0.009)
β	4.115***
	(0.039)
γ	0.630***
	(0.002)
δ	38.432***
	(0.042)
Observations	11,770,123
R^2	0.317

Note: *p<0.1; **p<0.05; ***p<0.01
Data as described in text. Regression uses data from 2017.

Table F2: Relationship Between Carbon Stock and Forest Age

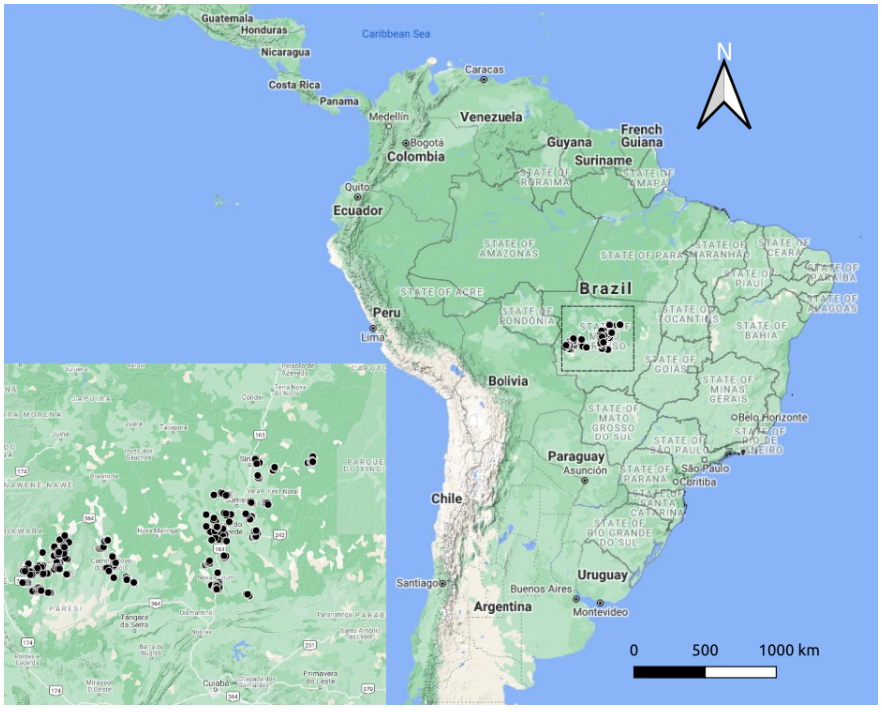
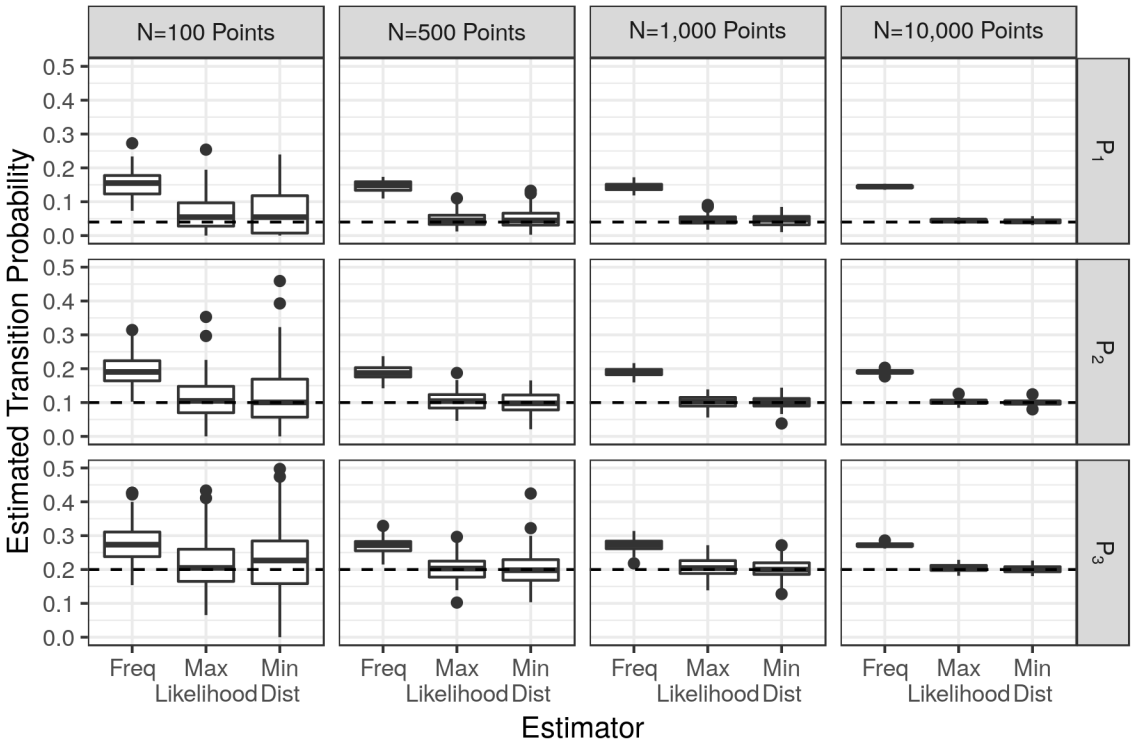
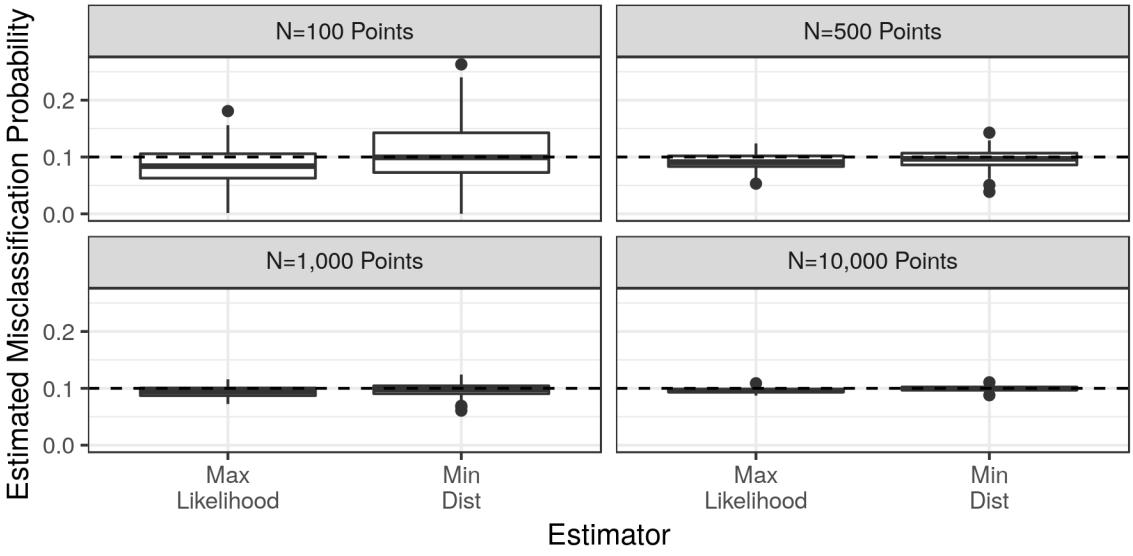


Figure F1: Map of Mato Grosso State and the Embrapa Sample Points

Figure F2: Map of Brazil and the Mapbiomas Sample Points, in the Brazilian Atlantic Forest

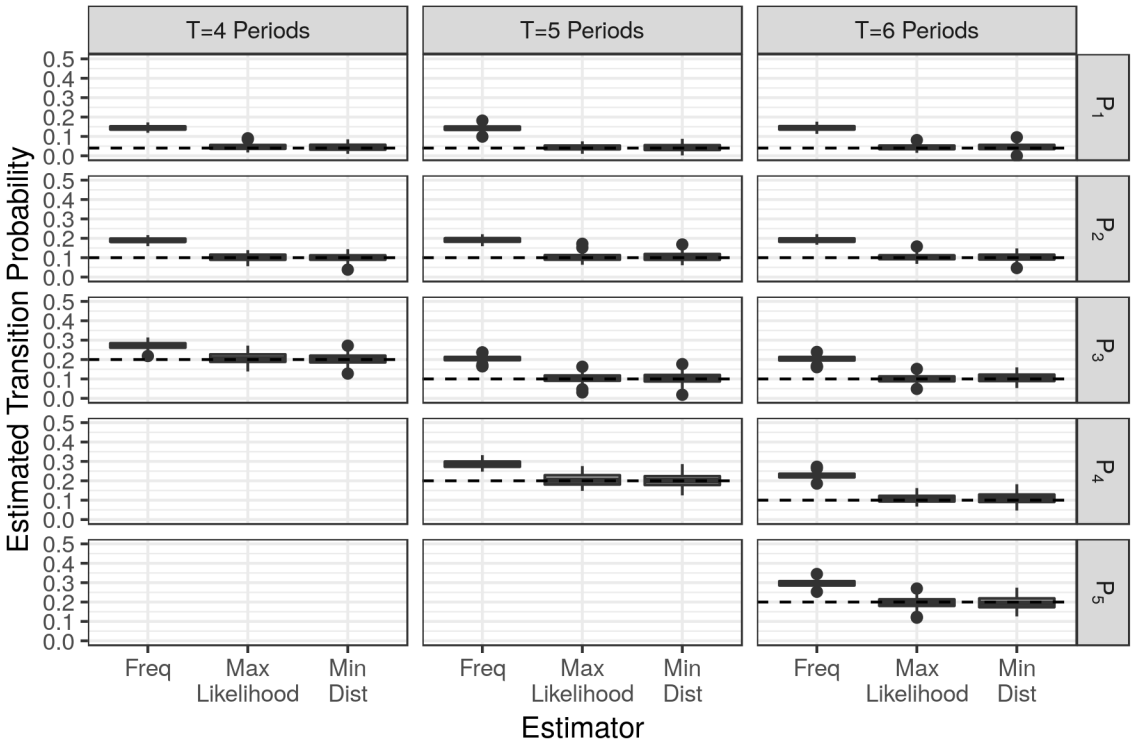


(a) Transition Probability, $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$

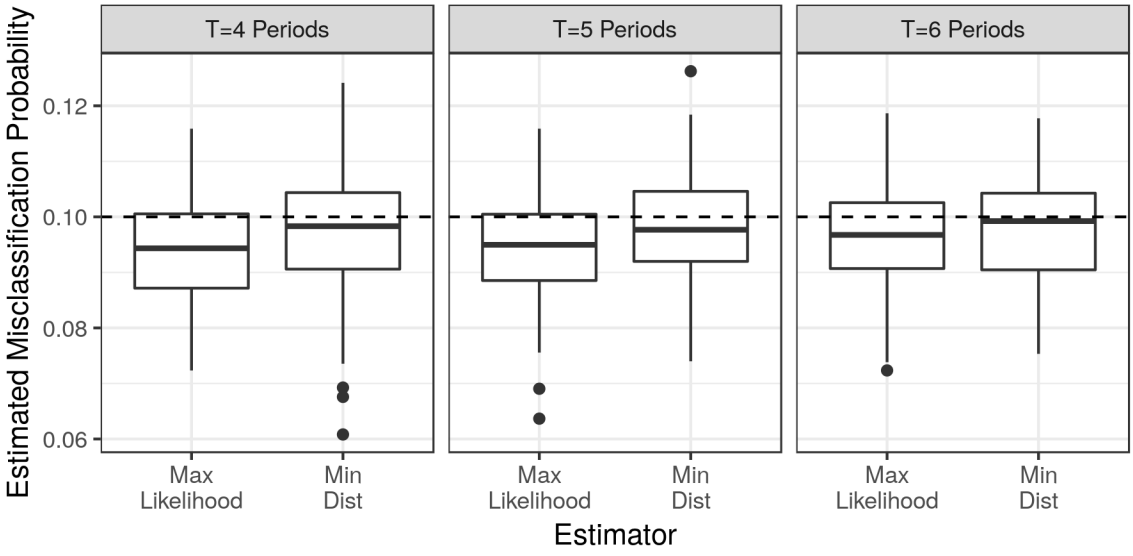


(b) Misclassification Probability, $\Pr[Y_{it} = 2|S_{it} = 1]$

Figure F3: Baseline Monte Carlo Simulation Results

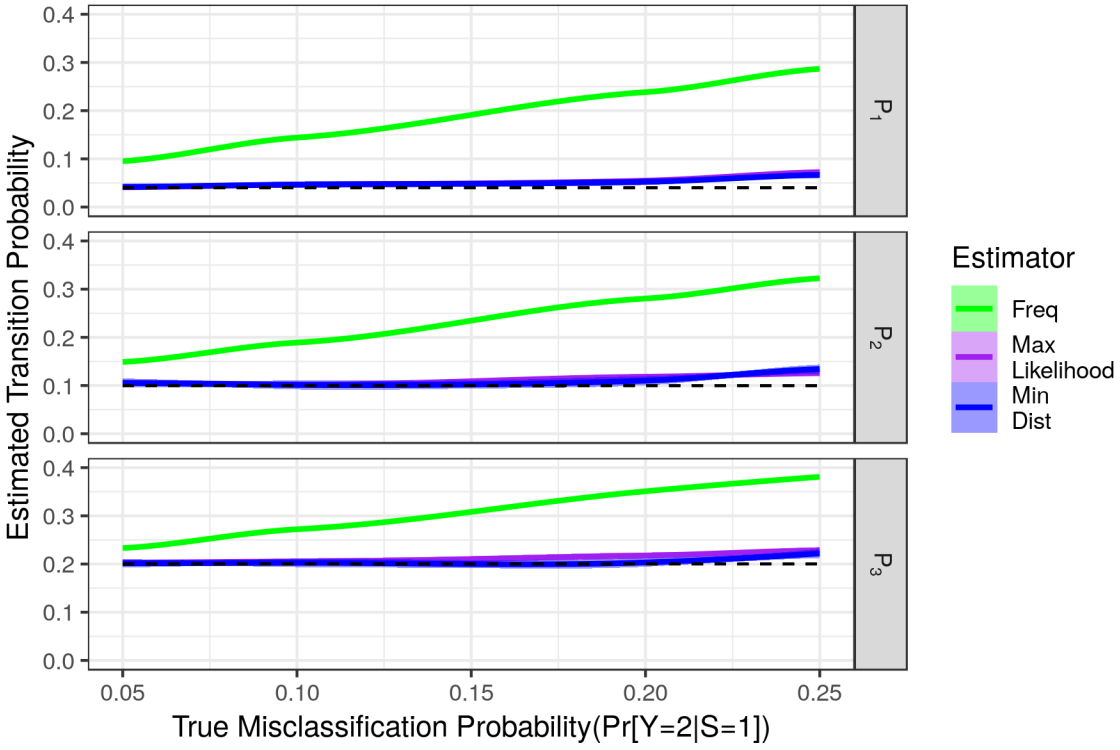


(a) Transition Probability, $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$

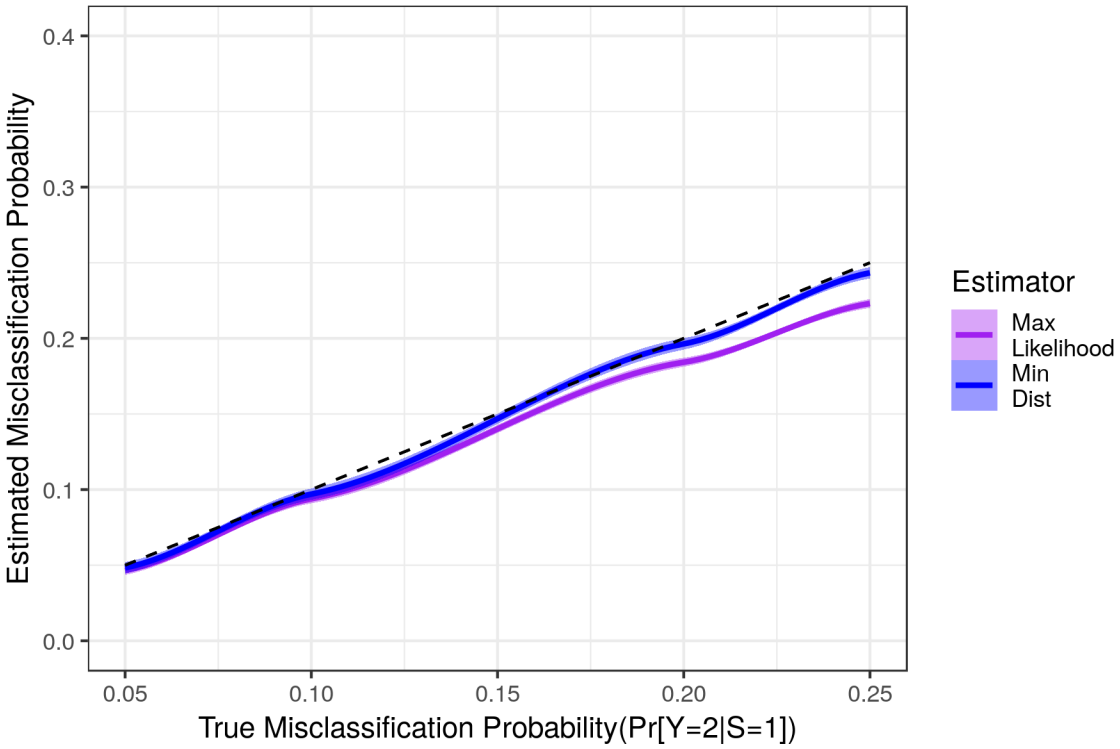


(b) Misclassification Probability, $\Pr[Y_{it} = 2|S_{it} = 1]$

Figure F4: Baseline Monte Carlo Simulation Results

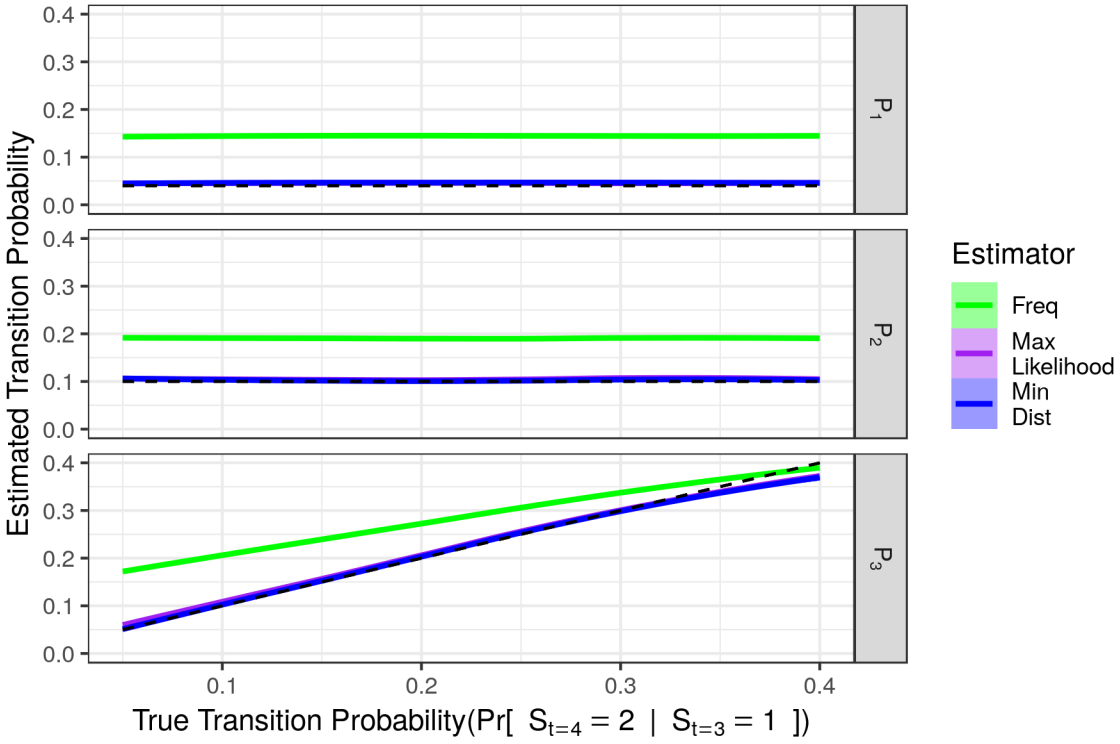


(a) Transition Probability

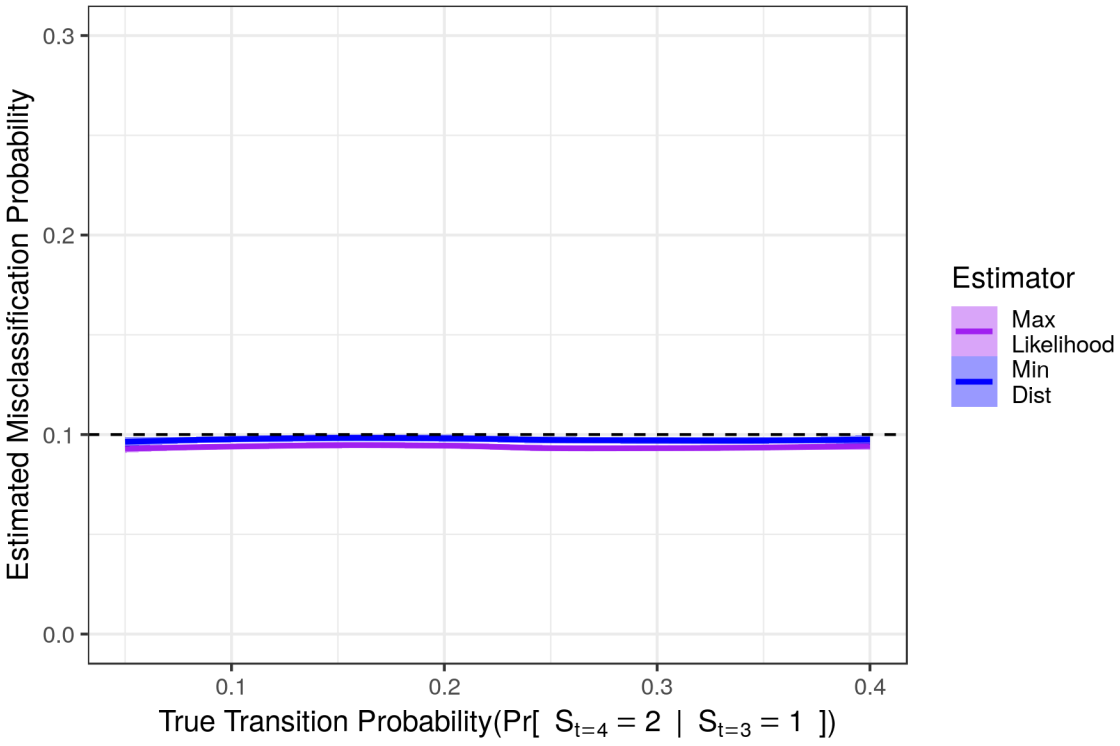


(b) Misclassification Probability

Figure F5: Monte Carlo Results for Varying Misclassification Probabilities

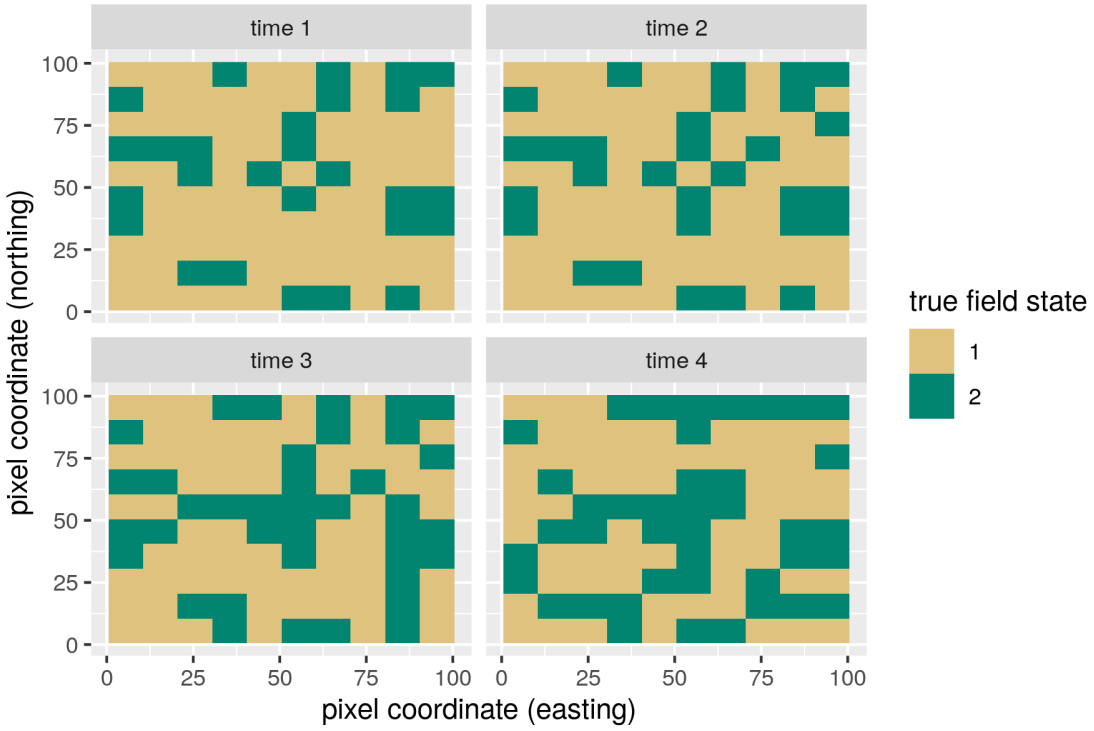


(a) Transition Probability

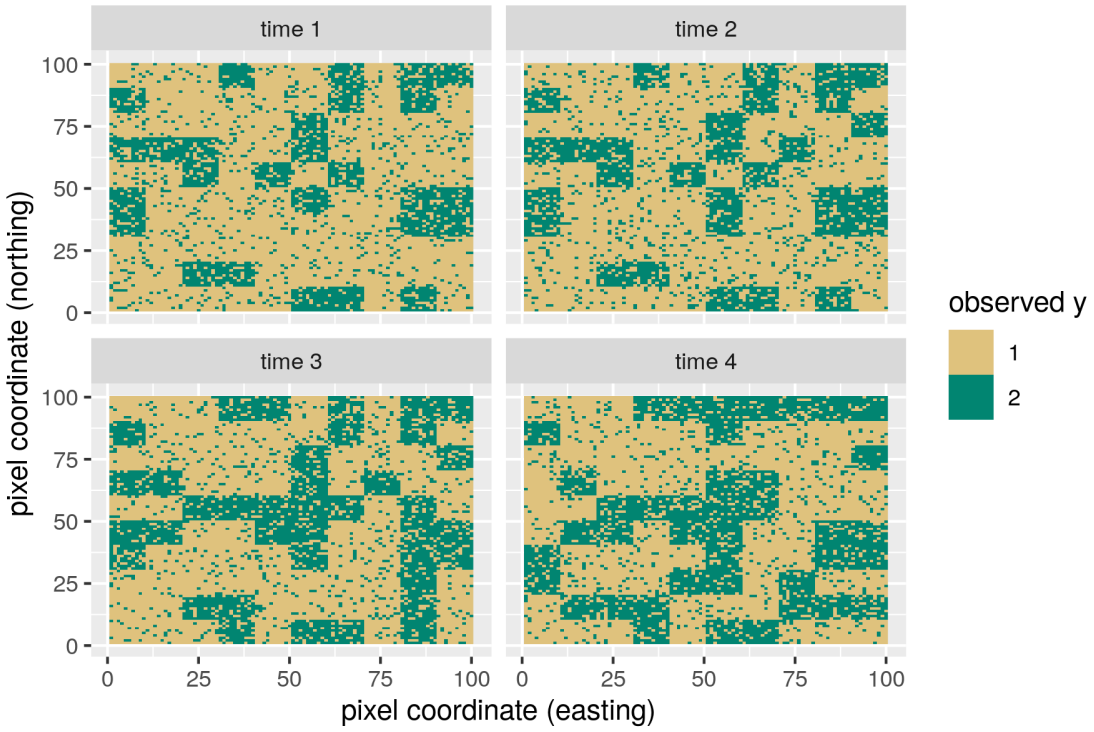


(b) Misclassification Probability

Figure F6: Monte Carlo Results for Varying Transition Probabilities

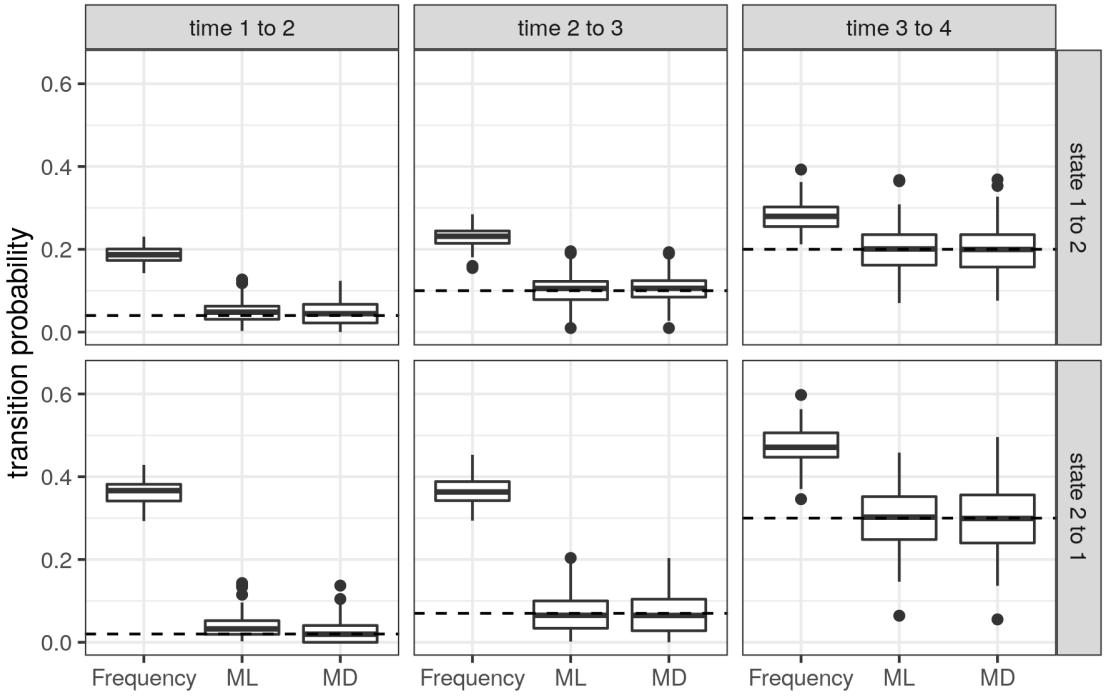


(a) Distribution of True Land Use, S_{it}

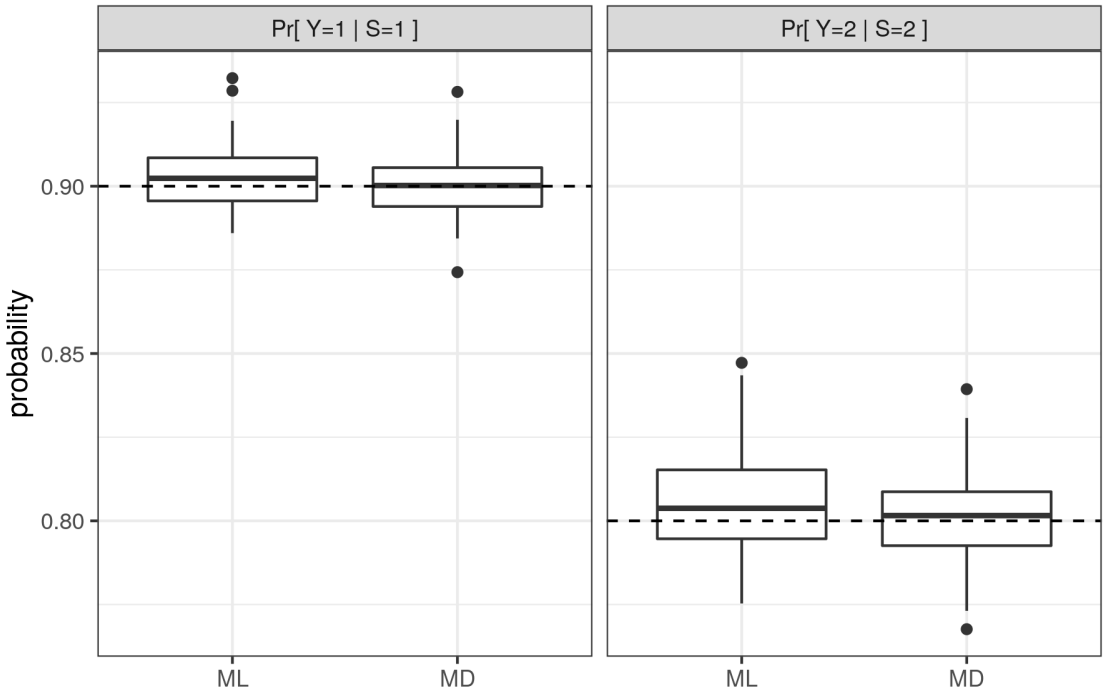


(b) Distribution of Observed Land Use, Y_{it}

Figure F7: Monte Carlo: Spatially Correlated Land Use, an Example

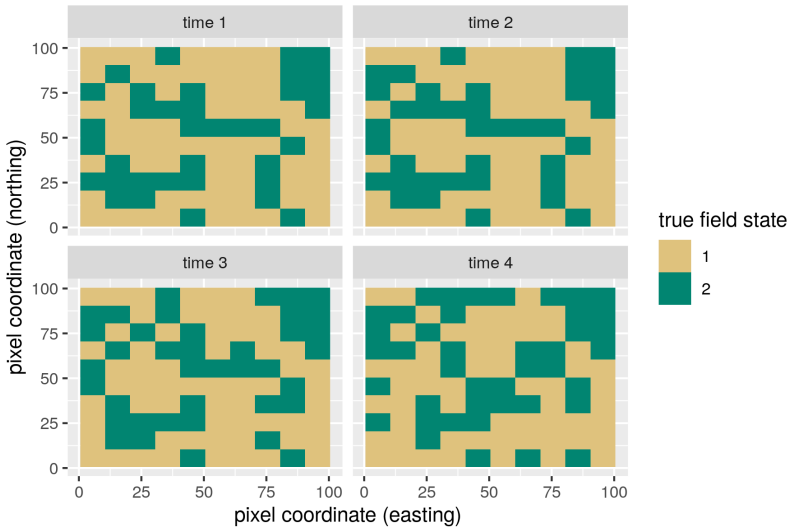


(a) Transition Probabilities

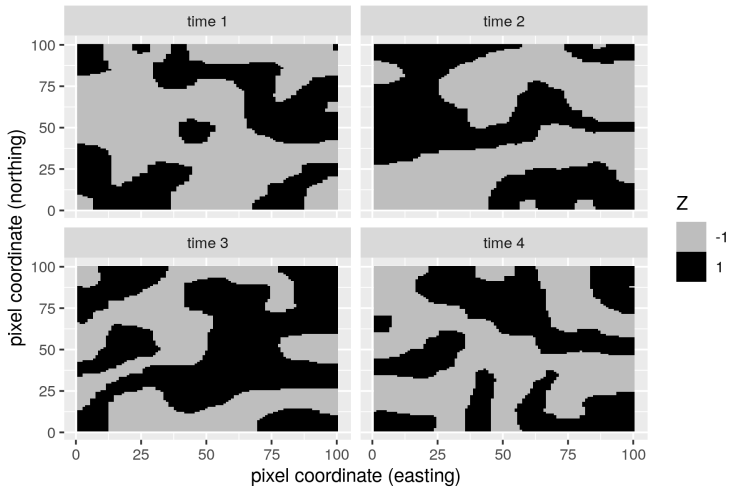


(b) Misclassification Probabilities

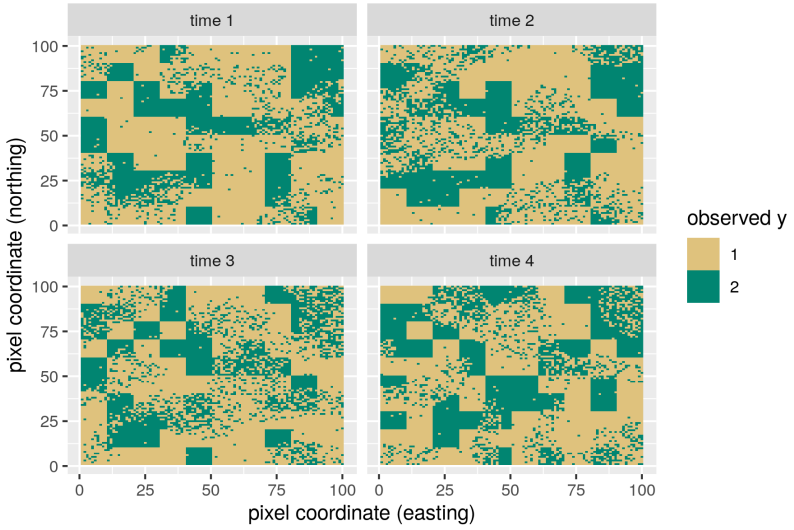
Figure F8: Monte Carlo: Spatially Correlated Land Use Results



(a) Distribution of True Land Use, S_{it}

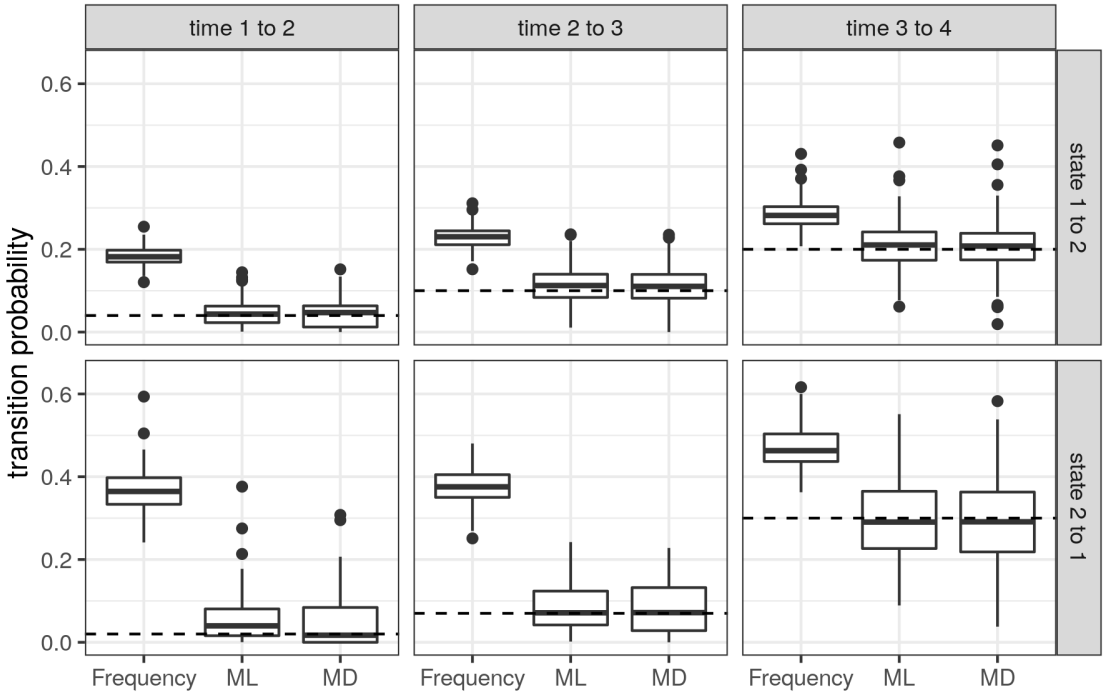


(b) Distribution of Z_{it}

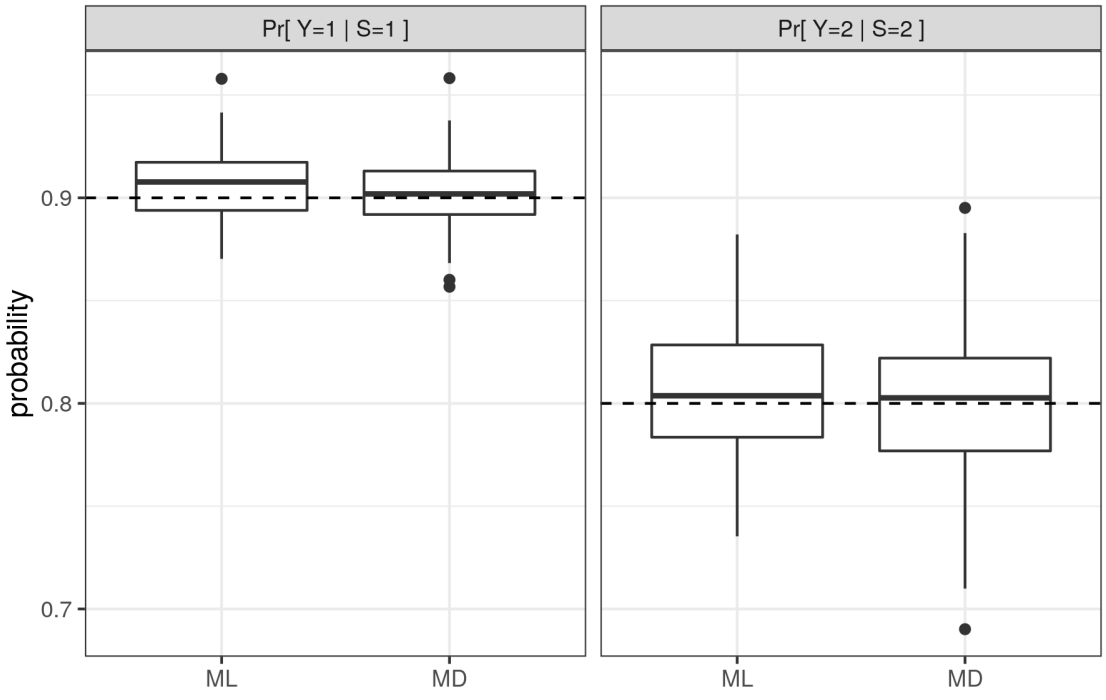


(c) Distribution of Observed Land Use, Y_{it}

Figure F9: Monte Carlo: Spatially Correlated Land Use and Misclassifications, an Example

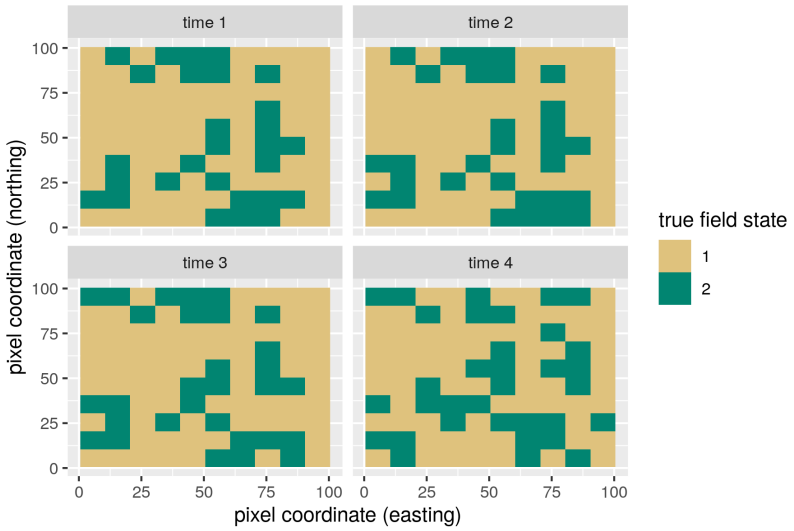


(a) Transition Probabilities

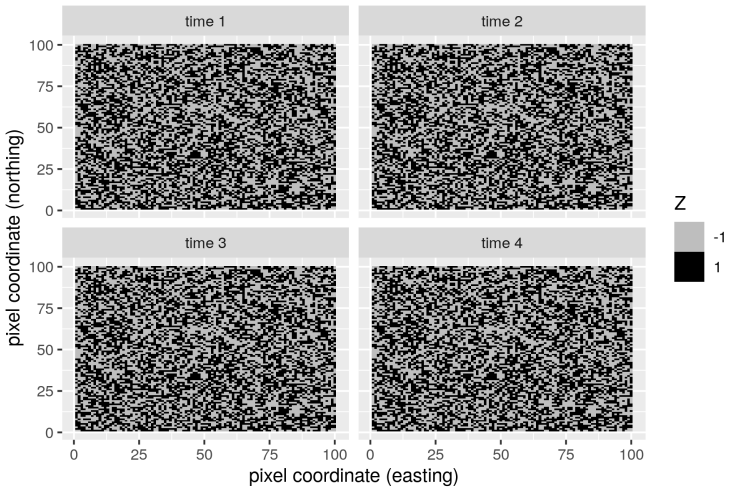


(b) Misclassification Probabilities

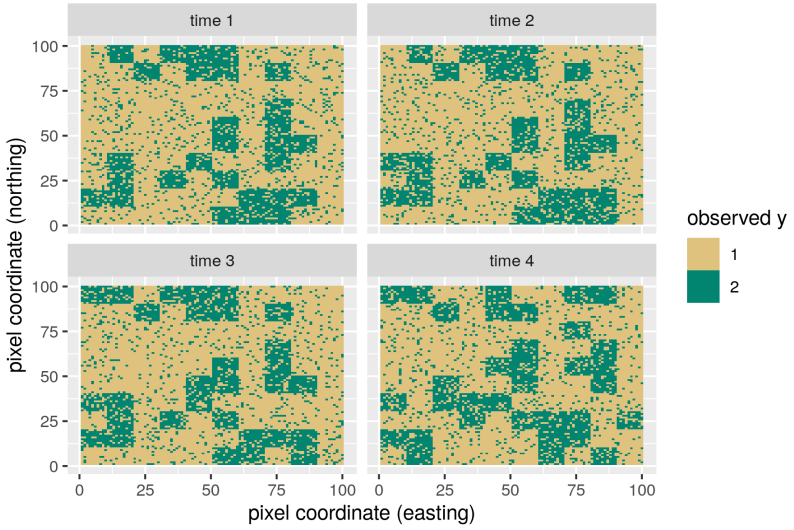
Figure F10: Monte Carlo: Spatially Correlated Land Use and Misclassifications Results



(a) Distribution of True Land Use, S_{it}

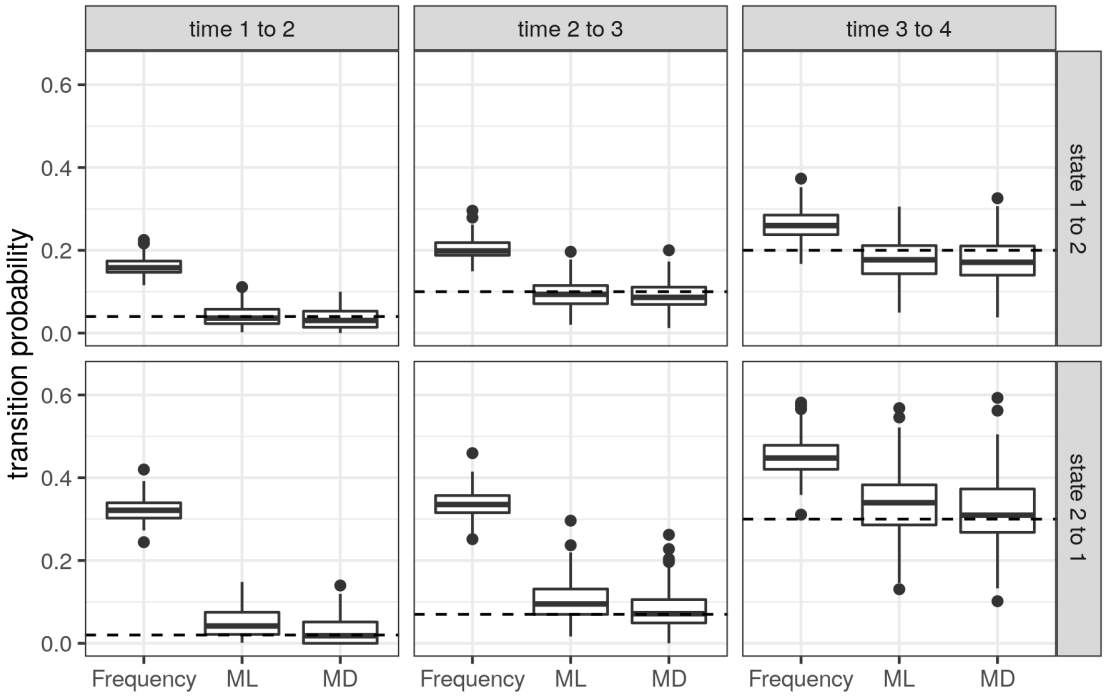


(b) Distribution of Z_{it}

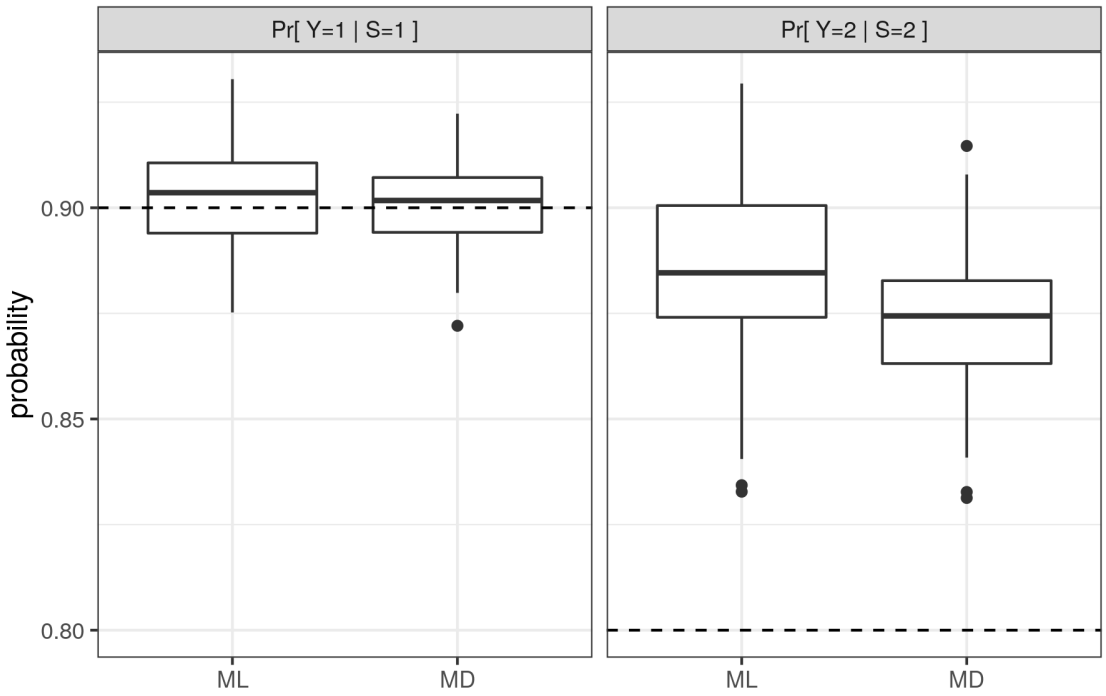


(c) Distribution of Observed Land Use, Y_{it}

Figure F11: Monte Carlo: Spatially Correlated Land Use and Serially Correlated Misclassifications, an Example

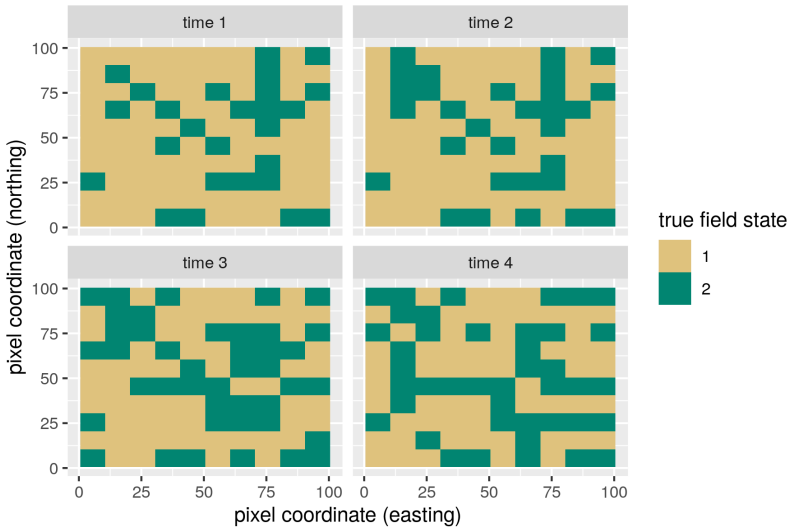


(a) Transition Probabilities

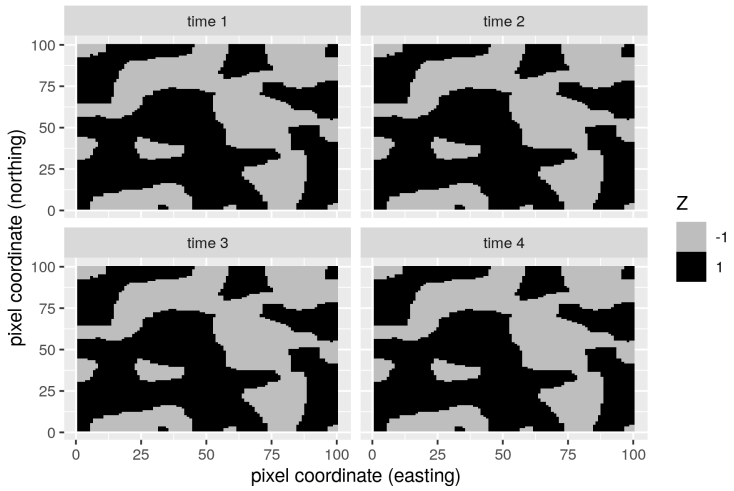


(b) Misclassification Probabilities

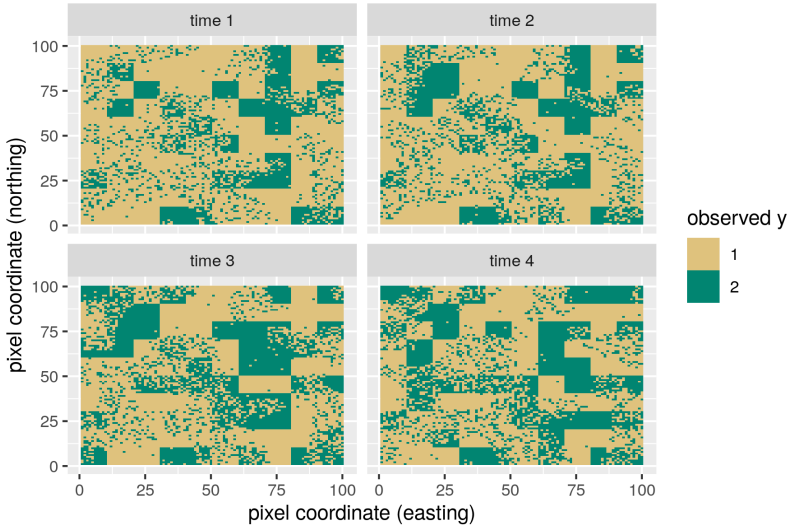
Figure F12: Monte Carlo: Spatially Correlated Land Use and Serially Correlated Misclassifications Results



(a) Distribution of True Land Use, S_{it}

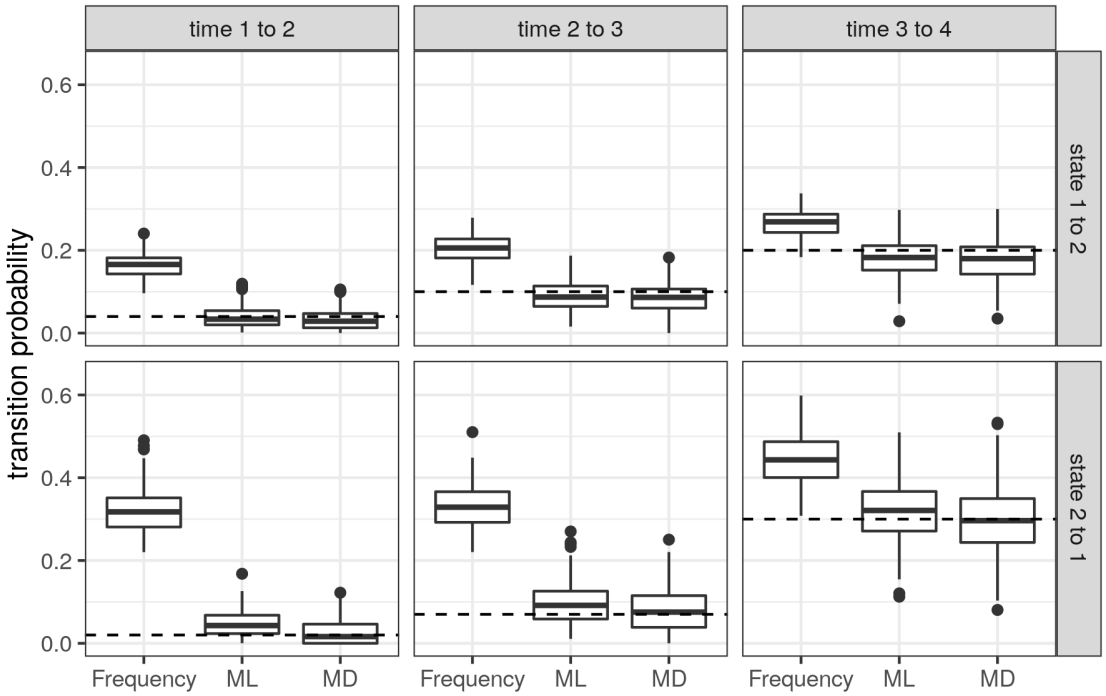


(b) Distribution of Z_{it}

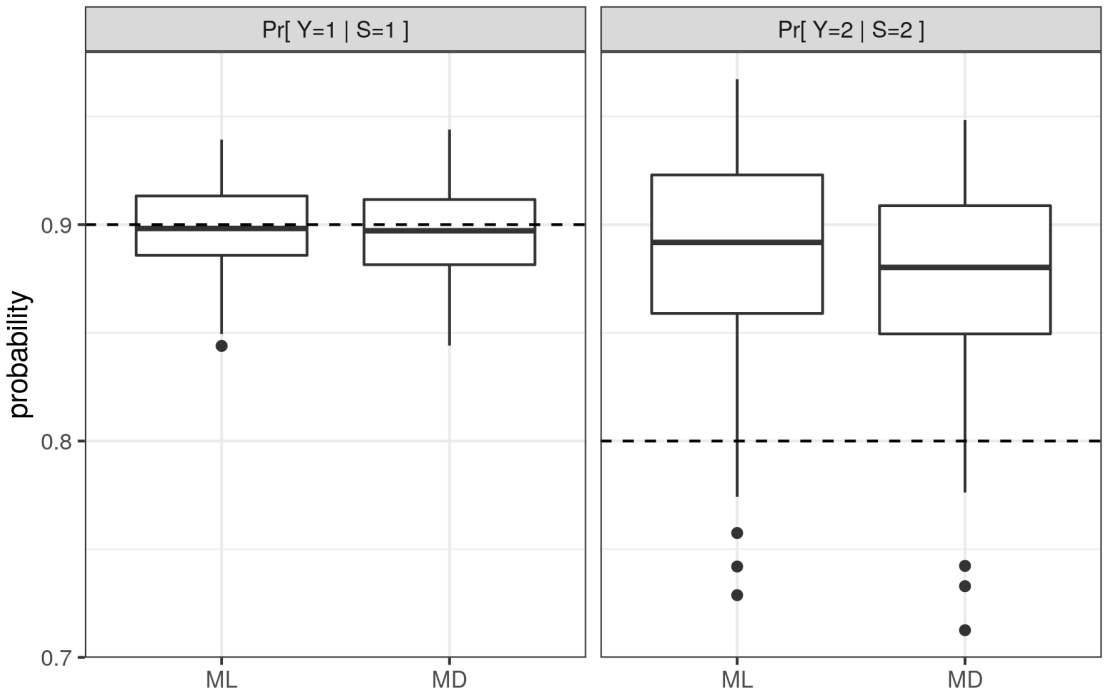


(c) Distribution of Observed Land Use, Y_{it}

Figure F13: Monte Carlo: Spatially Correlated Land Use, and Spatially and Serially Correlated Misclassifications, an Example



(a) Transition Probabilities



(b) Misclassification Probabilities

Figure F14: Monte Carlo: Spatially Correlated Land Use, and Spatially and Serially Correlated Misclassifications Results

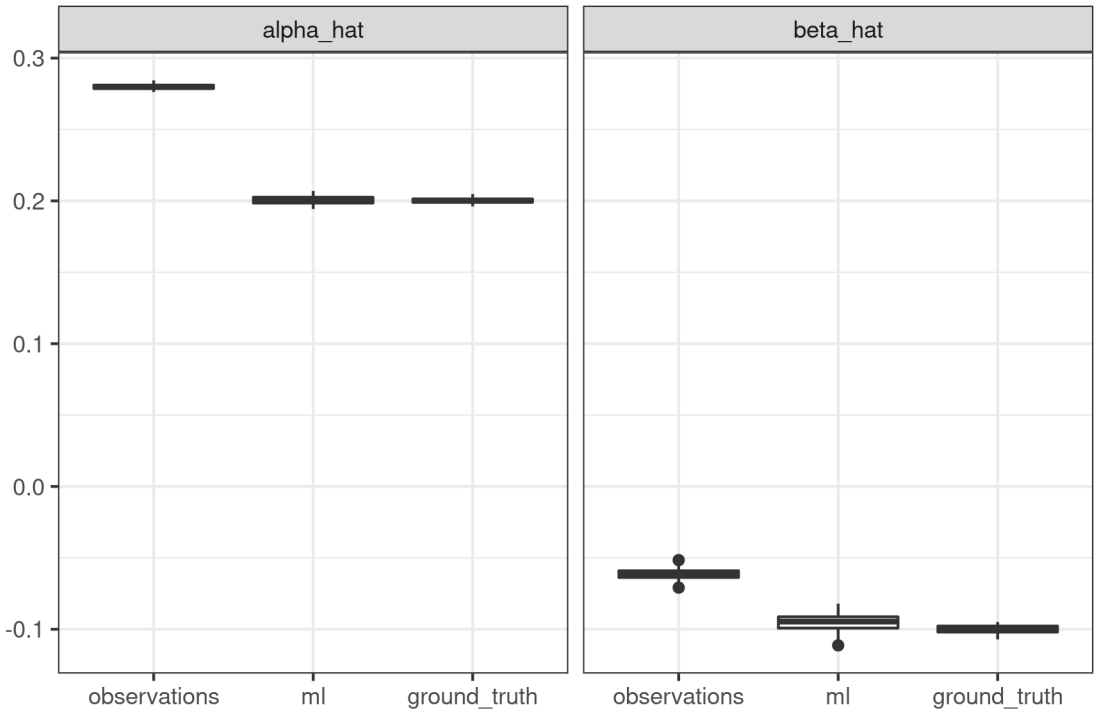


Figure F15: Regression Monte Carlo Results

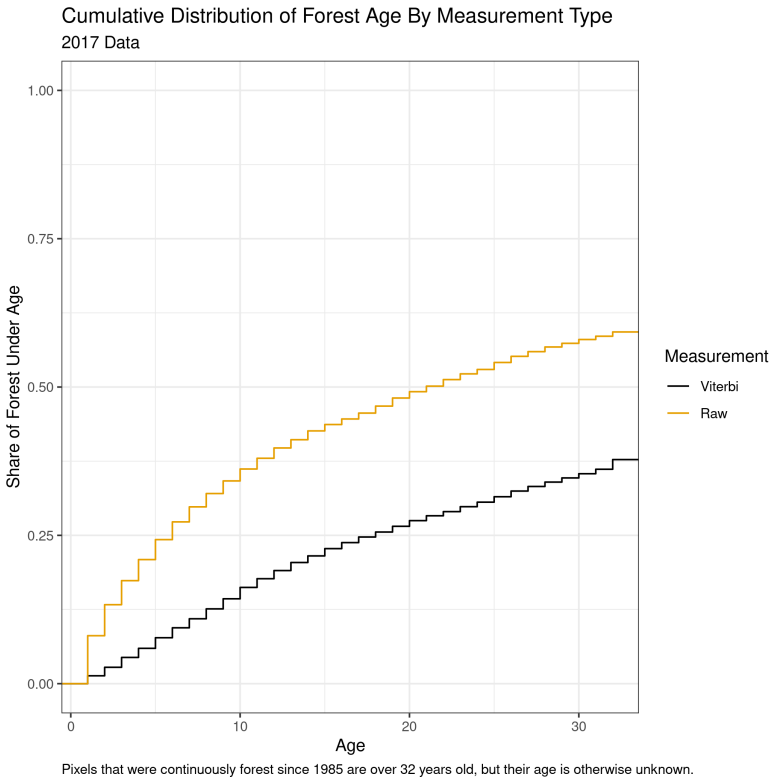


Figure F16: Forest Age Distribution

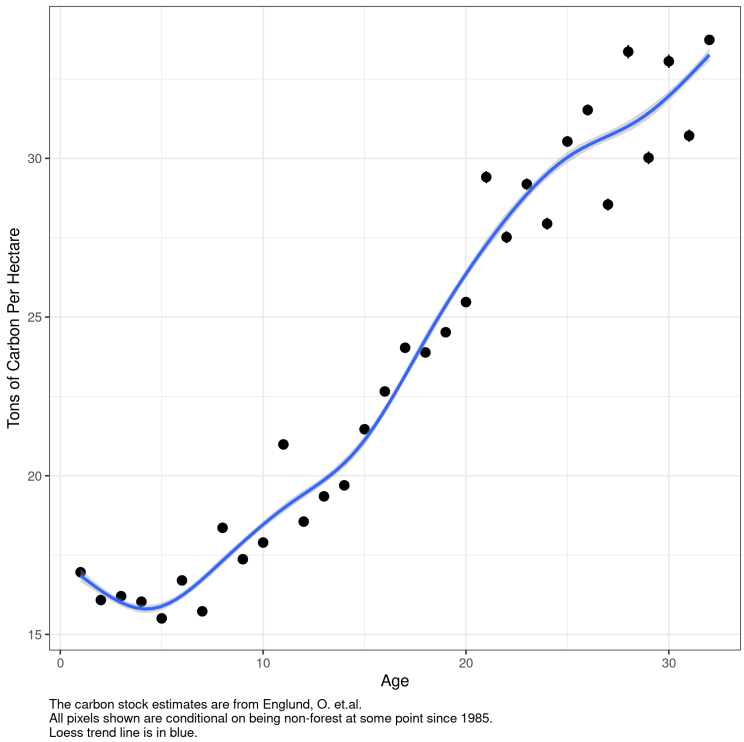


Figure F17: Carbon Stock by Age of Forest