University of Toronto
Department of Economics

Working Paper 672

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

By Eduardo Souza-Rodrigues, Adrian L. Torchiana, Ted Rosenbaum and Paul T. Scott

July 31, 2020

# 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 2020

## Abstract

Satellite-based image classification facilitates low-cost measurement of the Earth's surface composition. However, image classification techniques 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-level truth data is required to implement the correction. Our proposed correction produces consistent estimates of transition rates, confirmed by Monte Carlo simulations and panel validation data. In contrast, transition rates without correction for misclassification are severely biased.

**Keywords**: Measurement Error, Remote-Sensing Data, Land Cover, Hidden Markov Model

**JEL Codes**: C13, Q15, R14

[*] We are extremely grateful to Alexandre Camargo Coutinho and Daniel De Castro Victoria, who generously shared their gound-level land cover data with us. We also would like to thank Christian Abizaid, Anne Ruiz-Gazen, and Christine Thomas-Agnan for helpful discussions. We 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).

# 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] Compared to other data sources, remote sensing data have the advantages of providing a relatively low cost of accessing information that is difficult to obtain by other means, a high spatial resolution, and wide geographic and temporal coverage (Donaldson and Storeygard, 2016). Not surprisingly, they have been used widely – and increasingly – across a number of research fields, including economics, geography, biology, ecology, and political science, as well as being used 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; Lark et al., 2017; 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 misclassifications (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 (Assunção et al., 2019)) biases in transition rates can undermine efficient policy design and enforcement. Similarly, errors in satellite-based measures of pollution can lead to misleading information about pollution trends in different geographic areas and adversely affect air-quality regulations.[3]

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

[1] At the moment, there are approximately 5,000 satellites orbiting our planet, according to the Index of Objects Launched into Outer Space (http://www.unoosa.org/oosa/en/spaceobjectregister/index.html).

[2] Economists have been using satellite data to study crop choices and agricultural productivity (Holmes and Lee, 2009; Scott, 2013; Kudamatsu et al., 2016), climate change (Costinot et al., 2016), natural resources and deforestation (Burgess et al., 2012; Assunção et al., 2019), intensity and distribution of economic activities (Henderson et al., 2012; Nordhaus and Chen, 2014; Hu and Yao, 2019), urbanization and market boundaries (Goldblatt et al., 2018; Baragwanath et al., 2019), investments in housing innovation and poverty (Henderson et al., 2016; Marx et al., 2019), pollution and health outcomes (Fowlie et al., 2019), civil wars (Alix-Garcia et al., 2013), and the political economy of regional development (Gennaioli et al., 2012; Michalopoulos and Papaioannou, 2013), among many others. Geographers, biologists, and ecologists have also explored remote-sensing data to investigate land cover and degradation, terrestrial and marine ecosystems, biodiversity, and carbon emissions and carbon sequestration (Foley et al., 2005; Bonan, 2008; Geller et al., 2017).

[3] Fowlie et al. (2019) show that satellite-based measurement errors in pollution can lead to over-regulation in "clean" areas and under-regulation in "dirty" areas.

(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 hidden Markov model 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. 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 a 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 2) 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 land uses for each pixel in the data.

From the perspective of implementation, there are at least two attractive features to our approach. First, we do not require ground-truth data to implement the correction. Second, our 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 a Monte Carlo simulation study, and using rich longitudinal ground-level validation data. In the Monte Carlo study, we find that the HMM method estimates transition probabilities and misclassification probabilities accurately, including cases where the transition probabilities are time-varying. We also 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 and is less likely to result in estimates of transition probabilities that are at the edge of the unit interval $[0, 1]$.

We perform a validation exercise using a unique ground-level truth longitudinal data produced by the Brazilian Agricultural Research Corporation (Embrapa). The data describe land use in various private farmlands in the state of Mato Grosso, Brazil, from 2006 to 2010.[4] The state of Mato Grosso serves as an interesting setting 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 – factors that have attracted considerable attention from researchers and policy makers. This data is unprecedented in spatial and temporal coverage for the state – and arguably in general, given that longitudinal validation data are particularly difficult to obtain; typically, validation data are composed of a single or repeated cross-sections. Longitudinal ground-level data are crucial for validating our HMM approach, as they allow us to observe true land use transition rates and compare them to our estimates. The Embrapa data therefore provide a unique opportunity to test the performance of the HMM correction in practice.

In the validation exercise, we split the ground truth data into two disjoint sets: one which is used to train a land cover classifier; and another which is used to compare the estimated transition probabilities and the ground-level land use changes. The classifier's predictors are based on remote sensing data from MODIS, which are regularly used to study land cover (see, e.g., the discussion in Brown et al., 2013). After training the classifier, we implement the HMM corrections – the MD and ML estimators – and validate them against the ground truth data. We find that the HMM-based corrections estimate transition rates accurately, while transition rates computed without correction for misclassifications 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. Further, as a by-product, the HMM approach provides precise estimates of the land use misclassification probabilities (i.e. the classification errors of the uncorrected classifier).

To the best of our knowledge, the closest paper to ours is by Abercrombie and Friedl (2016), who also consider an HMM-based correction to errors in land use 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, and we propose consistent estimators for transitions and misclassification probabilities (together with corresponding sampling uncertainty measurements). We also allow for

[4] We use "land use" and "land cover" interchangeably in this paper. For details about the data set, see Coutinho et al. (2011) and Brown et al. (2013). While Brown et al. (2013) use that data to obtain substantial progress towards more refined crop-specific classification, we focus on land use transition estimates.

the estimation of time-varying transition probabilities from the data. Allowing for time-varying transition probabilities is crucial in many applications, e.g. when estimating how (and explaining why) deforestation processes may change over time.

This paper is organized as follows: Section 2 lays out the framework and the misclassification problem; Section 3 discusses the use of validation data and heuristic ad hoc solutions to correct misclassifications, and presents the hidden Markov model, followed by the HMM formal identification results; Section 4 describes the two estimation methods, the minimum distance estimator and the maximum likelihood estimator; Section 5 presents the Monte Carlo simulation studies; Section 6 describes the validation exercise using ground-level truth data from Brazil's Embrapa; and Section 7 concludes.[5]

## 2 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.

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. In our Embrapa validation exercise we use a binary state space $\mathcal{S} = \{\text{crops}, \text{pasture}\}$. Extensions to continuously distributed measurements, such as pollution or nighttime light, are possible, at the cost of more burdensome notation and additional technical details.[6] The true land use $S_{it}$ is not observed unless ground-level 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 keep the distinction in the notation to facilitate the presentation. 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 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 $t$. In other words, we can take $Y_{it} = f(R_{it})$, for some function $f$ that depends on the data used and the classification

[5] The Appendix presents relevant mathematical derivations for the HMM correction and the details of the EM algorithm. Code for replicating the Monte Carlo simulations in R is available at https://github.com/atorch/hidden_markov_model.

[6] For variables taking value on the real line, one needs to work in Hilbert spaces, with their corresponding operators, instead of in Euclidean spaces with transformation matrices, as we do here.

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 $\Pr[Y_{it}]$, 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.[7]

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 $y \neq s$. Errors in classifications may be the combined result of the specific sensor characteristics of the satellite, the angle of the satellite with respect to the sun and the earth's surface, and the atmospheric conditions, including cloud cover and haze (Lillesand et al., 2004).

In matrix notation, the equation above becomes

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

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 $\boldsymbol{\Upsilon}$ 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 $\boldsymbol{\Upsilon}$ 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 $\boldsymbol{\Upsilon}$ is time-invariant, but the results can be extended to misclassifications that may change over time.

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

[7] 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.

guarantee that the observed (estimated) transition $\Pr[Y_{it+1}|Y_{it}]$ is close to the true transitions $\Pr[S_{it+1}|S_{it}]$.

## 3 Correcting Satellite-Based Misclassifications

We now consider three possible corrections for misclassified remote-sensing data. The first correction is based on available cross-sectional validation data. The second is based on heuristic ad hoc corrections of implausible land use changes in the data. The third is based on our hidden Markov model approach.

### 3.1 Correction Based on Cross-Sectional Validation Data

Suppose we have access to a subset of (repeated cross-sectional) validation points for which we observe both $Y_{it}$ and $S_{it}$, i.e., $\{Y_{it}, S_{it} : i = 1, ..., N_s; t = 1, ..., T_s\}$. Given such data, we can directly estimate $\boldsymbol{\Upsilon}$. In fact, when fitting a machine learning model such as a land cover classifier, it is standard practice to split the sample into training and test sets. The training set is used to estimate the model, and predictions are then extrapolated to the test data and compared to the ground truth. The comparison produces the so-called "confusion matrix," which is a matrix with the joint distribution $\Pr[Y_{it}, S_{it}]$. The matrix $\boldsymbol{\Upsilon}$ of conditional probabilities $\Pr[Y_{it} \mid S_{it}]$ can be obtained directly from the confusion matrix. We can then recover the true land use shares provided the matrix $\boldsymbol{\Upsilon}$ is invertible:

$$\mathbf{P}_{S_t} = \boldsymbol{\Upsilon}^{-1}\mathbf{P}_{Y_t}. \tag{2}$$

In practice, we expect $\boldsymbol{\Upsilon}$ to be invertible because land use classifications are often sufficiently accurate to make the matrix diagonally dominant: $\Pr[Y_{it} = s|S_{it} = s]$ is typically much larger than 50 percent, and hence larger than the probability of misclassification $\sum_{y\neq s}\Pr[Y_{it} = y|S_{it} = s]$.

To verify whether the correction works in a particular case, we may split the test dataset in two. One subsample can be used to estimate $\boldsymbol{\Upsilon}$, while the other is used to check whether the corrected land use classifications are indeed better than the original classification (see Czaplewski, 1992). As discussed in the Introduction, a limitation of this approach is the potential lack of access to a sufficiently rich validation data set, particularly if our goal is to estimate models that vary spatially, e.g. by region. Moreover, if our goal is to estimate transition probabilities (rather than year-by-year land use shares), we need an approach that allows us to estimate more than just the marginal probabilities $\Pr[S_{it}]$.[8]

[8] To see why marginal probabilities are not enough to obtain transitions, take a binary case $\mathcal{S} = \{s_1, s_2\}$. By the

## 3.2 Heuristic Ad Hoc Corrections

Even in the absence of ground-level validation data, some errors in land cover classifications are readily apparent. A potential strategy is to come up with ad hoc rules to correct plausible errors. For instance, suppose $\mathcal{S} = \{\text{forest}, \text{non-forest}\}$. If the predictions for a particular point $i$ were $Y_{it} = \text{forest}$ for each of the first ten years of data, followed by non-forest for the eleventh year, followed by forest for four more years, then it is reasonable to assume that this point was covered in forest throughout the sample period and that the eleventh classification was almost certainly an error. This is conceivable given that it takes 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). By implementing such type of reclassifications, one can effectively smooth out implausible transitions in the data.[9]

While heuristic-based adjustments 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 (Friedl et al., 2010).

## 3.3 Correction Based on Hidden Markov Model Approach

We now turn to our proposed solution, which is based on Hu (2017). For each point $i$, we assume the stochastic process $\{Y_{it}, S_{it} : t = 1, 2, ...\}$ follows a hidden Markov process. Specifically, we assume 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 $\{Y_{it-j}, S_{it-j}\}$, $j \geq 0$, conditional on $S_{it+1}$. The conditional independence assumption means that if we know the true land use, 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).[10]

Law of Total Probability,

$$\Pr[S_{it+1} = s_1] = \Pr[S_{it+1} = s_1|S_{it} = s_1]\Pr[S_{it} = s_1] + \Pr[S_{it+1} = s_1|S_{it} = s_2]\Pr[S_{it} = s_2].$$

Even when the marginals $\Pr[S_{it+1}]$ and $\Pr[S_{it}]$ are known, there is only one equation to identify two unknowns $\Pr[S_{it+1} = s_1|S_{it} = s_1]$ and $\Pr[S_{it+1} = s_1|S_{it} = s_2]$.

[9] Ad hoc corrections like this are made by Mapbiomas, a publicly available panel database on land cover in Brazil (https://mapbiomas.org/). Researchers using NASS's Cropland Data Layer, which is an annual longitudinal data of land cover predictions for the contiguous United States, often make use of such types of ad hoc adjustments (see, e.g., Lark et al., 2015)

[10] Note that the observed process $\{Y_{it}\}$ does *not* necessarily follow a first-order Markov process. We also assume the processes are spatially independent: $\{Y_{it}, S_{it}\} \perp\!\!\!\perp \{Y_{jt}, S_{jt}\}$ for $i \neq j$ (where $\perp\!\!\!\perp$ indicates probabilistic independence); extending the model to allow for both spatial and time dependence is possible, but adds considerable complexity.

Formally,

$$\Pr[Y_{it+1}, S_{it+1} | \{Y_{it-j}, S_{it-j}\}_{j\geq 0}] = \Pr[Y_{it+1}|S_{it+1}] \times \Pr[S_{it+1}|S_{it}]. \tag{3}$$

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 HMM assumptions, then so must $\{f(R_{it}), S_{it}\}$ for any function $f$.[11] The stochastic process is summarized graphically in Figure 1, in which nodes represent random variables and edges indicate statistical dependence.

**Useful Identities.** Given the HMM assumptions, there are a series of identities that are helpful to obtain the identification results. For any two random variables $X, W \in \mathcal{S}$, define the $K \times K$ matrix $\mathbf{M}_{X,W}$ with the joint distribution $\Pr[X = s_l, W = s_k]$, $l, k = 1, ..., K$. 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 $\Pr[Y_{it+1} = y_{t+1}|X = s_k]$.[12]

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

$$\mathbf{M}_{Y_t,Y_{t-1}} = \mathbf{\Upsilon}\,\mathbf{M}_{S_t,Y_{t-1}}. \tag{4}$$

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

$$\mathbf{M}_{Y_{t+1},Y_t} = \mathbf{\Upsilon}\,\mathbf{M}_{S_{t+1},S_t}\,\mathbf{\Upsilon}^\intercal, \tag{5}$$

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 $Y_{it+1} = y_{t+1} \in \mathcal{Y}$,

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

Identification and estimation of the HMM is based on (4)–(6). See Appendix A for a derivation of

We therefore focus on temporal dependence within each spatial point and leave extensions incorporating spatial dependence for future research. The assumption of spatial independence is not uncommon in the literature (see, e.g., Abercrombie and Friedl, 2016).

[11] 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.

[12] 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 $Y_{it+1}$.

these equations.

### 3.3.1 Identification of the Hidden Markov Model

In this subsection, we outline the conditions needed to identify the Markov land use 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 $Y_{it}$. Since we use the identification results of Hu (2017), our assumptions parallel his.[13]

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.*

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

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.,* $\operatorname{rank}\left(\mathbf{M}_{Y_t,Y_{t-1}}\right) = K$.

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, which results in a full rank matrix.[14] This condition implies that both matrices $\boldsymbol{\Upsilon}$ and $\mathbf{M}_{S_t,Y_{t-1}}$ are invertible too; see equation (4).

Combining (4) and (6), we get

$$\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}\mathbf{M}^{-1}_{Y_t,Y_{t-1}} = \boldsymbol{\Upsilon}\,\mathbf{D}_{y_{t+1}|S_t}\,\boldsymbol{\Upsilon}^{-1}. \tag{7}$$

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 $\boldsymbol{\Upsilon}$ are the eigenvectors. Because each column of $\boldsymbol{\Upsilon}$ 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 $\boldsymbol{\Upsilon}$ and the diagonal matrix $\mathbf{D}_{y_{t+1}|S_t}$ from the joint distribution of the observed classifications $(Y_{it+1}, Y_{it}, Y_{it-1})$.

**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 $s, s' \in \mathcal{S}$.*

[13] These results are based on Section 2 of Hu (2017). See also Hu (2020).

[14] This is easy to see in the 2 × 2 example:

$$\mathbf{M}_{Y_t,Y_{t-1}} = \begin{bmatrix} \Pr[Y_{it} = y_1, Y_{it-1} = y_1] & \Pr[Y_{it} = y_1, Y_{it-1} = y_2] \\ \Pr[Y_{it} = y_2, Y_{it-1} = y_1] & \Pr[Y_{it} = y_2, Y_{it-1} = y_2] \end{bmatrix}.$$

Note that for (observed) persistent processes, the diagonal terms are greater than the off-diagonal terms.

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.[15]

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. $Y_{it+1} = \text{forest}$) 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,

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

for $y = forest$. Condition 4 is then satisfied.[16]

Next we turn to the eigenvectors:

**Condition 5.** $\Pr\left[Y_{it} = s^*|S_{it} = s^*\right] > \Pr\left[Y_{it} = s|S_{it} = s^*\right]$ *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 $\Pr\left[Y_{it}|S_{it} = s^*\right]$. In words, given that the true land use is $s^*$, the probability that the noisy measure equals $s^*$ is greater than the probability that $Y_{it}$ equals any other land use $s \neq s^*$. This condition is satisfied when $\boldsymbol{\Upsilon}$ is strictly diagonally dominant; as previously mentioned, accurate land use classifiers generate $\boldsymbol{\Upsilon}$ that is diagonally dominant in practice.[17]

Next, given identification of the misclassification probabilities $\boldsymbol{\Upsilon}$ from the eigenvalue-eigenvector decomposition (7), we identify the joint distribution $\Pr\left[S_{it+1}, S_{it}\right]$ under the assumption that $\boldsymbol{\Upsilon}$ is time-invariant. Specifically, we impose

**Condition 6.** $\Pr\left[Y_{it+1}|S_{it+1}\right] = \Pr\left[Y_{it}|S_{it}\right]$.

Given Condition 6 and equation (5), we obtain

$$\mathbf{M}_{S_{t+1},S_t} = \boldsymbol{\Upsilon}^{-1}\mathbf{M}_{Y_{t+1},Y_t}\left(\boldsymbol{\Upsilon}^{\intercal}\right)^{-1}, \tag{8}$$

which implies identification of $\mathbf{M}_{S_{t+1},S_t}$, and hence of both $\Pr\left[S_{it+1}|S_{it}\right]$ and $\Pr\left[S_{it}\right]$.

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

[16] In case the condition 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. On the other hand, when Condition 4 holds for more than one value $y$, the model becomes overidentified.

[17] This corresponds to Assumption 4.2 in Hu (2017). Alternatively, one could assume the misclassification probabilities $\Pr\left[Y_{it} = y|S_{it} = s\right]$ 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.

**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* $\Pr[Y_{it}|S_{it}]$, $\Pr[S_{it}]$, *and* $\Pr[S_{it+1}|S_{it}]$.

For completeness, note that when Condition 6 is not satisfied, we need an additional time period to identify misclassification probabilities at both $t$ and $t+1$; i.e. we need at least $T \geq 4$ periods of data.[18]

We close this section with an important remark.

*Remark* 1. (Viterbi Algorithm.) An HMM model is not a classifier *per se*, but it can be used to generate land use predictions using the Viterbi algorithm (van Handel, 2008, Chapter 3). Viterbi 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 $\Pr[S_1, S_2, \ldots, S_T|Y_1, Y_2, \ldots, Y_T]$ for any given pixel.[19]

Compared to heuristic adjustments, which aim to smooth out implausible transitions based on ad hoc rules, the Viterbi algorithm together with the HMM method smooths out observations using misclassification and transition probabilities. The strategy effectively makes corrections to obvious errors like the one described in Section 3.2 while allowing for the possibility that the data also contain less obvious errors. Furthermore, the strategy does not require researchers to make decisions about what sorts of transitions are implausible – the errors can be recovered from the data itself – although such restrictions can be implemented as restrictions on the estimated transition rates if desired.

## 4 Estimators for the HMM Correction

In this section, we consider two estimators for the HMM correction: a minimum distance (MD) estimator and a maximum likelihood (ML) estimator.

[18] When both transition probabilities and misclassification probabilities vary by time, equation (7) becomes $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}} \mathbf{M}^{-1}_{Y_t,Y_{t-1}} = \boldsymbol{\Upsilon}_t \, \mathbf{D}_{y_{t+1}|S_t} \, \boldsymbol{\Upsilon}_t^{-1}$, and equation (8) becomes $\mathbf{M}_{S_{t+1},S_t} = \boldsymbol{\Upsilon}^{-1}_{t+1} \mathbf{M}_{Y_{t+1},Y_t} \, (\boldsymbol{\Upsilon}_t^{\intercal})^{-1}$. So, we need one extra time period to identify $\mathbf{M}_{S_{t+1},S_t}$. Clearly, $\boldsymbol{\Upsilon}_1$ and $\boldsymbol{\Upsilon}_T$ are not identified, and neither are the first and last transition matrices.

[19] The probability path estimate $\Pr[S_1, S_2, \ldots, S_T|Y_1, Y_2, \ldots, Y_T]$ 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). Note that finding the most likely path is different from the problem of finding the most likely land use in a given period $\Pr[S_t|Y_1, Y_2, \ldots, Y_T]$.

## 4.1 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 (7)–(8). 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 (Hu, 2017).

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 $y \in \mathcal{Y}$,

$$\begin{aligned} g_{1yt}(\mathbf{M}, \boldsymbol{\Upsilon}) &= \left\| \mathbf{M}_{y_{t+2},Y_{t+1},Y_t} \mathbf{M}_{Y_{t+1},Y_t}^{-1} \boldsymbol{\Upsilon} - \boldsymbol{\Upsilon} \mathbf{D}_{y_{t+2}|S_{t+1}} \right\|, \\ g_{2t}(\mathbf{M}, \boldsymbol{\Upsilon}) &= \left\| \mathbf{M}_{Y_{t+1},Y_t} - \boldsymbol{\Upsilon} \mathbf{M}_t \boldsymbol{\Upsilon}^{\intercal} \right\|, \end{aligned} \tag{9}$$

where $\|\cdot\|$ is a matrix norm. Notice that $g_{1yt}$ is analogous to equation (7) with slight rearrangement, while $g_{2t}$ is analogous to equation (8). So, under the true misclassification probabilities and joint distributions, $\mathbf{M}$ and $\boldsymbol{\Upsilon}$, we have that $g_{1yt} = g_{2t} = 0$. (We omit $\mathbf{D}_{y_{t+2}|S_{t+1}}$as an argument to $g_{1yt}$ because it is a function of $\boldsymbol{\Upsilon}$ 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 $Q(\mathbf{M}, \boldsymbol{\Upsilon}) = g(\mathbf{M}, \boldsymbol{\Upsilon})^{\intercal} \mathbf{W} g(\mathbf{M}, \boldsymbol{\Upsilon})$, where $\mathbf{W}$ is a symmetric positive-definite weighting matrix. By construction, $Q(\mathbf{M}, \boldsymbol{\Upsilon}) \geq 0$, and the true matrices $(\mathbf{M}, \boldsymbol{\Upsilon})$ are the unique solution to the following minimization problem:

$$\min_{\mathbf{M},\boldsymbol{\Upsilon}} g(\mathbf{M}, \boldsymbol{\Upsilon})^{\intercal}\, \mathbf{W}\, g(\mathbf{M}, \boldsymbol{\Upsilon}), \tag{10}$$

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

The minimum distance estimator is the sample analog of (10):

$$(\widehat{\mathbf{M}}, \widehat{\boldsymbol{\Upsilon}}) = \arg\min_{\mathbf{M},\boldsymbol{\Upsilon}} \widehat{g}(\mathbf{M}, \boldsymbol{\Upsilon})^{\intercal}\, \widehat{\mathbf{W}}\, \widehat{g}(\mathbf{M}, \boldsymbol{\Upsilon}), \tag{11}$$

subject to the same constraints as above, where $\widehat{g}$ is a vector with elements defined in the same way as in (9), 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 $\mathbf{W}$. This is a standard minimum distance estimator defined over a finite-dimensional parameter space; under standard regularity conditions it is consistent and asymptotically normal (Newey and McFadden, 1994). As usual, inference must be adjusted when parameters are at or near the boundary (Politis and Romano, 1994; Andrews, 1999, 2000).[20]

In general, if we estimate a model with $K$ hidden land uses 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 $K = 2$ land uses and $T = 3$ time periods, we have 7 parameters to estimate in total.[21]

## 4.2 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 log likelihood function is

$$L = \sum_{i=1}^{N} \ln \Pr[Y_i], \tag{12}$$

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

$$\Pr[Y_i] = \sum_{s_1 \in \mathcal{S}} \cdots \sum_{s_T \in \mathcal{S}} \Pr[S_{i1} = s_1] \Pr[Y_{i1} | S_{i1} = s_1] \prod_{t=2}^{T} \Pr[S_{it} = s_t | S_{it-1} = s_{t-1}] \Pr[Y_{it} | S_{it} = s_t]. \tag{13}$$

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).

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). The EM algorithm is an iterative method that alternates between performing an expectation (E) step and a maximization (M) step, until convergence. The E-step computes the posteriors $\Pr[S_i | Y_i]$, where $S_i = (S_{i1}, ..., S_{iT})$, exploiting the HMM structure and Bayes rule. The M-step searches over the parameters to maximize the expected log-likelihood found on the E-step. See Appendix B for more details.[22]

[20] When Condition 4 is satisfied for more than one value of $y \in \mathcal{Y}$, the vector $g_{1yt}$ may be augmented accordingly. When that happens, or when the data on $Y_{it}$ has more than four time periods, $T \geq 4$, the model becomes overidentified.

[21] The total number of parameters before accounting for constraints for $T$ years 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).

[22] A key part in the E-step is the Baum-Welch algorithm. It calculates efficiently the probabilities

As a final observation, note that both ML and MD estimators can handle data that are missing at random (e.g. due to cloud cover) by restricting the observations to the cases where $Y_{it}$ are nonmssing. For instance, if $Y_{it}$ is missing at random with probability, say, 0.10, we need to throw out approximately $1-0.9^3 = 0.271$ of our data to compute $\widehat{\mathbf{M}}_{y_{t+1},Y_t,Y_{t-1}}$ and $\widehat{\mathbf{M}}_{Y_t,Y_{t-1}}$ when calculating the MD estimator.

# 5 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). Then, we fix the number of observations and evaluate how the estimators perform at different true transition probabilities, misclassification probabilities, and with different numbers of time periods.

## 5.1 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

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

$\Pr[S_{it}|Y_{i1}, Y_{i2}, \ldots, Y_{iT}]$, where $t \le T$, which is the posterior probability of the true land use at $t$ given the *full* history of classifications.

Recall that the elements of $\mathbf{\Upsilon}$ are $\Pr\left[Y_{it} = y | S_{it} = s\right]$ (with $Y_{it}$ along the rows and $S_{it}$ along the columns). That means that the probability of classifying land use $y = 1$ when the true land cover is actually $s = 2$ is just $\Pr\left[Y_{it} = 1 | S_{it} = 2\right] = 0.2$. Correct classification probabilities here are either 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 hidden Markov model 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. 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 $\mathbf{\Upsilon}$ 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, $\mathbf{W} = \mathbf{I}$.

## 5.2 Baseline Results

Table 1 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 2, 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 value is 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 (by factors between 6 and 42, depending on the sample size). 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.

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.[23] While the additional time periods require the estimation of additional parameters, the larger number of time periods could help improve the precision of the misclassification probability estimates. In Figure 3, we replicate the results from Figure 2 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.

These considerations suggest combining the MD and ML in practice, whenever possible, taking into account their strengths. For instance, one could run the MD estimator first (which is fast), and then use the MD estimate as an initial value for the (asymptotically more efficient) ML estimator.

## 5.3 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

[23] In other words, we set

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

transition probability for state $s = 1$ in the last period.

Figure 4 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. When we look at the estimates of the misclassification rate, the estimates are more similar for the MD and ML approaches, but the ML is more biased as the true misclassification rate increases.

Figure 5 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 4. These graphs show that both MD and ML estimators continue to perform well at estimating transitions and misclassifications with no notable differences between them (aside from those discussed above).

# 6 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).[24] We outline the data, the implementation of our methodology, and the validation results.

## 6.1 Ground-Level and Remote-Sensing Data

The ground-level 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 in an area extending from the coordinate

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

$(59^o 25' 14'' W, 14^o 2' 39'' S)$ [lower left] to the point $(54^o 25' 19'' W, 11^o 42' 16'' S)$ [upper right], within 14 municipalities in the most intensely cropped region of central Mato Grosso. The data is unprecedented in spatial and temporal coverage for the state – and arguably in general. (For details, see Coutinho et al. (2011) and Brown et al. (2013).)[25]

Not only are these data of unusually high quality for the Brazilian Legal Amazon (and in general), they are especially useful for us because longitudinal ground-level data are crucial for validating our HMM approach. They allow us to observe true land use transition probabilities and compare them to our estimates. We can also compare our HMM estimates of misclassification probabilities with a direct estimate of misclassifications $\Pr[Y_{it}|S_{it}]$. Also important to emphasize, longitudinal validation data are particularly difficult to obtain in practice; typically, validation data are composed of a single or repeated cross-sections, which cannot be used to verify transition probabilities. The Embrapa data therefore provide a unique opportunity to test the performance of the HMM correction in practice.

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, i.e. $\mathcal{S} = \{\text{crops}, \text{pasture}\}$. A small number of points do not fit into either of these categories (e.g. points classified as natural vegetation); we drop these, leaving us with 403 unique spatial points, each point observed for one to five years in 2006–2010 (unbalanced panel).[26]

We merge the ground truth land use data with remote-sensing data. Specifically, we use measurements from the sixteen-day composite Terra MODIS 250m.[27] 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 $R_{it}$ presented in Figure 1. As we discuss below, we use these variables to predict land use $Y_{it} = f(R_{it})$ in each year. We merged the MODIS data with the Embrapa ground-level data considering the September-to-August harvest years for consistency. In this way,

[25] While Brown et al. (2013) use that data to obtain substantial progress towards more refined crop-specific classification, we focus on land use transition estimates. 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). A total of 40 farmers or farm managers were interviewed as research participants (Coutinho et al., 2011; Brown et al., 2013).

[26] 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.

[27] 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).

the 2006 ground-level data, for instance, are merged with sensor data from September 2005 to August 2006.

After merging the MODIS and the truth ground-level datasets, we randomly split the panel data into two disjoint sets. The first ("training set") is used to train a machine learning land use classifier. With the second set of data ("test set"), we obtain the predicted land use from the classifier, estimate the HMM model, and then test the predictions of that HMM model on the ground-truth validation data. The training set contains 60 cross-sectional points (286 point-years); the test set contains 343 points (1715 point-years).

We opt for a larger fraction of the Embrapa data set to be part of the test set because to reflect the typical scenario faced by applied researchers using satellite-based data. These researchers will typically have access to large panels of remote sensing data and machine-learning-based classifications. Therefore, they will have access to a lot of (potentially misclassified) data points ($Y_{it}$) to use in estimating the HMM model.

## 6.2 Implementation

For our frequency estimator, 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. Specifically, we use the training set to fit our GBM classifier, and then generate out-of-sample predictions $Y_{it} = f(R_{it})$ on the held-out test data.[28] The out-of-sample predictions in the test set constitute our land use classification panel data, $\{Y_{it} : i = 1, ..., N, t = 1, ..., T\}$, that is used to calculate the transition rates based on sample frequencies.

After predicting $Y_{it}$ on the test set, we implement our HMM corrections using both the MD and the ML estimators described previously.[29] We consider two model specifications: a restricted model with time-invariant transition probabilities, and another in which the transitions are allowed to vary over time. 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]

[28] 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.

[29] For the MD estimator, we used both $y_{t+1} = \textit{crops}$ and $y_{t+1} = \textit{pasture}$, 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. 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 $\delta_q$ 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 in computing the standard errors and subsample only on the test data.

### 6.3 Validation Results

The out-of-sample performance of our GBM classifications is shown in Table 2. 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 $\boldsymbol{\Upsilon}$ 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. 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., $\Pr[Y_{it} = \text{pasture} \mid S_{it} = \text{pasture}] = 0.795$).[31] This implies that $\boldsymbol{\Upsilon}$ is diagonally dominant, as required for identification of the HMM approach (see Condition 5 in Section 3.3.1). This increases our confidence that the HMM is identified when applied to the Embrapa test data.

Figure 6 shows the estimated results for the restricted hidden Markov model (i.e., imposing time-invariant transition probabilities). The ground-truth data indicate that the probability that cropland continues to be cropland in the following year is 99.3%, while the probability of switching to pasture equals 0.7%. The ground-truth probability of maintaining pasture land is 86.2%, and the probability of switching from pasture to crops 13.8%. So, both land uses are persistent over time, and cropland is more persistent than pasture.

The GBM classification (i.e. 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-level transition probabilities than the frequency estimates. The confidence intervals in the figure indicate that these results hold even after accounting for sampling uncertainty. This is consistent with the simulation results discussed previously, in which the frequency estimator tends to overestimate switching rates when land use is persistent.

Figure 7 is analogous to Figure 6, 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

[31] For comparison, NASS's widely-used Cropland Data Layer has recall rates of roughly 92% for corn and soy, a recall of roughly 49% for alfalfa, and an overall accuracy of roughly 90%, in Iowa in 2018. See `https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/metadata_ia18.htm` for more detail.

rates can vary over time (though not substantially in this data set).

Next, we turn to the HMM estimates of the misclassification probabilities; Figure 8 shows the results computed using MD and ML for both the restricted and unrestricted. 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 2.) 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 apply the Viterbi algorithm based on the HMM maximum likelihood point estimates to increase the accuracy of the land use predictions for each pixel in the test set, as discussed in Remark 1 of Section 3.3. Indeed, Figure 9 shows that correcting classifications using the HMM estimates and the Viterbi algorithm increases the overall accuracy of the land use classifications to 96%, improving on our original classifier's accuracy of 92%.

## 7 Conclusion

Satellite-based data allow researchers to access an unprecedented number of rich datasets with substantial spatial and temporal coverage. These data have proved useful in the study of a variety of important phenomena, including the incidence of pollution levels, changes to urban areas, land use changes, deforestation and regeneration processes, the evolution of biodiversity, among others. 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. Based on the econometrics measurement error literature, we show how to obtain estimates of transition probabilities that account for misclassification. The method is based on a formal set of assumptions that can be analyzed on a case-by-case basis (avoiding therefore ad hoc adjustments on transitions) and its implementation does not require ground-level truth validation data.

We propose two estimators based on the identification results and find that they perform well in Monte Carlo simulation studies and in a validation study based on a high-quality ground-level land cover data. The results suggest that, in many circumstances, researchers working with remote sensing data can obtain more accurate estimates of transitions rates than those based on standard practices, using readily available data.

## 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., A. Bartlett, and D. Saah (2013, 01). The landscape of conflict: IDPs, aid and land-use change in Darfur. *Journal of Economic Geography 13*(4), 589–617.

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.

Baragwanath, K., R. Goldblatt, W. You, G. Hanson, and A. Khandelwal (2019). Detecting urban markets with satellite imagery: An application to india. Technical report, Columbia University.

Bonan, G. B. (2008). Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. *Science 320*(5882), 1444–1449.

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

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.

Burgess, R., M. Hansen, B. A. Olken, P. Potapov, and S. Sieber (2012). The political economy of deforestation in the tropics. *The Quarterly Journal of Economics 127(4)*, 1707–1754.

Cisneros, E., S. L. Zhou, and J. Börner (2015). Naming and shaming for conservation: Evidence from the brazilian amazon. *PloS one 10*(9), e0136402.

Costinot, A., D. Donaldson, and C. Smith (2016). Evolving comparative advantage and the impact of climate change in agricultural markets: Evidence from 1.7 million fields around the world. *Journal of Political Economy 124*(1), 205–248.

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.

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.

Gennaioli, N., R. La Porta, F. Lopez-de Silanes, and A. Shleifer (2012, 11). Human Capital and Regional Development *. *The Quarterly Journal of Economics 128*(1), 105–164.

Goldblatt, R., M. F. Stuhlmacher, B. Tellman, N. Clinton, G. Hanson, M. Georgescu, C. Wang, F. Serrano-Candela, A. K. Khandelwal, W.-H. Cheng, and R. C. Balling (2018). Using landsat and nighttime lights for supervised pixel-based image classification of urban land cover. *Remote Sensing of Environment 205*, 253 – 275.

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.

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.

Henderson, J. V., T. Regan, and A. J. Venables (2016). Building the city: Sunk capital, sequencing, and institutional frictions. Technical report, CEPR Discussion Paper 11211.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012, April). Measuring economic growth from outer space. *American Economic Review 102*(2), 994–1028.

Holmes, T. and S. Lee (2009). Economies of density versus natural advantage: Crop choice on the back forty. *Review of Economics and Statistics 94*(1), 1–19.

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 J. Yao (2019). Illuminating economic growth. Technical report, Johns Hopkins University.

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

Kudamatsu, M., T. Persson, and D. Stromberg (2016). Weather and infant mortality in africa. Technical report, IIES.

Lark, T. J., R. M. Mueller, D. M. Johnson, and H. K. Gibbs (2017). Measuring land-use and land-cover change using the U.S. department of agriculture's cropland data layer: Cautions and recommendations. *International Journal of Applied Earth Observation and Geoinformation 62*, 224 – 235.

Lark, T. J., J. M. Salmon, and H. K. Gibbs (2015). Cropland expansion outpaces agricultural and biofuel policies in the United States. *Environmental Research Letters 10*(4).

Lillesand, T., R. W. Kiefer, and J. W. Chipman (2004). *Remote Sensing and Image Interpretation* (5 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.

Michalopoulos, S. and E. Papaioannou (2013). Pre-colonial ethnic institutions and contemporary african development. *Econometrica 81*(1), 113–152.

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

Nordhaus, W. and X. Chen (2014, 05). A sharper image? Estimates of the precision of nighttime lights as a proxy for economic statistics1. *Journal of Economic Geography 15*(1), 217–246.

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

Scott, P. T. (2013). Dynamic discrete choice estimation of agricultural land use. *Working Paper*.

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

**Table 1:** Baseline Monte Carlo Simulation Results

| | | 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 2:** Confusion Matrix based on Embrapa Validation Data

| | 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 | |

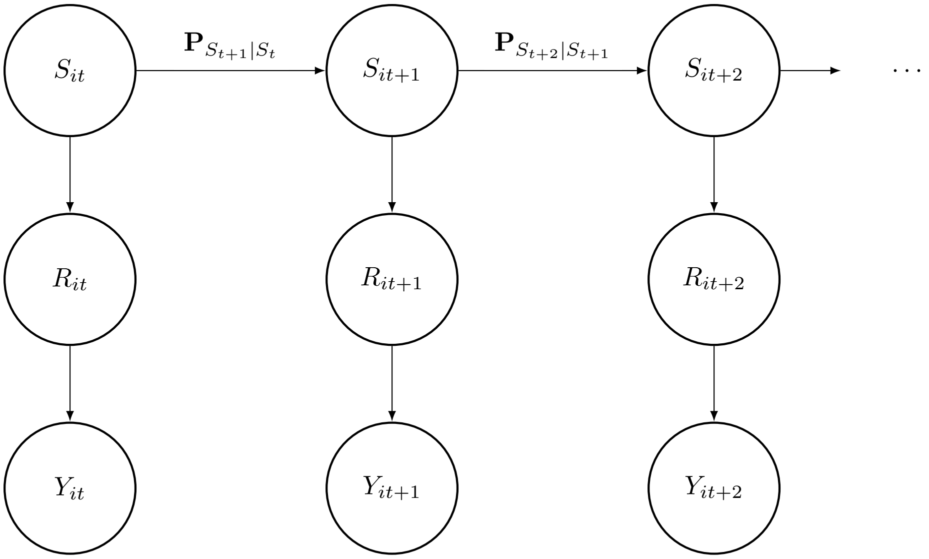
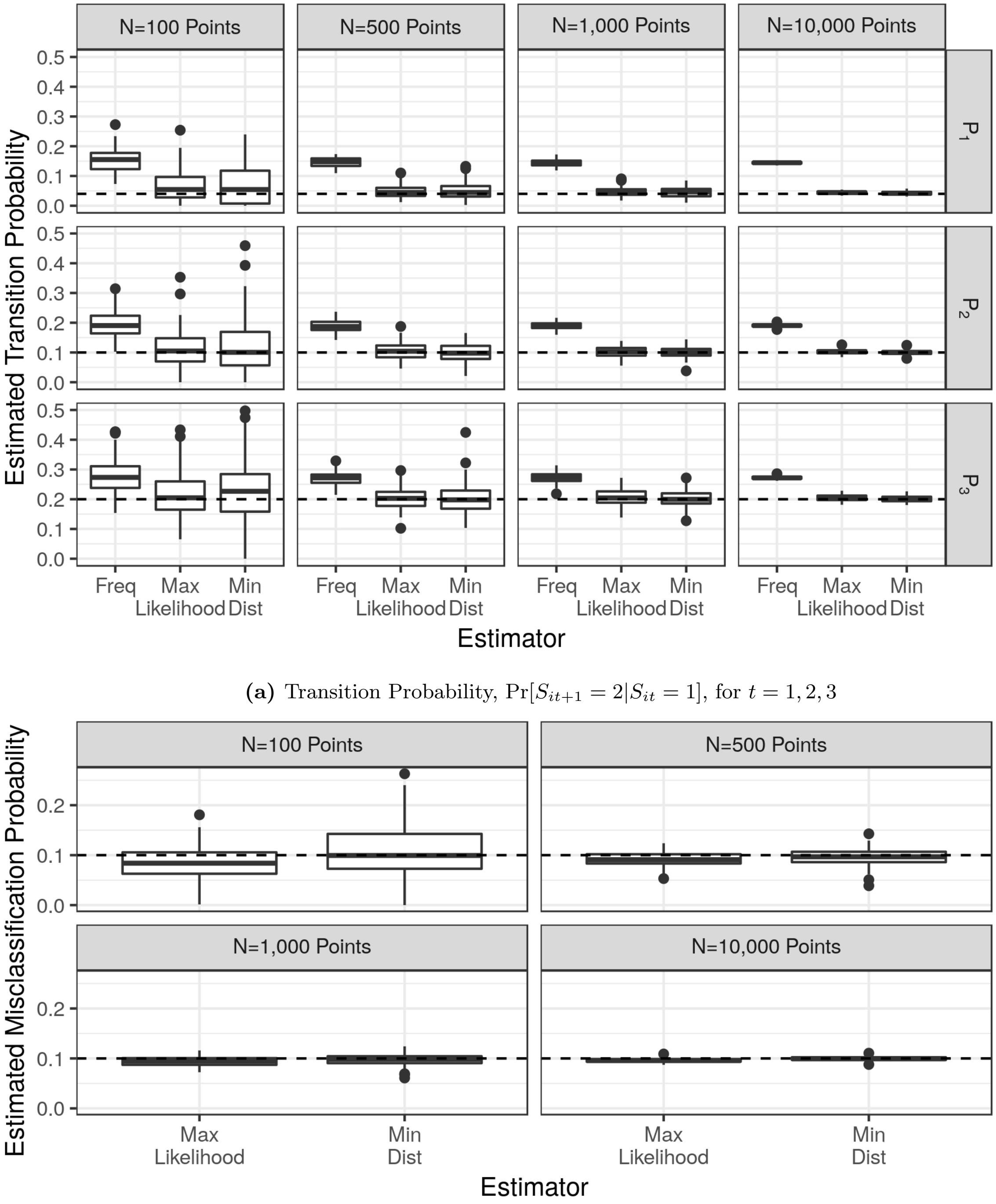


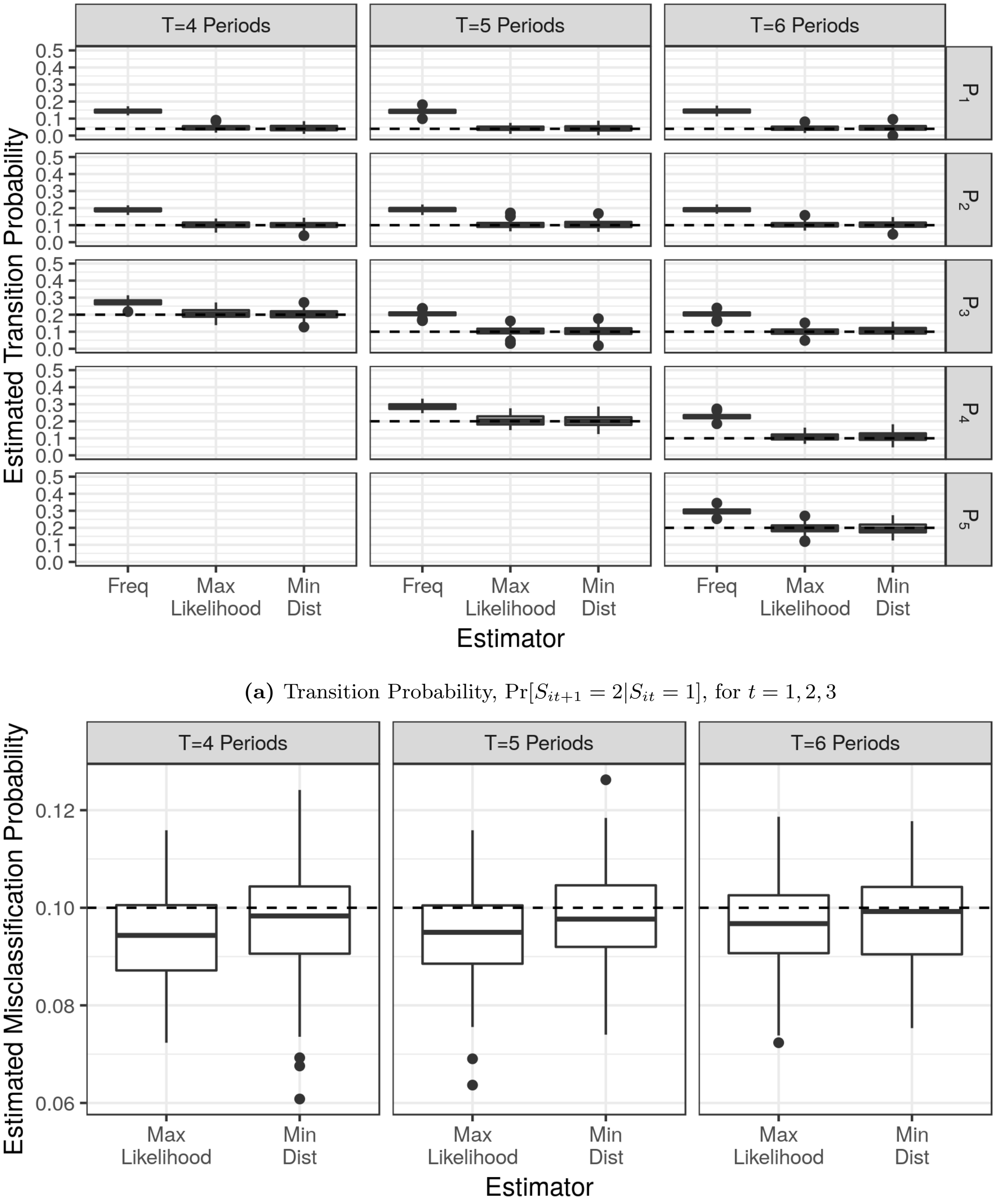
**Figure 1:** Dependency Graph for Hidden Markov Model



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

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

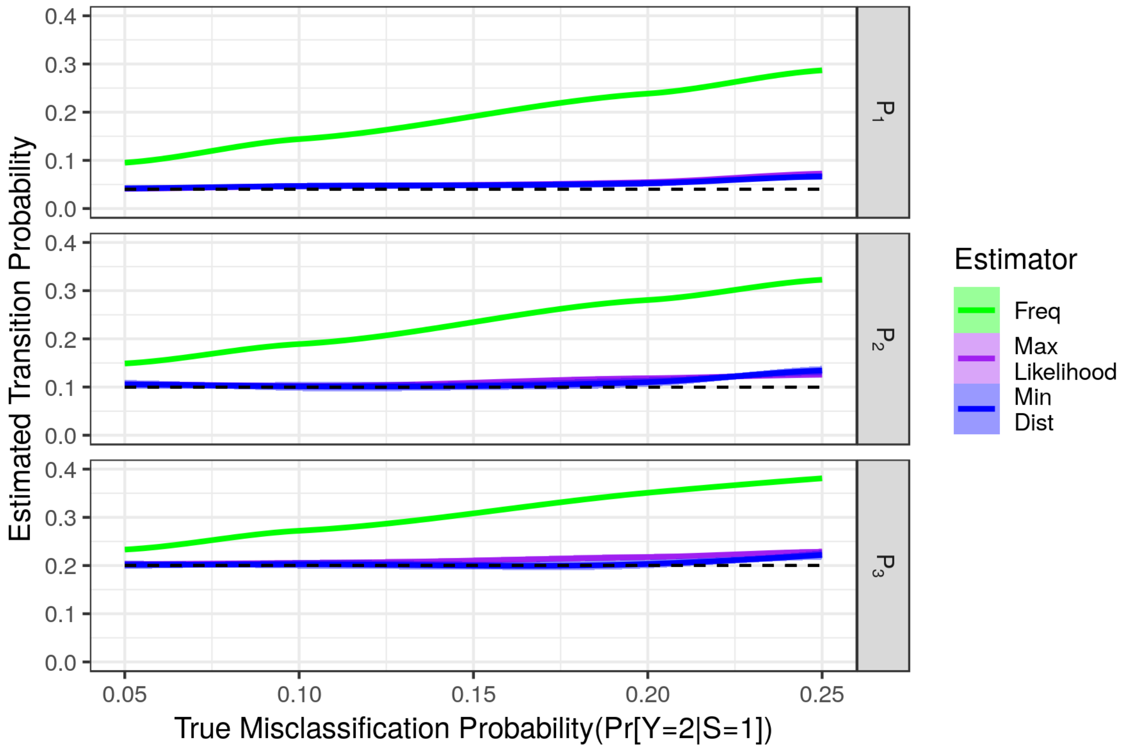
**Figure 2:** Baseline Monte Carlo Simulation Results



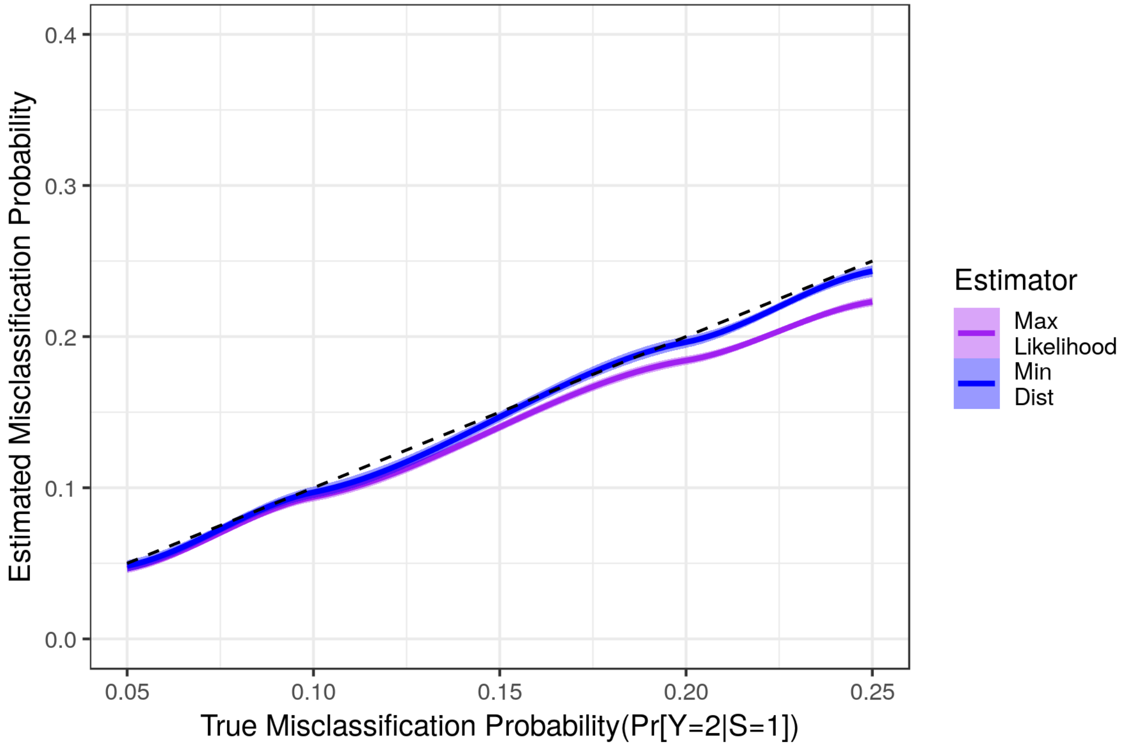
**(a)** Transition Probability, $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$

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

**Figure 3:** Baseline Monte Carlo Simulation Results

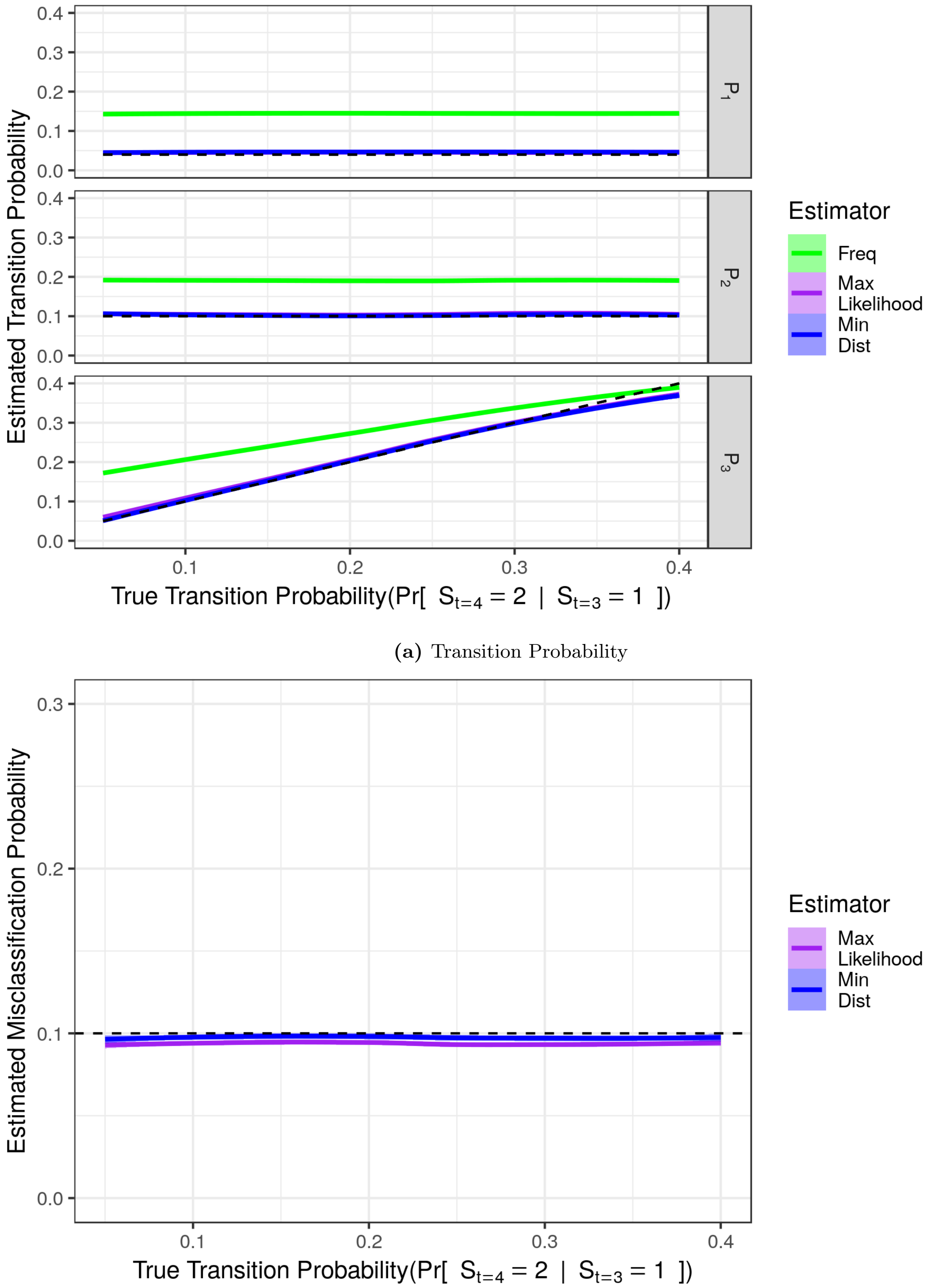


**(a)** Transition Probability



**(b)** Misclassification Probability

**Figure 4:** Monte Carlo Results for Varying Misclassification Probabilities



**(a)** Transition Probability

**(b)** Misclassification Probability

**Figure 5:** Monte Carlo Results for Varying Transition Probabilities

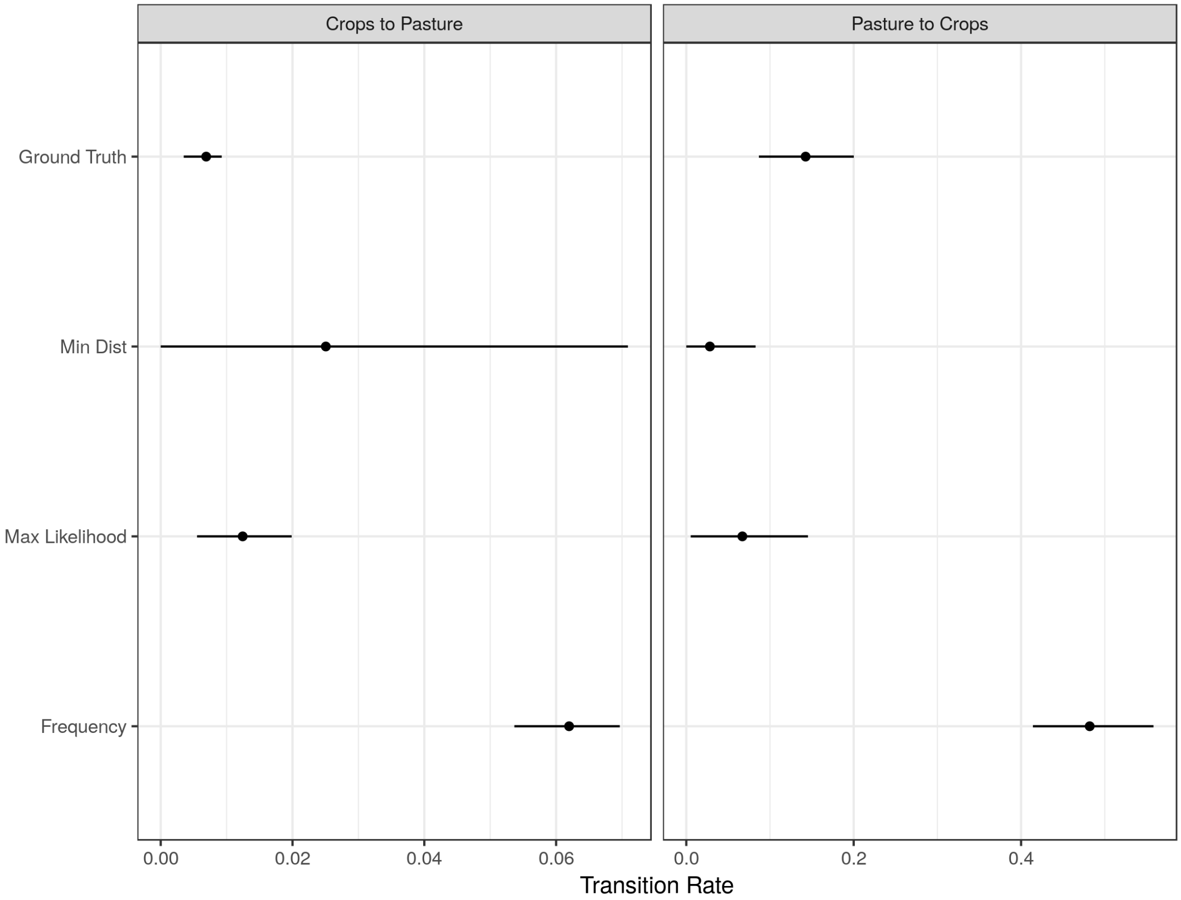


**Figure 6:** Time-invariant Transition Probabilities – Embrapa Validation Data

**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.

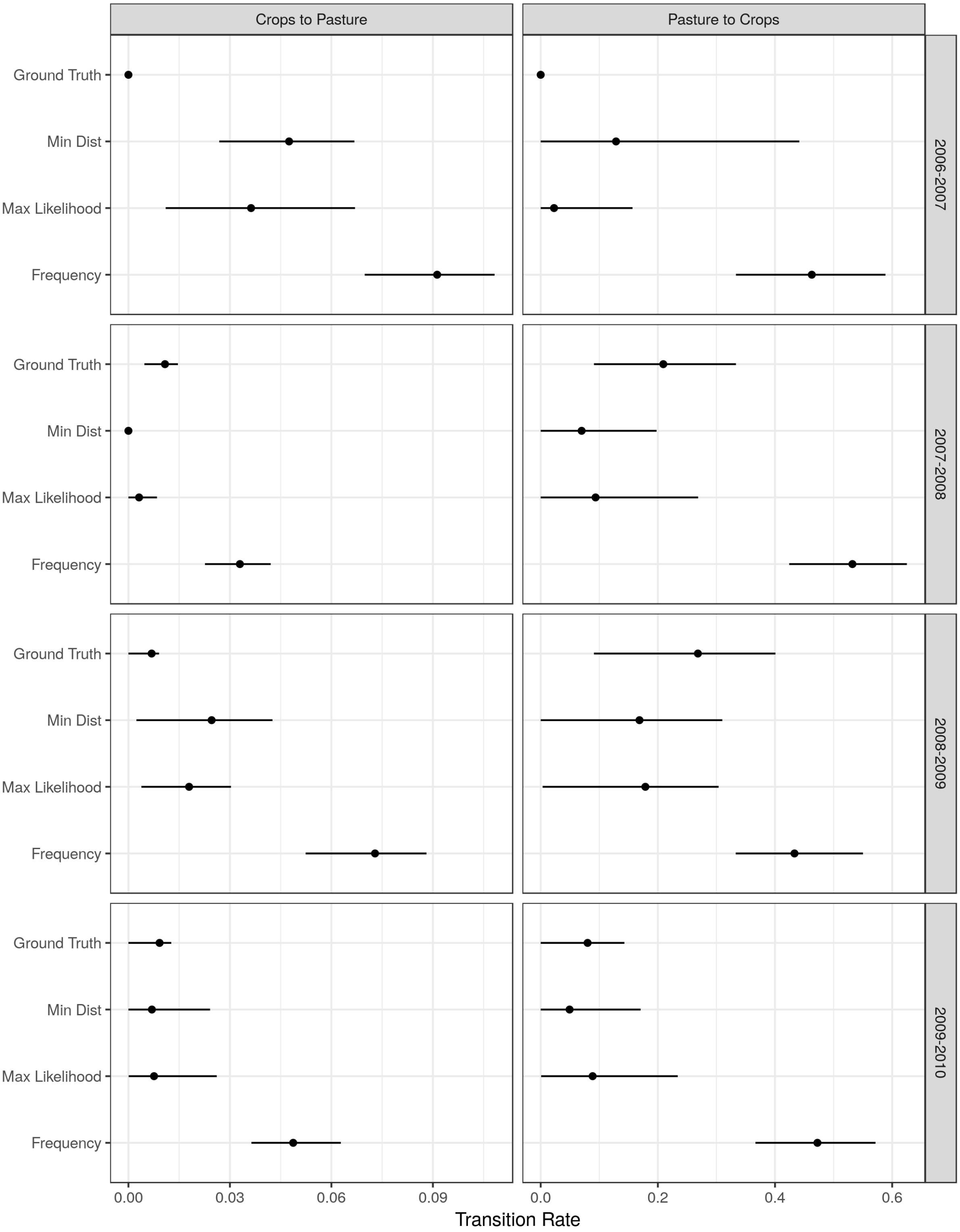


**Figure 7:** Time-varying Transition Probabilities – Embrapa Validation Data

**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.

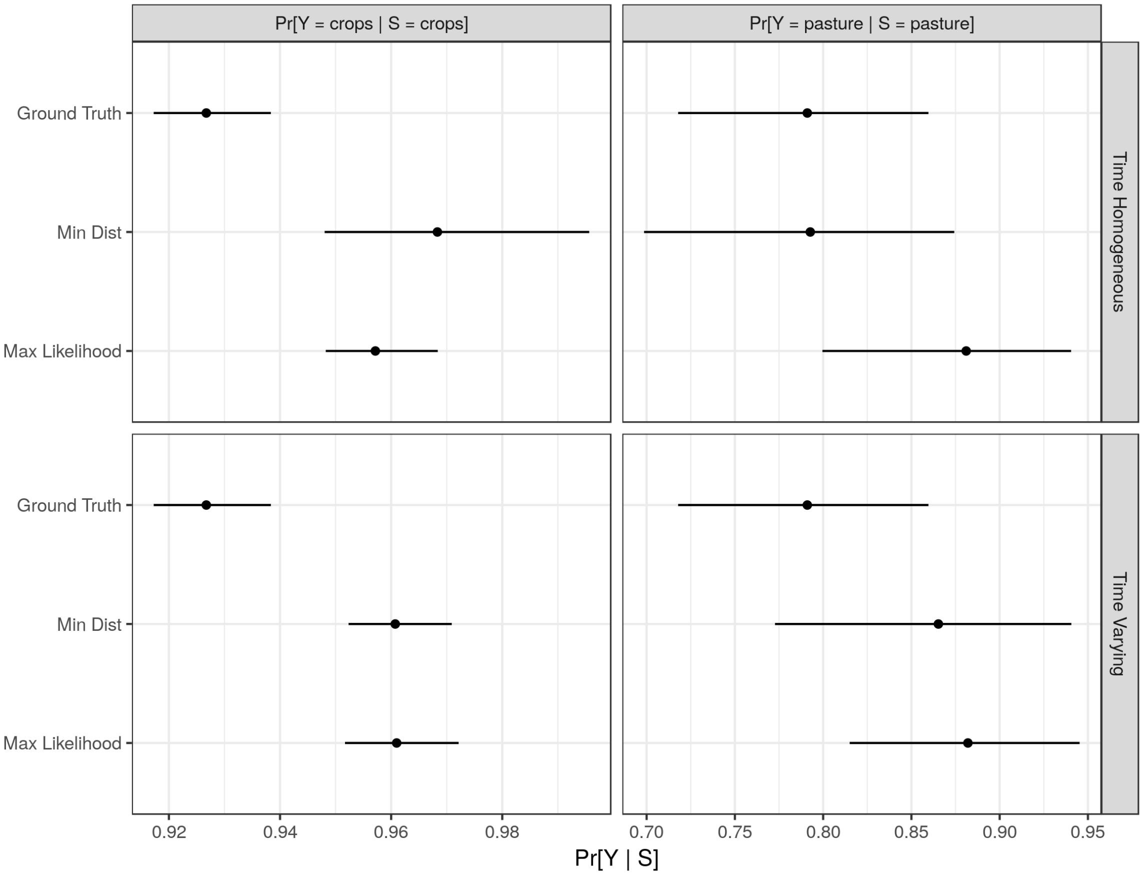


**Figure 8:** Misclassification Probabilities – Embrapa Validation Data

**Note:** *Ground Truth* correspondsdata 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.

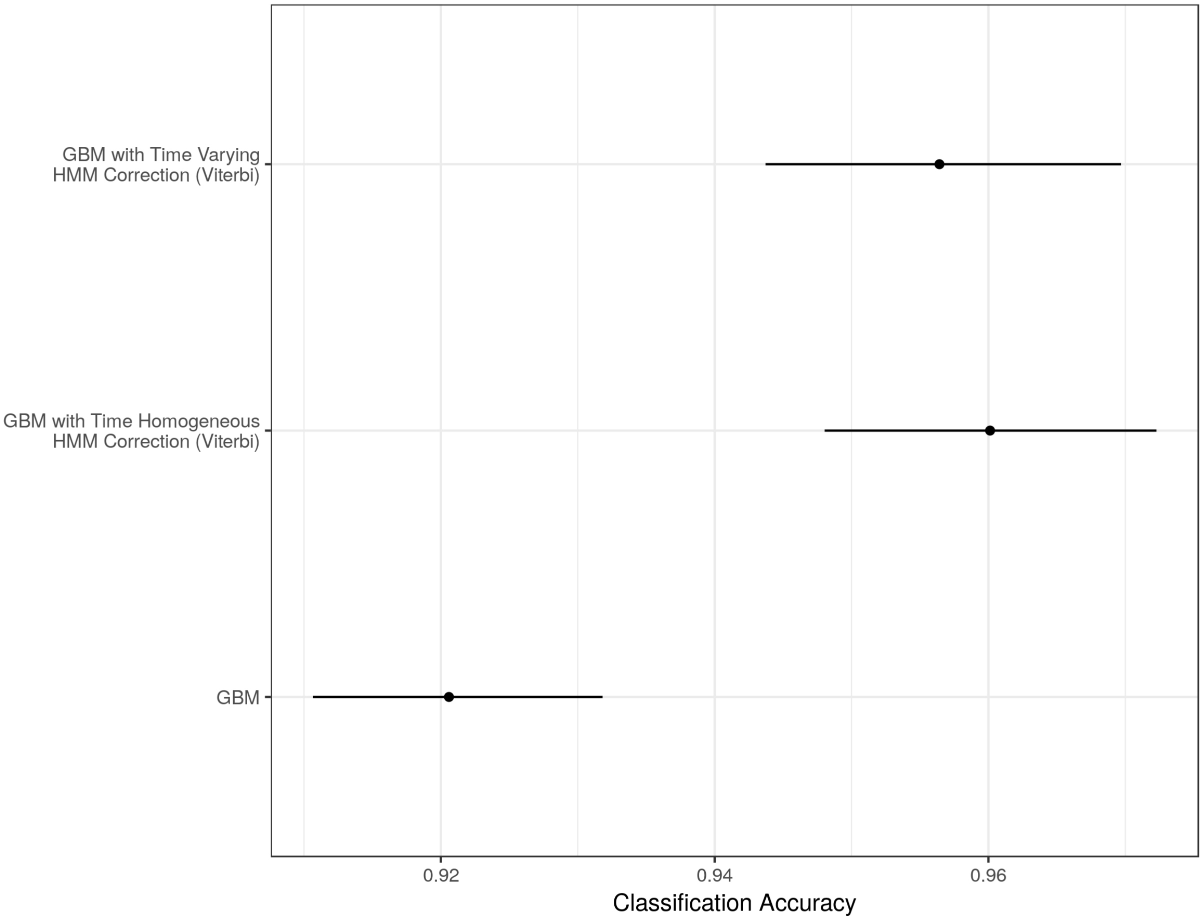


**Figure 9:** Classification Accuracy of GBM and HMM-Viterbi methods in the Embrapa Validation Data

**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.

# A Appendix: 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

$$\Pr\left[Y_{it}, Y_{it-1}\right] = \sum_{s\in\mathcal{S}} \Pr\left[Y_{it}|S_{it}=s\right]\Pr\left[S_{it}=s, Y_{it-1}\right]. \tag{14}$$

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

$$\begin{aligned}\Pr\left[Y_{it+1}, Y_{it}\right] &= \sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \Pr\left[Y_{it+1}|S_{it+1}=s'\right]\Pr\left[S_{it+1}=s'|S_{it}=s\right] \\ &\quad \times \Pr\left[Y_{it}|S_{it}=s\right]\Pr\left[S_{it}=s\right] \\ &= \sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \Pr\left[Y_{it+1}|S_{it+1}=s'\right]\Pr\left[S_{it+1}=s', S_{it}=s\right]\Pr\left[Y_{it}|S_{it}=s\right], \end{aligned} \tag{15}$$

where the first equality follows from the law of total probability and the HMM assumption (i.e., equation (3) 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

$$\Pr\left[Y_{it+1}, Y_{it}, Y_{it-1}\right] = \sum_{s\in\mathcal{S}} \Pr\left[Y_{it+1}|S_{it}=s\right]\Pr\left[Y_{it}|S_{it}=s\right]\Pr\left[Y_{it-1}, S_{it}=s\right], \tag{16}$$

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 (3)); the fourth equality rearranges the terms in the summations; and the fifth equality follows from the law of total probability.

In matrix notation, equations (14)–(16) are equivalent to the equations (4)–(6) presented in the main text.

# B Appendix: The EM Algorithm

We now briefly explain the EM algorithm. To simplify notation, let $\theta$ represent the collection of HMM parameters, i.e. $\theta$ is a list containing $\Pr\left[S_{i1}\right]$, $\Pr\left[S_{it+1}|S_{it}\right]$ for $t = 1, ..., T-1$, and $\Pr\left[Y_{it}|S_{it}\right]$, 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

$$l\left(\theta\right) \equiv \ln \Pr\left[Y = y; \theta\right] \tag{17}$$

and let

$$J\left(\theta, \theta'\right) \equiv \sum_{s} \Pr\left[S = s|Y = y; \theta'\right] \ln \left\{ \frac{\Pr\left[Y = y, S = s; \theta\ \right]}{\Pr\left[Y = y, S = s; \theta'\right]} \right\}. \tag{18}$$

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

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

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 (because transitions are rare), and that the true land use was forest for the entire 15 years. The HMM 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 $\boldsymbol{\Upsilon}$ are zero, for example, we do not need any smoothing. Identifying the parameters from observed data is therefore crucial in applications.

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 our assumption of spatial independence. The updated values of $\theta$ are

$$\begin{aligned} \mu^{(j+1)}[k] &= \frac{\sum_i \pi_{i1}[k]}{\sum_{i,s} \pi_{i1}[s]} \\ \mathbf{P}_t^{(j+1)}[k,l] &= \frac{\sum_i \pi_{it}[k,l]}{\sum_i \pi_{it}[k]} \\ \boldsymbol{\Upsilon}^{(j+1)}[y,k] &= \frac{\sum_{i,t:Y_{it}=y} \pi_{it}[k]}{\sum_{i,t} \pi_{it}[k]} \end{aligned} \tag{19}$$

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 $\boldsymbol{\Upsilon}$, 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.