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

Adrian L. Torchiana, Ted Rosenbaum,

Paul T. Scott, and Eduardo Souza-Rodrigues†

September 2019

Abstract

Satellite-based data provide wide unprecendent geographic and temporal coverage with high spatial resolution at relatively low financial cost. They have therefore been increasingly used in research and policy. However, errors in the classification techniques typically used to process these data can lead to misleading conclusions about transitions processess (e.g., deforestation, urbanization, and industrialization). In this paper, we propose a correction for transition estimates based on a hidden Markov model (HMM) to extract the signal (truth) from its noisy measurement (satellite-based classifications). We discuss the conditions for identification formally, and show that the proposed correction produces consistent estimates of transition rates using simulations and a high-quality validation dataset, whereas estimates that do not account for the classification errors are severely biased. An attractive feature of our approach is that no ground-level truth data is required to implement the correction.

Keywords: Misclassification, Measurement Error, Remote-Sensing Data, Satellite-Based Data, Transition Process, Land Cover, Hidden Markov Model, Minimum Distance Estimator, Maximum Likelihood Estimator, EM Algorithm

*We are extremely grateful to Alexandre Camargo Coutinho and Daniel De Castro Victoria, who generously shared their land cover validation 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 unprecendented 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 coverages (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). These classification errors can also affect estimates of the transition processes of outcomes of interest – our focus in this paper. Intuitively, errors in classifications may result in excessive transitions. For example, much of the apparent land cover change in satellite-based data may be the result of misclassifications (Abercrombie and Friedl, 2016), which may undermine calculations of carbon stock cycles and the establishment of efficient preservation policies. 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 transitions in land use. An alternative solution is to correct classification errors using validation data that can be treated as ground truth. But 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.

[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), 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. We focus in this paper on the changes over time.

In this paper we propose a different approach. We present a hidden Markov model (HMM) that corrects for misclassification bias in satellite-based transition probabilities. A hidden Markov model is the combination of an unobserved Markov process with observations that depend only on the contemporaneous hidden state. For instance, when studyng land use changes, the ground truth land use is the hidden state and remote sensor-based classifications are the observations.[4] 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. An attractive feature of our approach is that no ground-level truth data is required to implement the correction.

Based on Hu's seminal work on non-classical measurement error (summarized in Hu (2017)), 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; some of them are testable, which means researchers can verify them in particular cases.[5] We propose 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 based on the expectation-maximization (EM) algorithm (van Handel, 2008).

We investigate the performance of our strategy based on a Monte Carlo simulation study, and on 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 also perform a validation exercise using a unique ground-level truth longitudinal data produced by the Brazilian Agricultural Research Corporation (Embrapa). The data set contain information on land use in various private farmlands in the state of Mato Grosso, Brazil, from 2006 to 2010.[6] The state of Mato Grosso provides an interesting setting because it is a major center

[4] Hidden Markov models are used in many applications in which only imperfect measurements are available for some time-varying value of interest. See McLachlan and Peel (2000), chapter 13, for a textbook overview.

[5] For instance, we assume that misclassification in year t does not affect the probability of misclassification in year $t+1$, after conditioning on the true state in both years. This assumption would not be appropriate when using algorithms that explicitly rely on past classifications to generate the current year's classification, but it is arguably plausible when classifications have been generated separately for each year of data, as is common in practice.

[6] 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

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 probabilities and compare them to our estimates. As such the Embrapa data set provide a unique opportunity to test the performance of the HMM correction in practice.

In the validation exercise, we begin by generating land use classifications from remote sensing data from MODIS (which are regularly used to classify land cover around the globe), and using half of the ground truth data as training data. We find that the MODIS-based classification performs well in terms of the out-of-sample predictions (using the other half of the ground-level data as the validation data): The misclassifications are small in magnitude and are comparable to other validation studies. Then, we compare the transition probabilites estimated with no corrections for misclassifications and estimates based on our HMM correction. When transition probabilities are computed without correction for misclassifications, land use transition rates are substantially greater than the transitions observed in the ground truth data: Estimated transitions are approximately 54%–68% higher than the truth. In contrast, transition probabilities estimated using our HMM approach match the transition probabilities observed in the validation data far more accurately. Furthermore, the HMM correction estimates misclassification probabilities (surprisingly) well too, as a by-product of the method.

The closest paper to ours is Abercrombie and Friedl (2016), who also consider an HMM-based correction to errors in land use classifications. The main distinctions between our approaches are the following: We link the HMM procedure to formal identification results based on a set of explicit assumptions, providing transparent arguments for the conditions under which the approach can and cannot be implemented; we propose consistent estimators for transitions; and we allow for the estimation of time-varying transition probabilities from the data. Abercrombie and Friedl, in contrast, do not provide formal conditions under which their HMM correction is expected to be valid, and they use a model with *inputed* and *time-invariant* transition probabilities. We believe that being able to handle time-varying transition probabilities flexibly is important in many applications, e.g. when estimating how (and explaining why) deforestation processes may change over time.

more refined crop-specific classification, we focus on land use transition estimates.

This paper is organized as follows: Section 2 lays out the framework and the misclassification problem; Section 3 investigates the use of validation data to correct misclassifications, and then 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 study; Section 6 describes the validation exercise using ground-level truth data from Brazil's Embrapa; and Section 7 concludes.[7]

2 Framework

In this section, we present the basic framework, a simple correction based on vaidadtion data, the hidden Markov model, and the HMM identification results. Our running example is the land use classification problem, but results can be applied to other classification problems using panel remote-sensing data.

Let $S_{it} \in \mathcal{S}$ denote the ground truth land use at location i at time t. In our application, a location is usually a pixel or a spatial point. The set of possible values that S_{it} can take is $\mathcal{S} = \{s_1, ..., s_K\}$, $K < \infty$. E.g., $\mathcal{S}$ may include forest, pasture, and different types of crops. 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.[8] 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 the vegetation index, as well as 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 algorithm (see discussion below).

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

[7] The Appendix presents (i) some mathematical derivations, (ii) details on the EM algorithm, and (iv) a stylized exampe of how misclassifications in land use change may lead to mistakes in conservation policies. Codes for replicating the Monte Carlo simulations are available at https://github.com/atorch/hidden_markov_model.

[8] For variables taking value on the real line, one needs to work with operators in Hilbert spaces, instead of matrices in Euclidean spaces.

and a small number of time periods. Under standard regularity conditions, longitudinal data on Y_{it} can then 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. Although 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 estimation can therefore be done separately for each subregion.[9]

For pixel i at time period t, the probability of observing a (satellite-based) 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. In matrix notation, we have

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

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, while the misclassification probabilities are the off-diagonal terms. For now, we consider the case where $\boldsymbol{\Upsilon}$ is time-invariant, but we extend results for misclassifications that may change over time.

While the vector $\mathbf{P}_{Y_t}$ can be estimated in the data using frequency estimators, it is not possible to recover the true land use distribution $\mathbf{P}_{S_t}$ without additional information. Furthermore, there is no guarantee that the observed (estimated) transition $\Pr[Y_{it+1}|Y_{it}]$ is close to the true transitions $\Pr[S_{it+1}|S_{it}]$. To illustrate the potential importance of misclassification problems, we present a stylized exampe in the Appendix showing how misclassifications in land use may lead to inefficient conservation policies (Appendix C).

3 Correcting Satellite-Based Misclassifications

We now consider two possible corrections for misclassified remote-sensing data. The first correction is based on available validation data. The second one is based on our hidden Markov model approach.

[9] Incorporating continously distributed covariates, such as slope and altitude, is more challenging. One can apply standard kernel smoothing techniques, or parameterize the transition probability functions.

3.1 Correction Based on Validation Data

Suppose we have access to a subset of 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, it is standard practice to split the sample into training and validation sets. Supervised classification algorithms are applied to the training data to predict land uses. Predicted land uses are extrapolated to the validation data and then compared to the ground-truth data. The comparison produces the so-called "confusion matrix," which is a matrix with the joint distribution $\Pr[Y_{it}, S_{it}]$. The matrix $\boldsymbol{\Upsilon}$ with the conditional probabilities 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 validation dataset in two. One subsample can be used to estimate $\boldsymbol{\Upsilon}$, and the other, to check whether the corrected land use classification is 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 classifications), we need an approach that allows us to estimate more than just the marginal probabilities $\Pr[S_{it}]$.[10]

3.2 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 model (HMM). Specifically, we assume ground truth land cover $\{S_{it}\}$ follows a first-order Markov 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

[10] To see why marginal probabilities are not enough to obtain transitions, take a binary case $\mathcal{S} = \{s_1, s_2\}$. By the 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 though 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 $\Pr[S_{it+1} = s_1|S_{it} = s_2]$.

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).[11] 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.[12] The stochastic process is summarized graphically in Figure 1, in which nodes represent random variables and edges indicate statistical dependence.

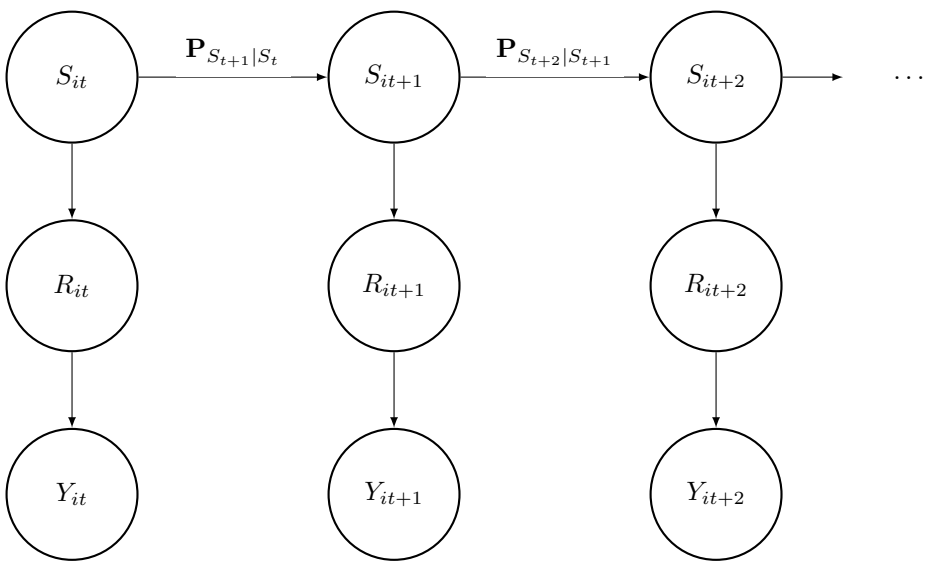


Figure 1: Dependency Graph for HMM

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 $y \in \mathcal{Y}$, define the matrix $\mathbf{M}_{y_{t+1},X,W}$, with elements $\Pr[Y_{it+1} = y, X = y_l, W = y_k]$, as well as the diagonal matrix $\mathbf{D}_{y_{t+1}|X}$, with diagonal entries $\Pr[Y_{it+1} = y|X = s_k]$. Then, from the joint

[11] Note that the 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. We therefore focus on temporal dependence within each spatial point.

[12] 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 model 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.

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

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

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

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

and, from the joint distribution of $(Y_{it+1}, Y_{it}, Y_{it-1})$,

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

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

3.2.1 Identification of the Hidden Markov Model

This subsection follows the identification results of Hu (2017). We adapt his general framework to our land use model. The objective is to identify the Markov land use transition process $\Pr[S_{it+1}|S_{it}]$, the initial distribution $\Pr[S_{it}]$, and the misclassification probabilities $\boldsymbol{\Upsilon}$ using a longitudinal data on Y_{it}. Assume we have data for at least $T \geq 3$ periods, and impose the following condition:

Condition 1. *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.[13] The condition implies that both matrices $\boldsymbol{\Upsilon}$ and $\mathbf{M}_{S_t,Y_{t-1}}$ are invertible too (see equation (4)). Next, 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 observable 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

[13] 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.

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 2. $\Pr[Y_{it+1} = y|S_{it} = s] \neq \Pr[Y_{it+1} = y|S_{it} = s']$ *for any given* $y \in \mathcal{Y}$ *whenever* $s \neq s'$, *and* $s, s' \in \mathcal{S}$.

Condition 2 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.

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,

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

for $y = forest$. Condition 2 is then satisfied.[14]

Next we turn to the eigenvectors.

Condition 3. $\Pr[Y_{it} = s^*|S_{it} = s^*] > \Pr[Y_{it} = s|S_{it} = s^*]$ *for any* $s \neq s^*$, *and* $s, s^* \in \mathcal{S}$.

Condition 3 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 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.

Next, given identification of $\boldsymbol{\Upsilon}$, we identify the joint distribution $\Pr[S_{it+1}, S_{it}]$ under the assumption that $\boldsymbol{\Upsilon}$ is time-invariant. More specifically, we impose

Condition 4. $\Pr[Y_{it+1}|S_{it+1}] = \Pr[Y_{it}|S_{it}]$.

[14] Condition 2 imposes assumptions on $\Pr[Y_{it+1}|S_{it}]$. Yet, because

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

Condition 2 can be derived based on restrictions imposed on the primitives $\Pr[Y_{it+1}|S_{it+1}]$ and $\Pr[S_{it+1}|S_{it}]$. 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 2 holds for more than one value y, the model becomes overidentified. We can compute (7) for, say, y and y', with $y \neq y'$, and check if we obtain the same eigenvectors $\boldsymbol{\Upsilon}$.

Given Condition 4 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 $\Pr\left[S_{it}\right]$.

We conclude that it is possible to estimate misclassification probabilities and true land use transition probabilities with no validation data, provided (a) we have access to a noisy classification observed in at least three consecutive time periods, (b) the stochastic process $\{Y_{it}, S_{it} : t = 1, 2, ...\}$ follows a hidden Markov model (HMM), and (c) Conditions 1–4 hold.

Theorem 1. *(Hu, 2017). Suppose equation (3) and Conditions 1–4 hold. Then, the joint distribution of the observed classifications* $(Y_{it+1}, Y_{it}, Y_{it-1})$ *uniquely identifies* $\Pr\left[Y_{it}|S_{it}\right]$, $\Pr\left[S_{it}\right]$, *and* $\Pr\left[S_{it+1}|S_{it}\right]$.

When Condition 4 is not satisfied, we need an additional time period to identify misclassification probabilities at both t and $t+1$ (i.e. we would need at least $T \geq 4$ periods of data).[15]

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.

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, in our experience, based on simulation exercises, the eigenvalue-eigenvector decomposition may result in estimated probabilities that are negative or greater than 1 in some data sets. This is more likely to happen when the sample size is small and the true parameters are close to 1 (e.g. transition probabilities of 0.99). For this reason, it is better to implement a constrained minimum distance estimator.

[15] When both transition probabilities and misclassification probabilities vary by time, equation (7) becomes $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}\mathbf{M}^{-1}_{Y_t,Y_{t-1}} = \boldsymbol{\Upsilon}_t\,\mathbf{D}_{y_{t+1}|S_t}\,\boldsymbol{\Upsilon}_t^{-1}$, i.e. the misclassification probability matrices have the same time subscript and we still have an eigenvalue-eigenvector decomposition. However, because this version of equation (7) uses Y_{t-1} and Y_{t+1} (in addition to Y_t) to identify $\boldsymbol{\Upsilon}_t$, it follows that $\boldsymbol{\Upsilon}_1$ and $\boldsymbol{\Upsilon}_T$ are not identified. Equation (8) becomes $\mathbf{M}_{S_{t+1},S_t} = \boldsymbol{\Upsilon}^{-1}_{t+1}\mathbf{M}_{Y_{t+1},Y_t}\left(\boldsymbol{\Upsilon}_t^{\intercal}\right)^{-1}$, and hence we cannot identify the first and last transition matrices.

Formally, for some $y \in \mathcal{Y}$, define the functions

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

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). To simplify notation, we consider $\mathbf{M}_{S_{t+1},S_t} = \mathbf{M}_t$ for all t, and collect all matrices into $\mathbf{M} = \{\mathbf{M}_t : t = 1, ...T-1\}$. 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 $t \in \{1, \ldots, T-1\}$, where $T \geq 3$. Define the vector $g = (g_1', g_2')'$, 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, 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 1.

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

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

subject to the same constraints as above, and where $\widehat{g}$ is the same as (9), but replaces $\mathbf{M}_{y_{t+1},Y_t,Y_{t-1}}$, $\mathbf{M}_{Y_t,Y_{t-1}}$, 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 $\widehat{\mathbf{M}}_{Y_{t+1},Y_t}$. 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 (Andrews, 1999). When Condition 2 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.

In general, if we estimate a model with K hidden land uses from T years of data, we have to optimize over $K + T\,K^2$ parameters subject to $T\,K + 1$ equality constraints.[16]

[16] 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).

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 for i integrates-out the hidden states[17]

$$l_i = \ln \sum_{s_t \in \mathcal{S}: t=1,...T} \Pr[Y_{i1}, ..., Y_{iT}, S_{i1} = s_1, ..., S_{iT} = s_T], \tag{12}$$

and the log likelihood function for the entire panel data is $L = \sum_{i=1}^{N} l_i$. The ML estimator chooses the transition probabilities for S_{it}, the initial distribution of S_{it}, and the misclassification probabilities that maximizes the function L. Because directly maximizing L is difficult, we follow the literature and use the expectation-maximization (EM) algorithm (van Handel, 2008). See Appendix B for more details.

Note that the combination of the initial distribution and the transition probabilities for S_{it}, $t \in \{1, \ldots, T-1\}$ in (12), carries the exact same information as the joint distributions $\left\{\mathbf{M}_{S_{t+1},S_t}\right\}$ estimated using the MD estimator.

5 Monte Carlo Studies

In this section, we present Monte Carlo experiments to investigate the finite-sample properties of the MD and EM 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 estimator performs 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

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

[17] For instance, under the HMM assumptions, we have for $T = 3$,

$$\begin{aligned} \ln \Pr[Y_{it+2}, Y_{it+1}, Y_{it}] &= \ln \sum_{s'' \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} \Pr\left[Y_{it+1} | S_{it+1} = s''\right] \Pr\left[S_{it+1} = s'' | S_{it} = s'\right] \\ &\quad \times \Pr\left[Y_{it} | S_{it} = s'\right] \Pr\left[S_{it} = s' | S_{it-1} = s\right] \Pr\left[Y_{it-1} | S_{it-1} = s\right] \Pr\left[S_{it-1} = s\right]. \end{aligned}$$

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 for state $s = 2$ the same over time, $\Pr[S_{it+1} = 2|S_{it} = 2]$.

The misclassification probabilities are time-invariant and given by

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

Recall that the elements of $\boldsymbol{\Upsilon}$ are $\Pr[Y_{it} = y|S_{it} = s]$ (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 $\Pr[Y_{it} = 1|S_{it} = 2] = 0.2$. Correct classification probabilities here are either 0.9 (for $s = 1$) and 0.8 (for $s = 2$), which are highly accurate when compared to typical out-of-sample predictions in real data settings.

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 changes. 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 verify the performance of the proposed HMM corrections, based on the MD and EM estimators, we generated samples with $N = 100$, $N = 500$, and $N = 1000$ spatial grid points, observed for $T = 4$ time periods.[18] For each sample size, we generate 50 Monte Carlo replications. In each replication, we estimate the observed transitions for Y_{it} using frequency estimators, and run both MD and EM 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 $\boldsymbol{\Upsilon}$ 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

[18] In a typical empirical application, the points might be arranged in a 100-by-100 grid. Our assumption of spatial independence allows us to treat each point separately, simplifying the empirical analysis.

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 EM estimators (on the columns).

As expected, the performances of the MD and EM 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 EM often dominates the MD estimator by having smaller biases. Also, especially for smaller sample sizes, the EM has much smaller standard deviations than the MD estimator. This is not surprising given that the EM estimator is a maximum likelihood estimator, and so it is efficient. This can be seen graphically in Figure 2, where we show the distribution across replications of the estimated transition probabilities and misclassification probabilities for the first unobserved state, $s = 1$, using box and whisker plots. The true parameter value is marked by a dotted line. 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 EM) 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 EM takes longer to converge (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 constrast, 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.[19] While the additional time periods require

[19] In other words, we set

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

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 $N = 1000$ observations and $T = 4, 5$, and 6 time periods. As these graphs show, the results are similar across the different number of time periods.

Two remarks about the performances and practical implementations are in order.

Remark 1. The MD estimator may suffer from an additional difficulty (though this one is relatively less severe): if the observed joint distribution of Y_t and Y_{t-1} ever includes a row or column of zeros (which is more likely at smaller sample sizes, especially if $\mathcal{S}$ is large and includes rare land uses), the matrix $\widehat{\mathbf{M}}_{Y_t,Y_{t-1}}$ will be singular, and the minimum distance objective function as stated in equation (11) will be undefined. In this case, we would have to modify the definition of g_{1yt} so that $\widehat{\mathbf{M}}_{Y_t,Y_{t-1}}$ no longer needs to be inverted; this would require computing $\boldsymbol{\Upsilon}^{-1}$ on each objective function evaluation. This modification can be problematic in applications where the misclassification probability matrix is close to singular, but that will not be the case if Y_{it} is the output of an accurate classifier.

Remark 2. The EM has an additional advantage over the MD if Y_{it} is missing at random (e.g. due to cloud cover): in this case, the EM algorithm can make full use of the entire panel, whereas the frequency estimators $\widehat{\mathbf{M}}_{y_{t+1},Y_t,Y_{t-1}}$ have to be computed from the subset where Y_{it+1}, Y_{it}, and Y_{it-1} are all non-missing. If we are in a setting where, for example, Y_{it} is missing at random with probability, say, 0.10, we will have to throw out approximately $1 - 0.9^3 = 0.271$ of our data to compute $\widehat{\mathbf{M}}_{y_{t+1},Y_t,Y_{t-1}}$ for the MD estimator. (A similar statement applies to $\widehat{\mathbf{M}}_{Y_t,Y_{t-1}}$.)

These considerations suggest combining the MD and EM 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 (more efficient) EM estimator.

5.3 Varying Parameter Configurations

We now fix the sample size at $N = 1000$ 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 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 estimator predicts many more transitions than actually occur. In contrast, the EM and the MD 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 misclasification rate, the estimates are more similar for the EM and MD approaches, but the EM 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 EM and MD 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 a unique validation data from the Brazilian Agricultural Research Corporation (Embrapa).[20] The data set contain information on land use at 409 spatial points observed 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 $(59^o25'14''W, 14^o2'39''S)$ [lower left] to $(54^o25'19''W, 11^o42'16''S)$ [upper right], whithin 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).)[21]

Not just these are high quality ground-level data for the Brazilian Legal Amazon (and in general), they are especially useful for us because longitudinal ground-level data are crucial for val-

[20] We are grateful to Alexandre Camargo Coutinho and Daniel De Castro Victoria, who generously shared the dataset.

[21] 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. Their data were collected via farmer or farm manager interviews. The cropping practices were recorded for each individual sites and integrated into a Gegraphic 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).

idating 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 $\Pr[Y_{it}|S_{it}]$. No less 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 for any given point i. The Embrapa data set 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 here, $\mathcal{S} = \{\text{crops}, \text{pasture}\}$. 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).

To implement a frequency estimator for land use transitions, we trained a machine-learning classifier using remote-sensing data. Specifically, the remote-sensing data is from the sixteen-day composite Terra MODIS 250m.[22] MODIS data provide information of five variables that we used 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 – they correspond to our vector of variables R_{it}, and that are presented in Figure 1. We use them to predict land use $Y_{it} = f(R_{it})$ in each year. We merge the MODIS data with the Embrapa ground-level data, considering the September-to-August harvest years for consistency. In this way, 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 a 50% training set and 50% test set. (The training set contains 201 cross-sectional points, and the test set, 202 points.) Using the training data set, we predict the land cover based on the gradient boosting method (GBM) (Hastie et al., 2009, Chapter 10). The GBM is a machine learning technique for classification problems, based on multiple decision trees learning algorithms, and that allows for an arbitrary differentiable loss function, in order to improve predictions from single decision trees.[23]

[22] 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).

[23] In our data, we select the number of trees in the GBM by cross-validation, which is a standard technique for selecting tuning parameters for machine learning models. (We impose a maximum of 8000 trees allowed).

Once we obtain the best fit of the GBM algorithm in the training data, we generate out-of-sample predictions for the validation data (the "test" data set). The out-of sample performance of our GBM classifications is shown in Table 2. This table presents the so-called "confusion matrix," which counts the number of correctly and incorrectly classified points. From these points we can calculate the matrix $\boldsymbol{\Upsilon}$, with the misclassification probabilities.

The GBM algorithm has a high accuracy rate: Overall, it correctly predicts land use in 98% of the points. It predicts crops almost perfectly, with probability 99.9% (i.e., $\Pr[Y_{it} = \textit{crops} \mid S_{it} = \textit{crops}] = 0.999$), while it predicts pasture less precisely, but still with a high probability of 78.5% (i.e., $\Pr[Y_{it} = \textit{pasture} \mid S_{it} = \textit{pasture}] = 0.785$).

Next, we implement the HMM method using the training data. Recall that the HMM approach does not require the use of the validation data to correct the transitions (though that would be possible in this particular context). Given that correct classification probabilities based on Table 2 are all well above 50%, the validation data suggests that $\boldsymbol{\Upsilon}$ is indeed diagonally dominant, which implies that Condition 3 presented in Section 3.2.1 holds – this increases our confidence that the HMM is identified.

An important practical limitation here is that the number of points in pasture is small (see the "Total" column in Table 2). To handle this issue of (relative) small data, we fit a model with time-invariant transition probabilities. This reduces the total number of parameters to be estimated from 11 to 5, and should increase the estimation precision.[24]

We now compare the true (time-invariant) transition probabilities from Embrapa ground-level data, the frequency estimator (based on our GBM predictions Y_{it}), and the HMM correction. The ground-level true transition is

$$\mathbf{P}_{S_{t+1}|S_t} = \begin{pmatrix} 0.987 & 0.013 \\ 0.234 & 0.766 \end{pmatrix}.$$

This means that the probability that a point i with cropland at year t continues to be cropland in the next year $t+1$ is 98.7%; while the probability of switching to pasture equals 1.3%. The probability of switching from pasture to crops is higher: 23.4%; and the probability of maintaining pasture land is 76.6%.

The frequency estimator generates the following transition:

$$\widehat{\mathbf{P}}_{Y_{t+1}|Y_t} = \begin{pmatrix} 0.98 & 0.02 \\ 0.393 & 0.607 \end{pmatrix}.$$

[24] In order to increase precision further, we estimate the HMM using the panel of test points from 2006 to 2016, i.e. for more years than we observe S_{it}. Implicitly, we are assuming that the Embrapa land uses S_{it} follow a first-order Markov chain whose transition probabilities remained fixed from 2006 to 2016.

Similarly to the Monte Carlo simulations presented in Section 5, land use misclassification results in higher probabilities on the off-diagonal elements, indicating excessive transitions in land use. Transitions from crops to pasture here is 2%, instead of 1.3%; and transitions in the opposite direction, from pasture to crops, is estimated to be 39.3%, instead of 23.4%. These transitions are, respectively, 54% and 68% higher than the ground-level true transitions.

Next, we discuss the HMM corrections. We focus on the EM estimator because the Monte Carlo simulations indicate that it performs better in finite (small) samples than the MD estimator. Indeed, the MD parameter estimates for the transitions are exactly at the edges of $[0, 1]$. This is consistent with the simulations in Section 5, where we learned that the minimum distance estimates are more likely than the EM estimates to hit the boundary of the parameter space, particularly at small sample sizes.

The EM estimates for the transitions are

$$\widehat{\mathbf{P}}^{\text{EM}}_{S_{t+1}|S_t} = \begin{pmatrix} 0.988 & 0.012 \\ 0.197 & 0.803 \end{pmatrix}$$

Clearly, the EM estimates are substantially closer to the true ground-level transitions than the frequency estimates.

Finally, the HMM estimates of the misclassification probabilities are

$$\widehat{\boldsymbol{\Upsilon}}^{\text{EM}} = \begin{pmatrix} 0.999 & 0.22 \\ 0.001 & 0.78 \end{pmatrix}.$$

These estimates are close to the true misclassifications obtained from the out-of-sample predictions based on the GBM classifications. (See the last column in Table 2.) As a matter of fact, they agree almost exactly, which is remarkable given that the HMM is calculated using only the (training) panel data with observed classifications $\{Y_{it}\}$, and with no information on S_{it}.

The Embrapa dataset is large relative to typical validation data, especially for the Legal Amazon. Yet, for the purposes of fitting an HMM and computing transition probabilities, it is relatively small – particularly after splitting it into training and test sets, and taking into account that pasture is relatively rare. In order to get a sense of the variability in the estimates, we take bootstrap samples from the test set, and refit the HMM on each sample. This produces a collection of transitions probabilities $\mathbf{P}^b_{S_{t+1}|S_t}, \widehat{\mathbf{P}}^b_{Y_{t+1}|Y_t}, \widehat{\mathbf{P}}^{b,\text{EM}}_{S_{t+1}|S_t}$, where b indexes the bootstrap sample, for $b = 1, ..., 100$. In the interest of speed, we hold the GBM training set fixed, i.e. the model that produces Y_{it} is held fixed throughout this exercise.

The results of the bootstrap exercise are shown in Figures 6 and 7. In the horizontal axis we

show the transition probabilities in the bootstrapped test data, and in the vertical axis we show the corresponding estimated transition probabilities. Each point corresponds to a single bootstrap sample. The small number of pasture observations in the test panel is reflected by the considerable variability in $\Pr[S_{it+1} = \text{pasture} \mid S_{it} = \text{pasture}]$. Nonetheless, we see that the frequency estimates are on average too low, whereas the EM estimates appear to be much closer to the true transitions.

7 Conclusion

Satellite-based data allow researchers to access an unprecedent 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 contribute to the applied research by drawing from the econometrics literature on measurement errors and proposing methods to estimate transition process, such as land use changes, from remote-sensing data. Based on hidden Markov models, we show how to obtain estimates of transition proabababilities that account for misclassification, as opposed to traditional methods that either ignore the problem altogether, propose ad-hoc solutions, or yet require (prohibitively) costly validation data. The method that we propose avoid these issues since it 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 it does not require ground-level truth validation data. We propose two estimators based on the HMM framework, and investigate their performances in Monte Carlo simulation studies, as well as in a validation study based on a high-quality ground-level land cover data in a region of interest inside the Brazilian Legal Amazon that has been experiencing rapid land use changes from agricultural expansion.

We consider two promising and interesting topics for future research: expanding the HMM correction to allow for spatial correlation in the grid points, and incorporating richer sets of covariates when estimating transition probabilities, possibly based on semiparametric specifications.

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.

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.

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.

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.

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

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.

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

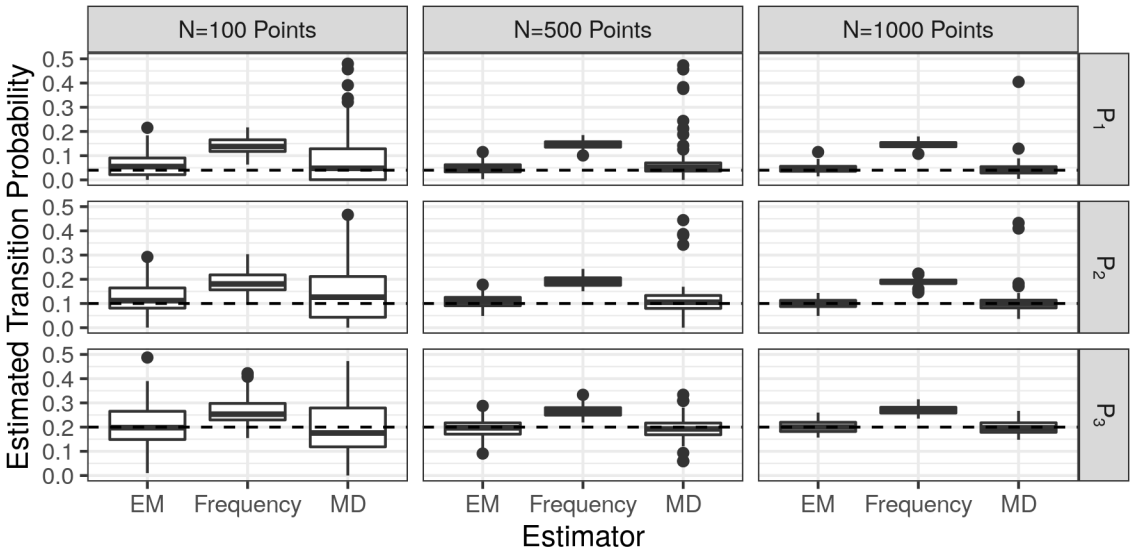
van Handel, R. (2008). Hidden Markov Models: Lecture notes. `https://www.princeton.edu/~rvan/orf557/hmm080728.pdf`. [Online; accessed 2017-06-06].

Table 1: Baseline Monte Carlo Simulation Results

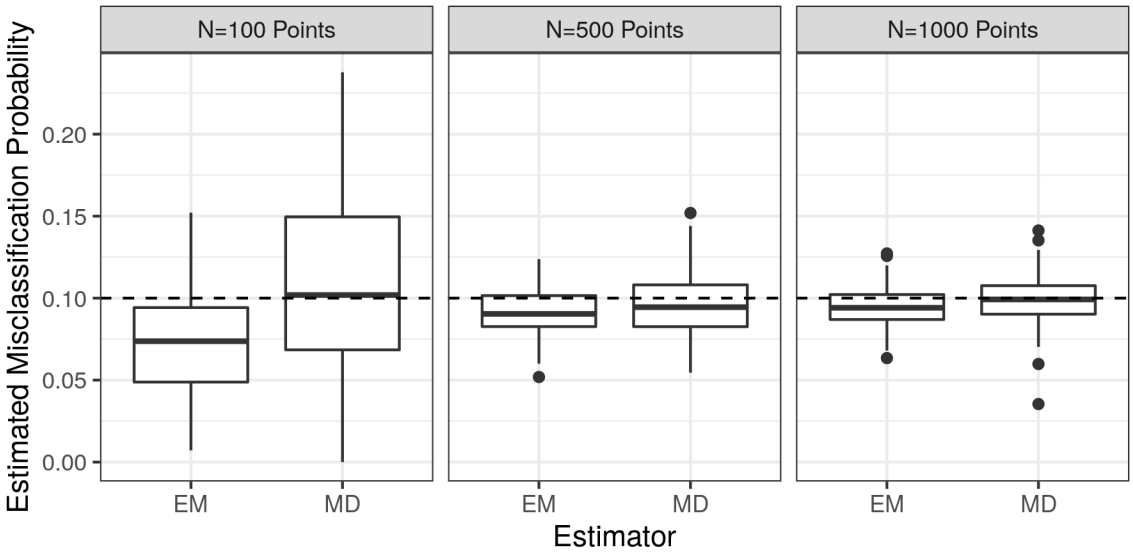
		N=100			N=500			N=1000		
		Freq	MD	EM	Freq	MD	EM	Freq	MD	EM
	Bias	-0.071	-0.077	-0.033	-0.068	-0.029	-0.009	-0.072	-0.009	-0.008
$P_{S_1} = .9$	s.d.	0.038	0.168	0.059	0.015	0.076	0.025	0.012	0.037	0.021
	RMSE	0.080	0.184	0.067	0.070	0.081	0.027	0.073	0.038	0.022
	Bias		0.010	-0.025		-0.004	-0.008		-0.001	-0.004
$\Upsilon(2,1) = .1$	s.d.		0.058	0.033		0.018	0.013		0.015	0.012
	RMSE		0.058	0.041		0.019	0.016		0.015	0.012
	Bias		0.292	-0.035		0.065	-0.018		-0.001	-0.022
$\Upsilon(1,2) = .2$	s.d.		0.333	0.115		0.196	0.058		0.096	0.042
	RMSE		0.442	0.119		0.206	0.061		0.096	0.047
	Bias	0.102	0.086	0.023	0.106	0.031	0.009	0.105	0.006	0.005
$P_1(1,2) = .04$	s.d.	0.036	0.194	0.048	0.017	0.082	0.022	0.013	0.041	0.016
	RMSE	0.108	0.212	0.053	0.107	0.088	0.024	0.106	0.042	0.017
	Bias	0.552	0.229	0.195	0.542	0.065	0.086	0.540	0.065	0.082
$P_1(2,1) = .02$	s.d.	0.123	0.274	0.196	0.057	0.122	0.114	0.042	0.104	0.088
	RMSE	0.565	0.356	0.276	0.545	0.138	0.143	0.542	0.123	0.120
	Bias	0.090	0.212	0.022	0.091	0.049	0.008	0.089	0.004	0.001
$P_2(1,2) = .1$	s.d.	0.047	0.341	0.070	0.021	0.161	0.026	0.015	0.051	0.020
	RMSE	0.101	0.401	0.073	0.093	0.168	0.027	0.090	0.051	0.020
	Bias	0.464	0.138	0.086	0.475	0.046	0.048	0.469	0.038	0.035
$P_2(2,1) = .02$	s.d.	0.120	0.233	0.140	0.055	0.080	0.076	0.036	0.065	0.052
	RMSE	0.479	0.270	0.163	0.478	0.092	0.089	0.471	0.075	0.063
	Bias	0.066	0.246	0.010	0.065	0.081	-0.003	0.071	0.014	0.002
$P_3(1,2) = .2$	s.d.	0.058	0.387	0.083	0.024	0.251	0.039	0.016	0.116	0.025
	RMSE	0.088	0.456	0.083	0.070	0.262	0.038	0.073	0.116	0.025
	Bias	0.355	0.099	0.091	0.363	0.034	0.047	0.363	0.027	0.038
$P_3(2,1) = .02$	s.d.	0.105	0.233	0.131	0.040	0.073	0.062	0.032	0.057	0.042
	RMSE	0.370	0.252	0.159	0.365	0.080	0.078	0.364	0.063	0.056

Table 2: Confusion Matrix based on Embrapa Validation Data

	GBM Classification (Y_{it})			*Percentage of*
Embrapa Data (S_{it})	Crops	Pasture	**Total**	*Correctly Predicted*
Crops	857	1	858	0.999
Pasture	17	62	79	0.785
Total	874	63	937	

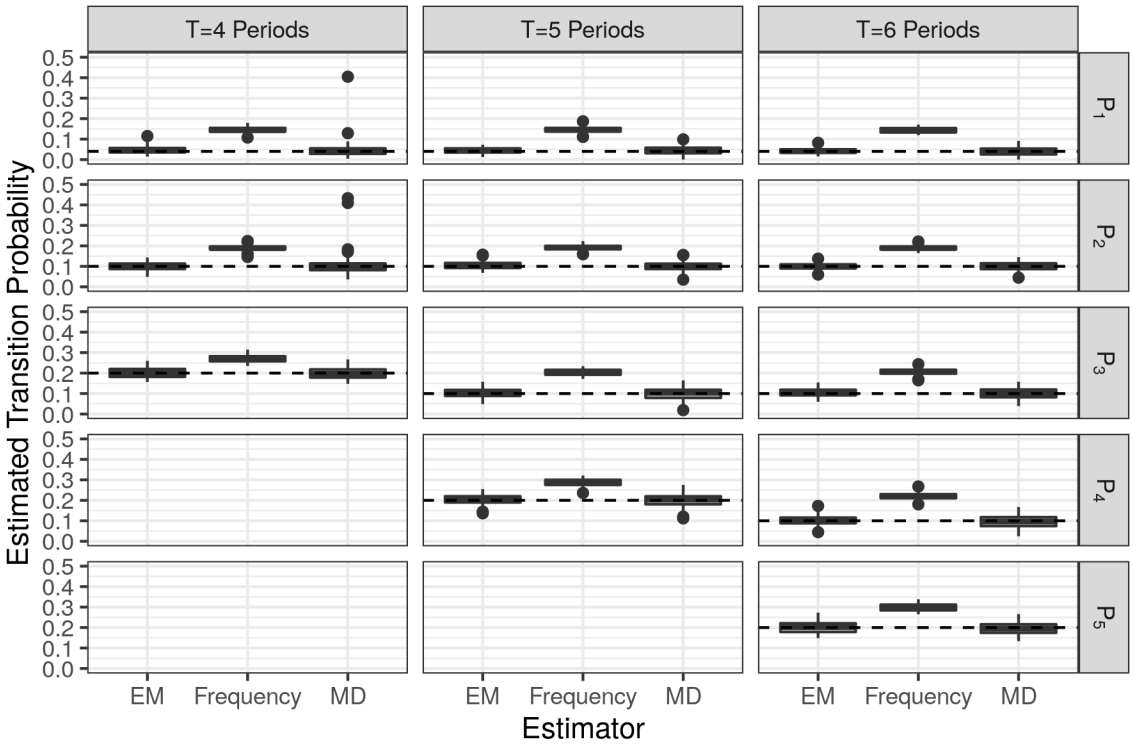


(a) Transition Probability

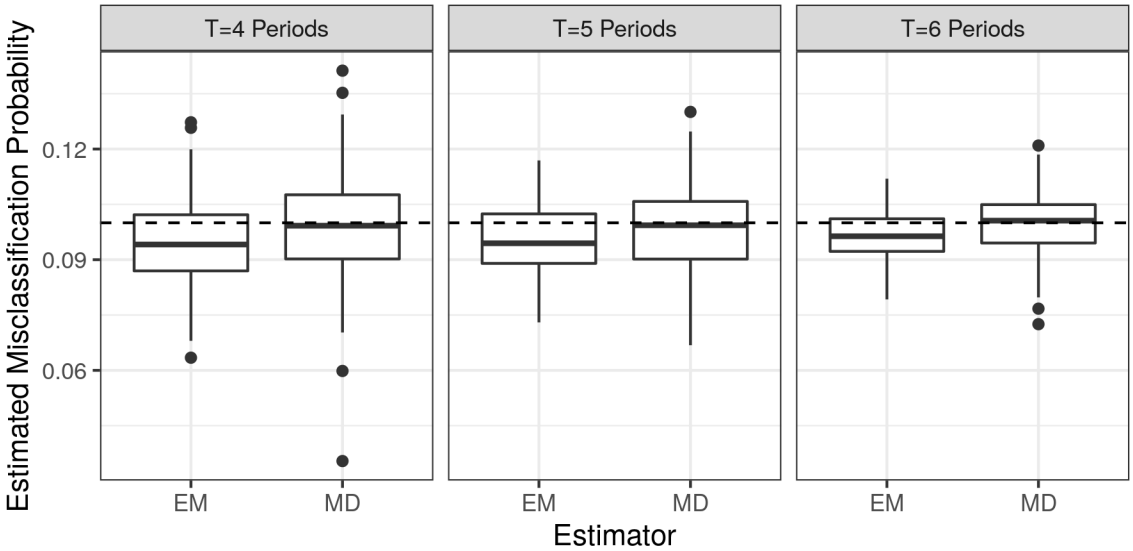


(b) Misclassification Probability

Figure 2: Baseline Monte Carlo Simulation Results (Unobserved State 1)

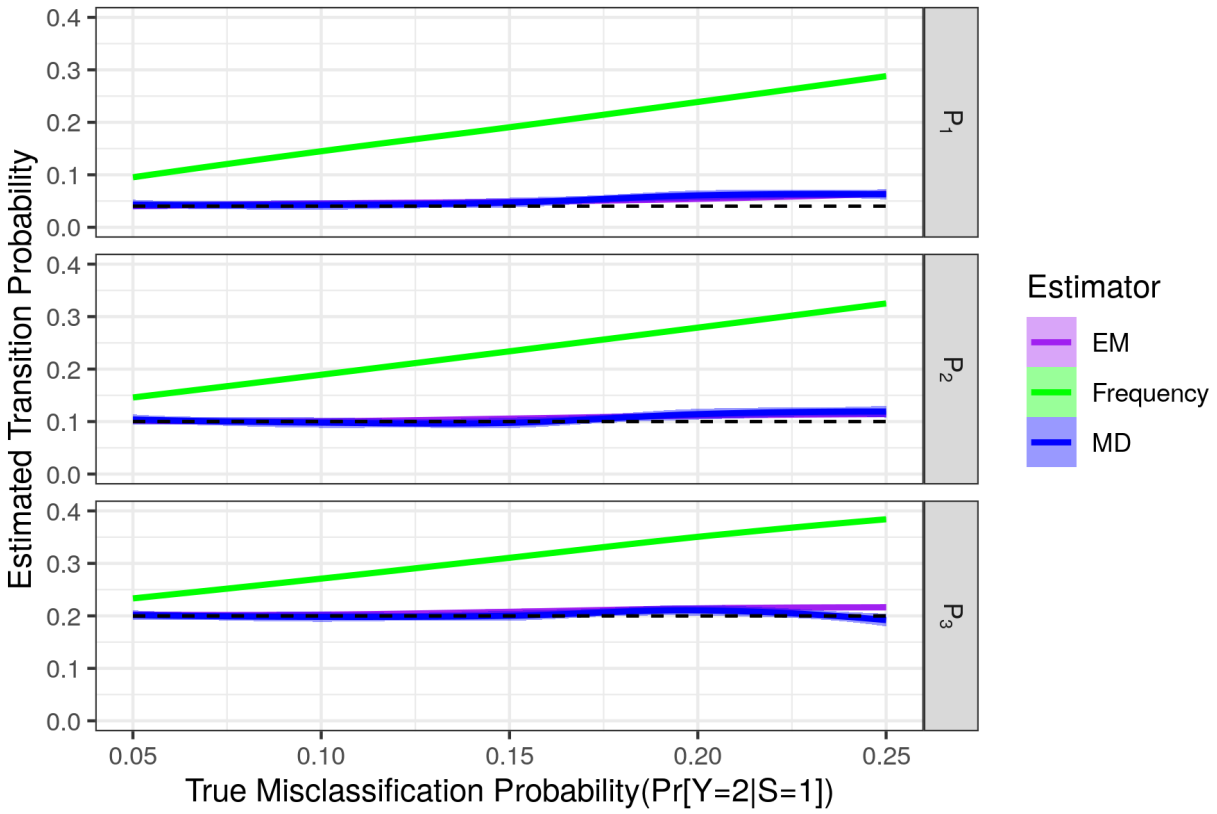


(a) Transition Probability

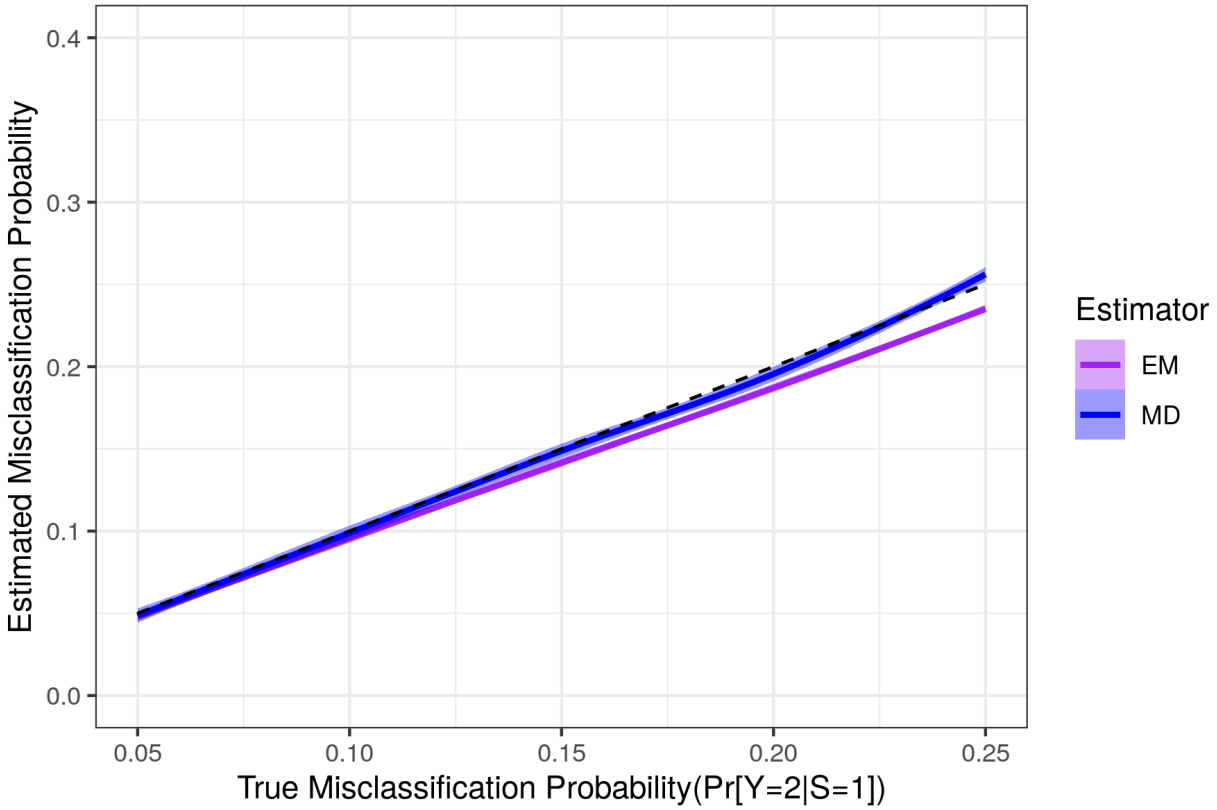


(b) Misclassification Probability

Figure 3: Baseline Monte Carlo Simulation Results (Unobserved State 1

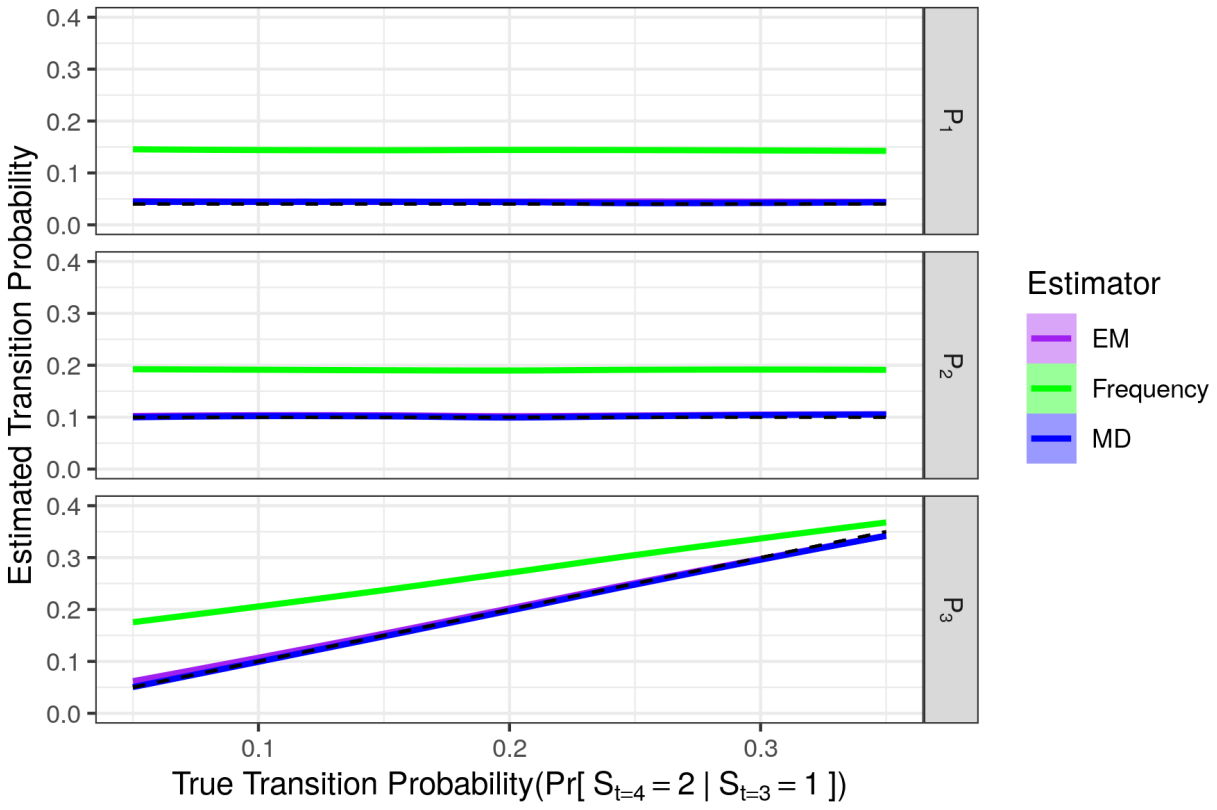


(a) Transition Probability

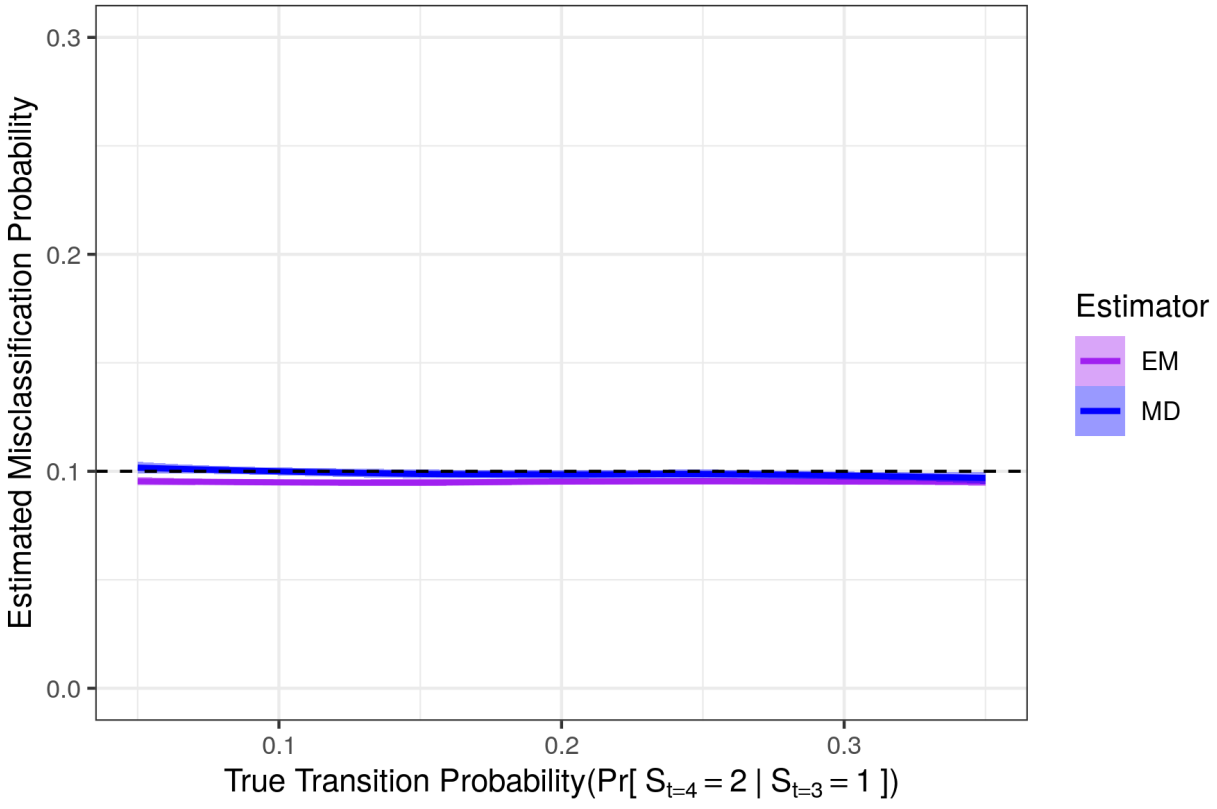


(b) Misclassification Probability

Figure 4: Monte Carlo Results for Varying Misclassification Probabilities



(a) Transition Probability



(b) Misclassification Probability

Figure 5: Monte Carlo Results for Varying Transition Probabilities

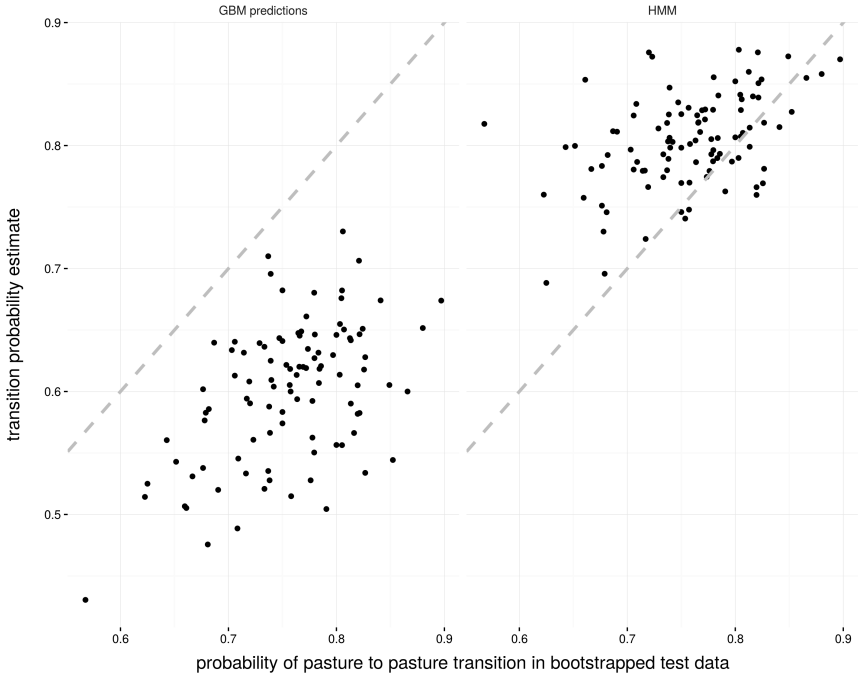


Figure 6: Validation: HMM and GBM Estimates of Pasture Transition Probability

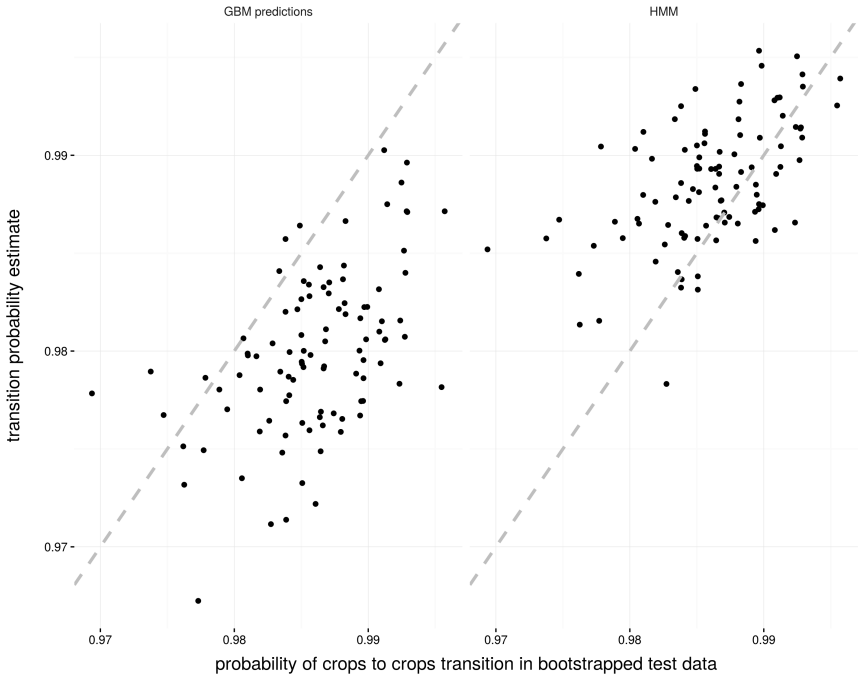


Figure 7: Validation: HMM and GBM Estimates of Crop Transition Probability

A Appendix: Mathematical Derivation of Useful Identities

Under the HMM assumptions, the joint distribution of (Y_{it}, Y_{it-1}) satisfies

$$\Pr\left[Y_{it}, Y_{it-1}\right] = \sum_{s\in\mathcal{S}} \Pr\left[S_{it} = s, Y_{it-1}\right] \Pr\left[Y_{it}|S_{it} = s\right]. \tag{13}$$

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

$$\begin{aligned}\Pr\left[Y_{it+1}, Y_{it}\right] &= \sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \Pr\left[Y_{it+1}|S_{it+1} = s'\right] \Pr\left[S_{it+1} = s'|S_{it} = s\right] \\ &\quad \times \Pr\left[Y_{it}|S_{it} = s\right] \Pr\left[S_{it} = s\right] \\ &= \sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \Pr\left[Y_{it+1}|S_{it+1} = s'\right] \Pr\left[S_{it+1} = s', S_{it} = s\right] \Pr\left[Y_{it}|S_{it} = s\right]. \end{aligned} \tag{14}$$

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

$$\Pr\left[Y_{it+1}, Y_{it}, Y_{it-1}\right] = \sum_{s\in\mathcal{S}} \Pr\left[Y_{it+1}|S_{it} = s\right] \Pr\left[Y_{it}|S_{it} = s\right] \Pr\left[Y_{it-1}, S_{it} = s\right], \tag{15}$$

because

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

In matrix notation, equations (13)–(15) 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 θ represent the collection of HMM parameters, i.e. θ is a list containing $\Pr\left[S_{i1}\right]$, $\Pr\left[S_{it+1}|S_{it}\right]$, and $\Pr\left[Y_{it}|S_{it}\right]$, for all $t =$

$1, ..., T-1$. 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

$$l(\theta) \equiv \ln \Pr[Y = y; \theta] \tag{16}$$

and let

$$J(\theta, \theta') \equiv \sum_s \Pr\left[S = s | Y = y; \theta'\right] \ln \left\{ \frac{\Pr[Y = y, S = s; \theta\ ]}{\Pr[Y = y, S = s; \theta']} \right\}. \tag{17}$$

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 $T \geq t$. 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 θ are

$$\begin{aligned} \mu^{(j+1)}[k] &= \frac{\sum_i \pi_{i1}[k]}{\sum_{i,s} \pi_{i1}[s]} \\ \mathbf{P}_t^{(j+1)}[k,l] &= \frac{\sum_i \pi_{it}[k,l]}{\sum_i \pi_{it}[k]} \\ \boldsymbol{\Upsilon}^{(j+1)}[y,k] &= \frac{\sum_{i,t:Y_{it}=y} \pi_{it}[k]}{\sum_{i,t} \pi_{it}[k]} \end{aligned} \tag{18}$$

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.

C Online Appendix: A Policy-Relevant Illustrative Example

In this Appendix, we consider a policy-relevant illustration of how misclassification probabilities can affect conservation policies, and lead to inefficient outcomes.

Consider the problem faced by a regulator in a country with two regions of the same size, A and B, each of which is covered by land whose state is in $\mathcal{S} = \{\text{forest}, \text{deforested}\}$. Imagine that the regulator has sufficient resources to protect only a fixed area of forest (e.g. by increasing enforcement in a 500 km^2 area), with the goal of preventing as much deforestation as possible: she can declare a new protected area either in region A or in region B (but not both), and she would therefore like to identify the region with the larger deforestation rate.[25] Obtaining ground truth panel data $\{S_{it}^A\}$ and $\{S_{it}^B\}$ would be prohibitively expensive; instead, the regulator has access only to noisy land use classifications $\{Y_{it}^A\}$ and $\{Y_{it}^B\}$ based on remotely-sensed data.

Suppose Y_{it}, S_{it} satisfy the hidden Markov model assumptions presented in the main text, with

[25] We take as given that the regulator's action is effective, and moreover that the amount of deforestation it prevents is increasing in the deforestation probability.

the following time-homogeneous transition probabilities and misclassification probabilities:[26]

$$\mathbf{P}_A = \begin{pmatrix} 0.95 & 0.05 \\ 0.01 & 0.99 \end{pmatrix} \quad \mathbf{P}_B = \begin{pmatrix} 0.9 & 0.1 \\ 0.01 & 0.99 \end{pmatrix}$$
$$\boldsymbol{\Upsilon}_A = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \quad \boldsymbol{\Upsilon}_B = \begin{pmatrix} 0.98 & 0.02 \\ 0.02 & 0.98 \end{pmatrix}.$$

The deforestation probability $\Pr[S_{it+1} = \text{deforested} \mid S_{it} = \text{forest}]$ is 0.05 in region A and 0.1 in region B. Given the large differences in the deforestation rates, the regulator's problem might appear straightforward: It is better to protect region B to conserve a larger area of forest.

Notice, however, that the misclassification probabilities differ across the two regions, with region A having a lower deforestation probability but a larger misclassification rate. Misclassifications cause observed transition probabilities to be erroneously high, and, in this case, the difference between $\boldsymbol{\Upsilon}_A$ and $\boldsymbol{\Upsilon}_B$ is large enough to push the mismeasured deforestation rate for region A above the observed deforestation rate for B. In a simulation using panels of 10000 points per region (imagine a 100-by-100 raster of spatial points), each of which is observed for 3 years (with an initial distribution $\Pr[S_{i1} = \text{forest}] = 0.7$ in both regions), the frequency estimates obtained from the transitions in Y_{it} are

$$\widehat{\mathbf{P}}_A^{\text{Freq}} = \begin{pmatrix} 0.82 & 0.18 \\ 0.25 & 0.75 \end{pmatrix} \quad \widehat{\mathbf{P}}_B^{\text{Freq}} = \begin{pmatrix} 0.874 & 0.126 \\ 0.059 & 0.941 \end{pmatrix}.$$

Clearly, based on these estimates, the regulator would decide to protect area A, instead of B.

By contrast, the EM estimates of the transition probabilities are highly accurate and produce the correct ordering between the two regions, with region A's estimated deforestation rate below that for region B:

$$\widehat{\mathbf{P}}_A^{\text{EM}} = \begin{pmatrix} 0.943 & 0.057 \\ 0.019 & 0.981 \end{pmatrix} \quad \widehat{\mathbf{P}}_B^{\text{EM}} = \begin{pmatrix} 0.9 & 0.1 \\ 0.002 & 0.998 \end{pmatrix}.$$

In this example, a regulator following the frequency estimator approach to transition probability estimation would focus enforcement resources on the wrong region, whereas a regulator using the HMM strategy would reach the correct decision, and would therefore prevent twice as much deforestation.

In the more general setting where the initial distribution over S_{it} varies by region, an approach

[26] In order to keep the presentation straightforward, we drop the time subscript and assume that transition probabilities are time-homogeneous within each region.

based on looking just at the marginal distribution of Y_{it} could easily fail: one region might have a very large forest share but a low deforestation probability; another region might have a lower forest share but a much higher deforestation probability. If the regulator has only enough resources to protect a fixed area of forest (smaller than the total forest area in either region), she would want to focus enforcement on the region whose current deforestation probability is largest: this could be handled by an HMM with time-varying deforestation probabilities, whereas focusing enforcement on the region with the largest decrease in $\Pr[Y_{it} = \text{forest}]$ could lead to the wrong decision.

Finally, note that in problems involving N regions, if misclassification probabilities vary spatially, a correction based on validation data would require N separate ground truth datasets, which may be extremely expensive and time-consuming, whereas the HMM approach could be applied using a single classifier trained on a single training set, with $\boldsymbol{\Upsilon}$ allowed to vary by region, in order to inform the regulator.