Supplemental Material for "Improving Estimates of Transitions from Satellite Data: A Hidden Markov Model Approach"

Adrian L. Torchiana, Ted Rosenbaum,

Paul T. Scott, and Eduardo Souza-Rodrigues*

December 26, 2022

This Supplemental Material consists of the following sections: Section A presents relevant mathematical derivations for the identification of the hidden Markov model (HMM). Section B discusses the measurement error in observed transition probabilities that is implied by the HMM model, and its consequences for regression analyses. Section C presents the details of the expectation-maximization (EM) and Viterbi algorithms, used respectively for the maximum likelihood estimation and for the computation of the most likely trajectory of land uses for each pixel in the data. Section D shows the Monte Carlo simulation studies. Section E provides additional details on the carbon stock empirical application.

*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).

A Mathematical Derivation of Useful Identities

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

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

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

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

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

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

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

because

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

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

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

B Measurement Error under the HMM Assumptions

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

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

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

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

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

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

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

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

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

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

where the second equality holds by the Bayes rule. Note again that we re-incorporate the subscripts m and $t+1$ explicitly in the definition of β_{mt+1}, as this term may vary over regions and time periods. The terms on the right-hand side of (B6) should be understood to condition on pixels in region m.

Assuming the classifier is accurate (in the sense that correct classification is more likely than misclassification), the first term on the right-hand side of (B6) is positive, which implies that β_{mt+1} is between zero and one. Clearly, β_{mt+1} depends on the misclassification probabilities and on the ratio of the shares of true and observed forested areas in region m in the previous period t. In general, the greater the percentage of correct classifications, and the higher the share of true forested areas relative to the share of observed forested areas, the greater the β_{mt+1}.

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

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

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

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

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

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

$$D_{mt} = \alpha + \beta D^*_{mt} + V_{mt}, \tag{B8}$$

where

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

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

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

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

B.1 Consequences of HMM Measurement Error for Regression Models

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

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

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

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

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

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

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

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

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

If we regress the mismeasured D_{mt} on X_{mt} using OLS to estimate γ_1, we obtain biased results for two reasons. First, if X_{mt} and ξ_{mt} were uncorrelated, OLS would be unbiased for $\delta_1 = \beta\gamma_1 \neq \gamma_1$, and so it would be biased for γ_1 given that β is between zero and one when the land use classifier is accurate. That leads to an attenuation bias. (When the land use classifier is not accurate, β could be negative, reversing the sign of the estimates.) Second, because X_{mt} may correlate with the unobservable ξ_{mt}, through the term $h(.)$. As discussed previously, $h(.)$ reflects the nonclassical

nature of the measurement error in transition probabilities (i.e., it reflects the fact that D^*_{mt} and V_{mt} likely correlate). This correlation can be positive or negative; when the correlation is positive and sufficiently large, it can bias OLS upward. In sum, when transition probability is the dependent variable of interest in a linear regression model, its measurement error most likely biases (without necessarily attenuating) the OLS estimator.

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

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

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

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

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

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

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

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

However, given that $D^*_{mt} - D_{mt} = -\alpha + (1-\beta)D^*_{mt} - V_{mt}$, where $(1-\beta) > 0$, and that D^*_{mt} likely

correlates with V_{mt} (see equation (B8)), the expectation of the first-order term may be different from zero.

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

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

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

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

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

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

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

C The EM and the Viterbi Algorithms

We now briefly explain the EM and the Viterbi algorithms.

C.1 The EM Algorithm

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

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

and let

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

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

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

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

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

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

where $t \leq T$. In words, the model allows us to condition on a long sequence of noisy land use classifications at a given spatial point, and make probabilistic statements about the point's true land

use at any period in that history. This is valuable if we are interested in land cover at a specific point: the fact that we condition on the entire sequence $Y_{i1}, Y_{i2}, \ldots Y_{iT}$ can potentially improve predictions when compared to classifiers that use only contemporaneous data to predict land use. For instance, suppose we have 15 years of data at a particular spatial point, and that the land use set is $\mathcal{S} = \{\text{forest}, \text{deforested}\}$. Imagine that our land use prediction model outputs $Y_{it} = \text{forest}$ for the first 10 years, followed by deforestation for a single year, followed by four years of forest.

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

In our application, the M step of the EM algorithm has a closed-form solution. Denote the posterior probabilities by $\pi_{it}[k] \equiv \Pr[S_{it} = k | Y = y; \theta^{(j)}]$ and $\pi_{it}[k, l] \equiv \Pr[S_{it} = k, S_{it+1} =$

$l|Y = y; \theta^{(j)}]$; these can be computed in an efficient forward-backward pass over time using the Baum-Welch algorithm (i.e., the E step), and the calculations can be done in parallel across spatial points given that we are not modeling spatial dependence (i.e., not conditioning on other pixels' land uses). The updated values of θ are

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

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

C.2 The Viterbi Algorithm

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

Briefly, the probability path estimate $\Pr\left[S_1, S_2, \ldots, S_T|Y_1, Y_2, \ldots, Y_T\right]$ can be expressed in terms of initial, transition and misclassification distributions by exploiting the HMM structure and the Bayes formula. Based on such expression, the maximization problem can be solved recursively, as

the Bellman equation in dynamic optimization problems, solving for one variable only in each step (see Section 3.3 in van Handel, 2008). Notice that finding the most likely path is different from the problem of finding the most likely state in a given period $\Pr[S_t|Y_1, Y_2, \ldots, Y_T]$, which is calculated efficiently by the Baum-Welch algorithm, as noted previously.

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

D Monte Carlo Studies

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

D.1 Basic Setup

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

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

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

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

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

The misclassification probabilities are time-invariant and given by

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

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

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

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

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

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

D.2 Baseline Results

Table E1 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 E3, where we show the distribution across replications

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

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

We also verify the performance of the estimator with $T = 5$ and $T = 6$. Relative to our $T = 4$ period baseline, we fix the transition probabilities for the first and last period and set the transitions for the middle periods equal to each other.[1] While the additional time periods require the estimation

[1]Specifically, we set

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

of additional parameters, the larger number of time periods could help improve the precision of the misclassification probability estimates. In Figure E4, we replicate the results from Figure E3 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.

D.3 Varying Parameter Configurations

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

Figure E5 presents the results for when we vary the misclassification probability for state 1, $\Pr[Y_{it} = 2|S_{it} = 1]$ (i.e., $\Upsilon(2, 1)$), between 5 and 25 percent, while holding other parameters fixed. The top panel shows the behavior of the estimates of the transition probabilities $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$, and the bottom panel shows the behavior of the estimates of the misclassification probability $\Pr[Y_{it} = 2|S_{it} = 1]$. The lines are non-parametric loess regression lines with a shaded 95% confidence interval, where the data is fit from the different Monte Carlo simulations.

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

Figure E6 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 E5. These graphs show that both MD and ML estimators continue to perform well at estimating transitions and misclassification rates with no notable differences between them (aside from those discussed above).

D.4 Spatial Dependence and Serial Correlation

We now incorporate spatial dependence and serial correlation in our Monte Carlo exercises. For each specification, we run 100 simulations.[2]

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

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

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

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

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

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

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

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

$$\Upsilon = 0.5 \cdot \Pr[Y_{it} \mid S_{it}, Z_{it} = 1] + 0.5 \cdot \Pr[Y_{it} \mid S_{it}, Z_{it} = -1] = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}, \tag{D15}$$

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

We do not include Z_{it} in the data set, so when we estimate the model, we *do not* condition on Z_{it}. In this way, any spatial dependence and serial correlation in Z_{it} will be translated into spatial and serial correlation in misclassification. Note that both MD and ML estimators will estimate the (unconditional) Υ given by (D15). (Note also that had we conditioned on Z_{it} when estimating the parameters, we would return to the standard HMM model presented in the main text – but now we

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

We model spatial dependence in Z_{it} in the following way. First, we assume Z_{it} is generated according to the Ising model, which specifies a join distribution of binary random variables over the lattice in a given time period t (see, e.g., Hastie et al., 2015, Chapter 9).[3] The degree of spatial correlation is controlled by the Ising "temperature" parameter, denoted here by β. When $\beta = 0$, there is no spatial correlation; when $\beta > 0$ there is positive spatial correlation, and the higher the value that β takes, the stronger the spatial correlation. Misclassifications are spatially (but not temporally) correlated when $\beta > 0$ and Z_{it} is i.i.d. over time. In our simulations, we set either $\beta = 0$ or $\beta = 2$. (The qualitative results are similar when we consider other values for β.)

To incorporate serial correlation in misclassifications, we allow Z_{it} to correlate over time; to simplify, we assume perfect correlation (i.e., Z_{it} is time-invariant). In this way, we force misclassification probabilities to depend on past values of (Y_{it}, S_{it}).[4] Importantly, when that happens, the main HMM assumption – equation (2) – is violated and our correction is not guaranteed to work.

Results. We start discussing the results for when there is spatial correlation in the true transition process, $\Pr[S_{it+1}|S_{it}]$, but no spatial, nor serial correlation in misclassification. Figure E7 presents an example of the initial distribution, and the evolution of both the true and the observed land uses in the lattice, in panels (a) and (b), respectively. It is clear from the figure that the pixels' outcomes are spatially correlated, and that misclassifications have no spatial nor temporal dependence (given that the same probability distribution $\Pr[Y_{it}|S_{it}]$ holds everywhere and in all time periods).

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

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

Figure E8 presents the estimated results across the simulations. As expected, both MD and ML estimators are unbiased for both the transition and misclassification probabilities, given that the HMM is correctly specified at the pixel level, while the frequency estimator is severely biased. In addition, since neither the MD nor the ML estimators make use of all the possible information available (namely, the pixels' spatial correlation), their variances are larger here when compared to the i.i.d. case presented in Section D.2.

Next, we add spatial dependence in misclassification (but still no serial correlation). As mentioned previously, we incorporate the variable Z_{it} in the data generating process (but not in the data), assuming it is i.i.d. over time and fixing the "temperature" parameter to be $\beta = 2$. Figure E9 presents an example of the evolution of the true land uses (in panel (a)), the distribution of Z_{it} (in panel (b)), and the evolution of the observed land uses (in panel (c)). The evolution of true land use is similar to the previous example; the distribution of Z_{it} is highly correlated in the lattice as well, leading to spatially correlated classification errors in any given time period, as can be seen by contrasting the panels (a) and (c) of the figure.

Figure E10 presents the estimated results for this scenario. As before, both MD and ML estimators are unbiased. But now they have even higher variances as neither spatial correlation in land uses nor in misclassification are incorporated explicitly in the estimation strategy. The frequency estimator continues to be highly biased for transitions.

In the next simulations, we drop the spatial correlation in Z_{it} and make this variable constant over time. This translates into misclassifications that are serially (but not spatially) correlated. Importantly, this renders the HMM model misspecified. Figure E11 shows the evolution of the true land uses (in panel (a)), the distribution of Z_{it} (in panel (b)), and the evolution of the observed land uses (in panel (c)) for one simulated example. We observe the same patterns as in the previous case with two differences: there is no spatial correlation in classification errors, but the errors tend to persist over time. In terms of the estimated results, presented in Figure E12, the MD and ML estimators are now biased for transitions (though not substantially) and for misclassification probabilities (particularly so for $\Pr[Y_{it} = 2|S_{it} = 2]$). Yet, the frequency estimator is significantly

more biased than the HMM corrections.

Finally, we incorporate spatially dependent and time-invariant Z_{it}, imposing therefore both spatially and serially correlated misclassifications. Figure E13 shows one simulated example, and Figure E14 presents the estimated results. As expected, the MD and ML estimators are biased, given that the HMM model is misspecified, but not substantially so for the transition probabilities (though it is more biased for the misclassification probabilities). Once again, the frequency estimator shows significant biases for the estimated transition process.

D.5 Regression Analysis

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

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

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

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

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

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

In Figure E15, we show the distribution of the estimated parameters α (the baseline) and β (the "treatment effect") from 100 Monte Carlo simulations. We consider three different scenarios:

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

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

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

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

E Additional Details on the Carbon Stock Application

E.1 Distribution of Forest Age

In Figure E16, we plot the cumulative distributions of the forest age for both the raw and the HMM-based approaches. The graph illustrates that the forest age predicted by the raw data is significantly younger than that predicted by the HMM-based estimates. That is a direct result of the high deforestation and reforestation rates obtained from the raw data: a pixel is more likely to

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

E.2 Relationship Between Carbon and Forest Age

We estimate the carbon stock for a given forest age using data on the 2017 carbon stock from Englund et al. (2017) and our HMM-based estimates of forest age for each pixel. First, we show informally the relationship between the carbon stock and the age of the forest for 2017 in Figure E17. As expected, the graph shows an increase in the carbon stock as the forest age.[5]

Next, we estimate the following regression model:

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

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

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

The results are presented in Table E2. We find that the average baseline level of carbon in forests is approximately 4 tons greater than in non-forest pixels. Every additional year the pixel remains forested adds approximately 0.6 tons of carbon, on average. For forests that are over 32 years old, the average amount of carbon in a pixel is approximately 51.5 tons ($= \alpha + \beta + \delta$).

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

References

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

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

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.

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

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

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

		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 E1: Baseline Monte Carlo Simulation Results

	Carbon Stock
α	8.929***
	(0.009)
β	4.195***
	(0.039)
γ	0.628***
	(0.002)
δ	38.318***
	(0.042)
Observations	11,770,115
R^2	0.317

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

Table E2: Relationship Between Carbon Stock and Forest Age

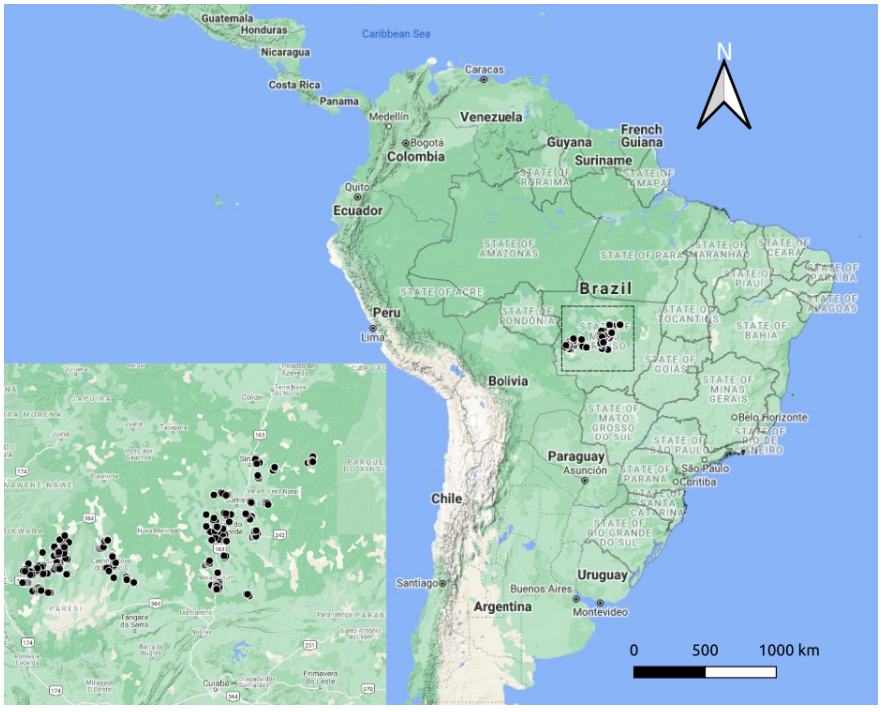
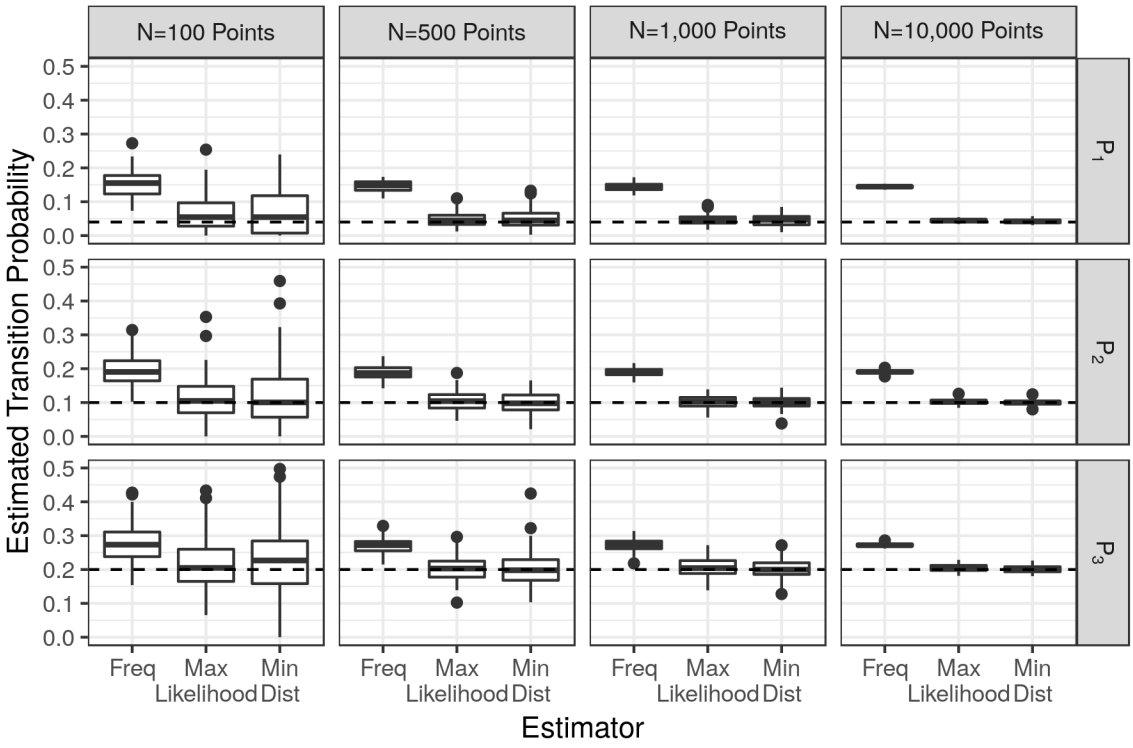
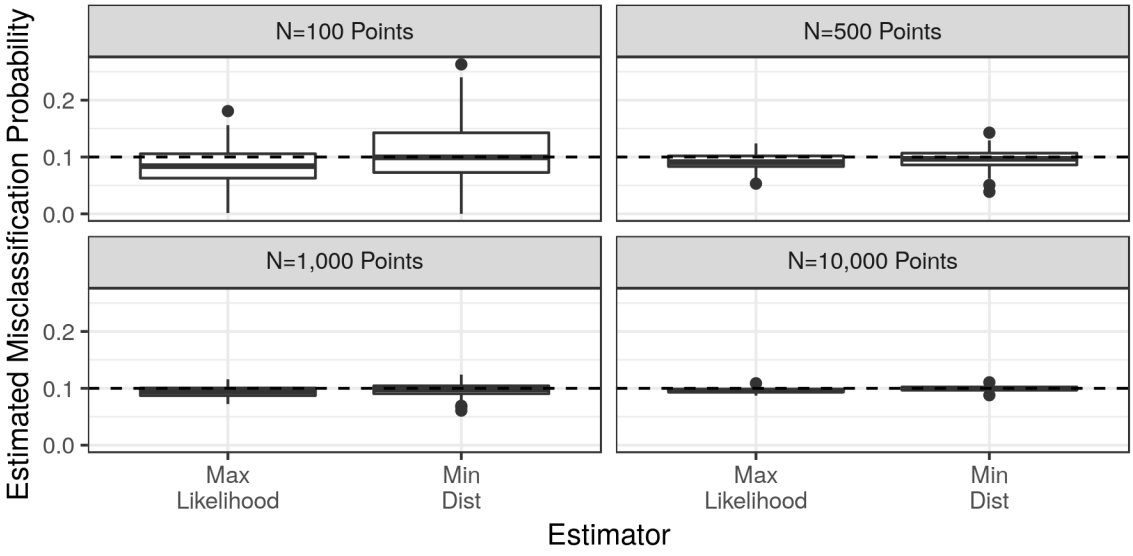


Figure E1: Map of Mato Grosso State and the Embrapa Sample Points

Figure E2: Map of Brazil and the Mapbiomas Sample Points, in the Brazilian Atlantic Forest

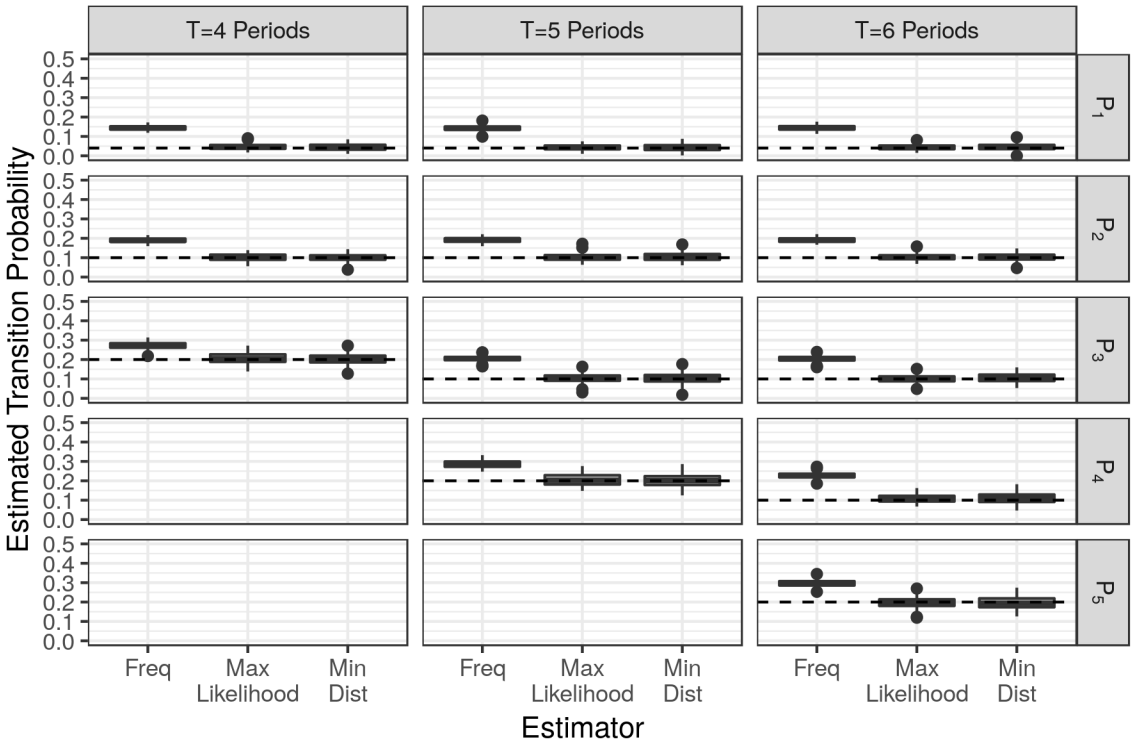


(a) Transition Probability, $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$

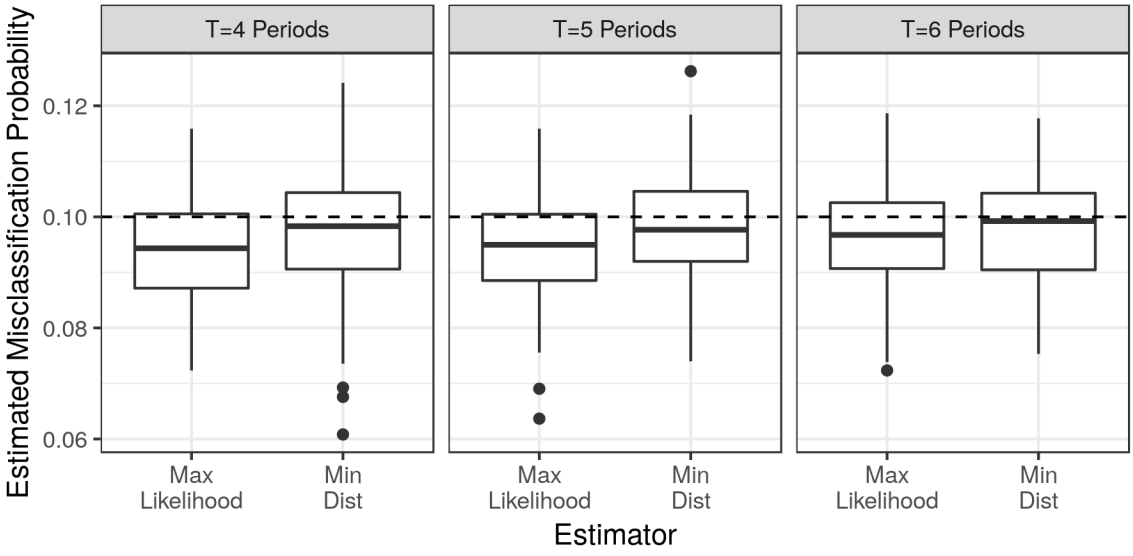


(b) Misclassification Probability, $\Pr[Y_{it} = 2|S_{it} = 1]$

Figure E3: Baseline Monte Carlo Simulation Results

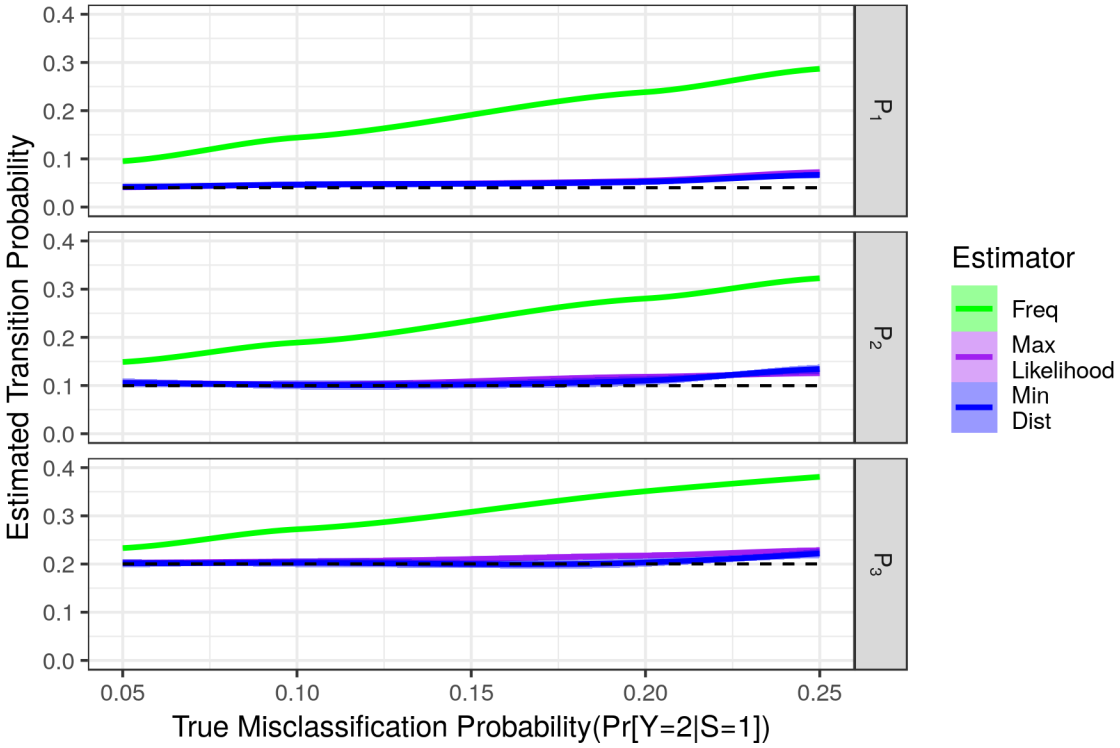


(a) Transition Probability, $\Pr[S_{it+1} = 2|S_{it} = 1]$, for $t = 1, 2, 3$

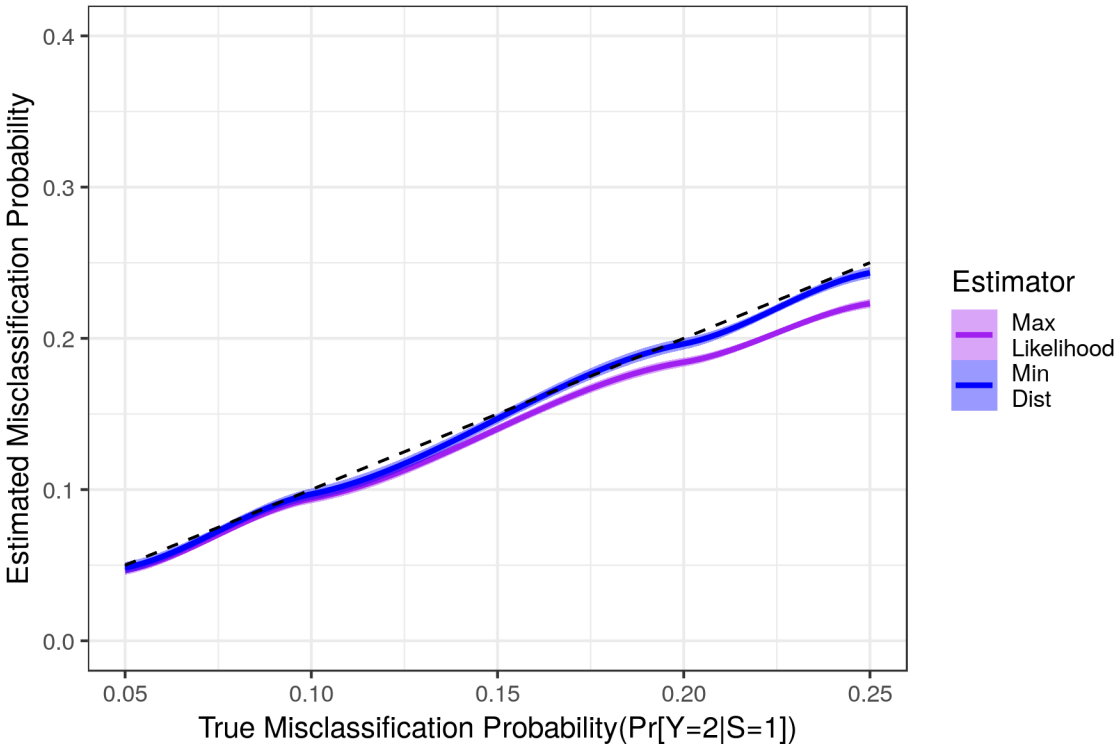


(b) Misclassification Probability, $\Pr[Y_{it} = 2|S_{it} = 1]$

Figure E4: Baseline Monte Carlo Simulation Results

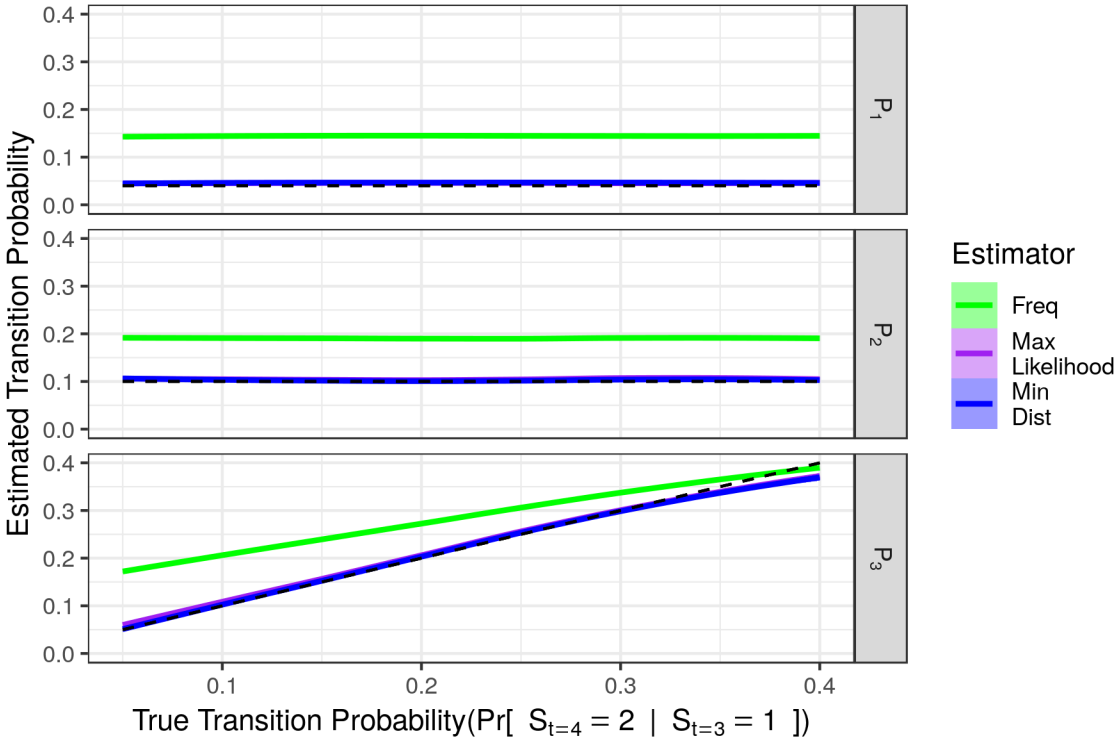


(a) Transition Probability

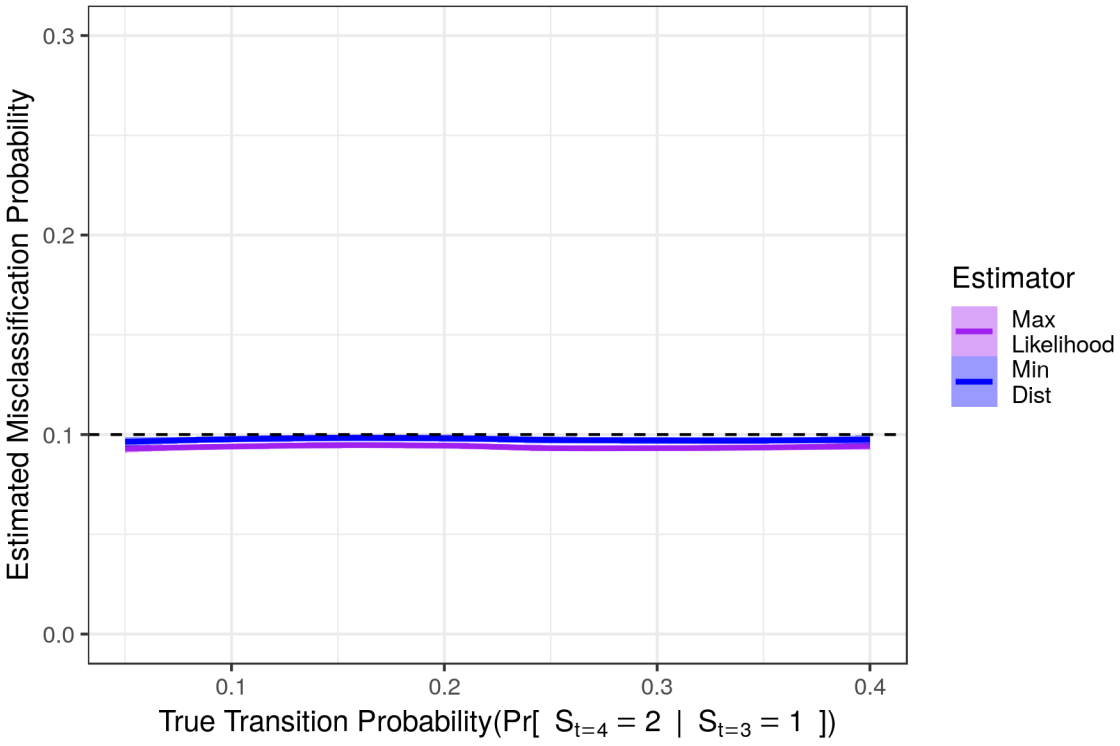


(b) Misclassification Probability

Figure E5: Monte Carlo Results for Varying Misclassification Probabilities

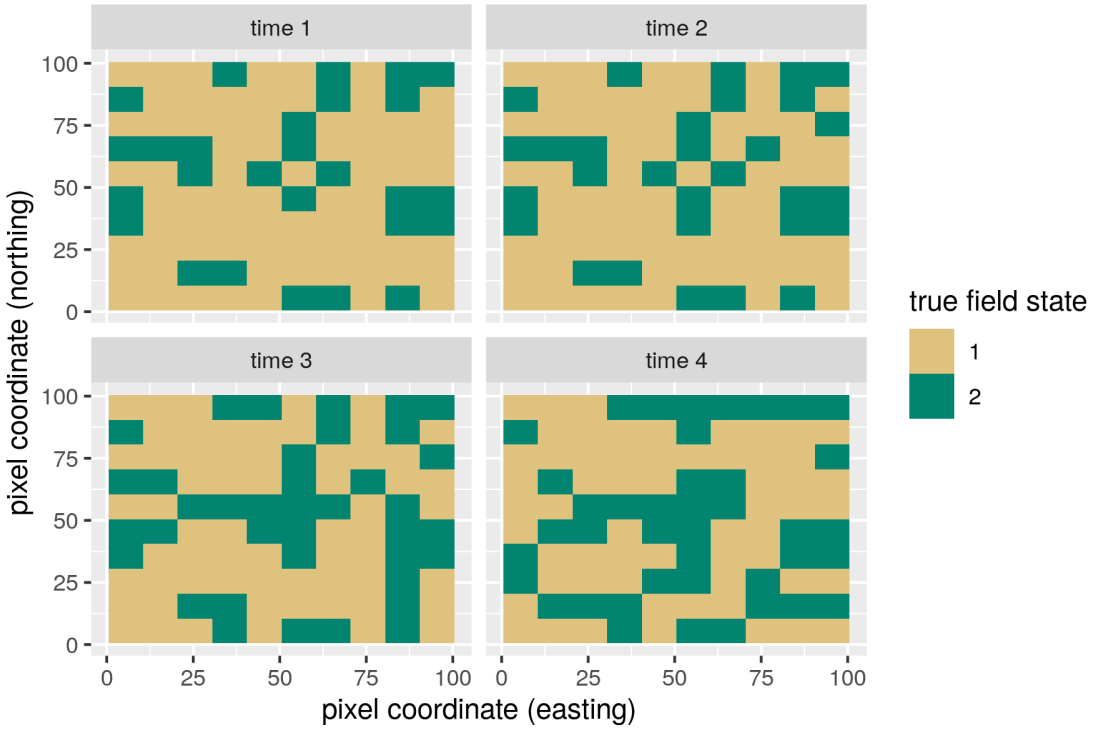


(a) Transition Probability

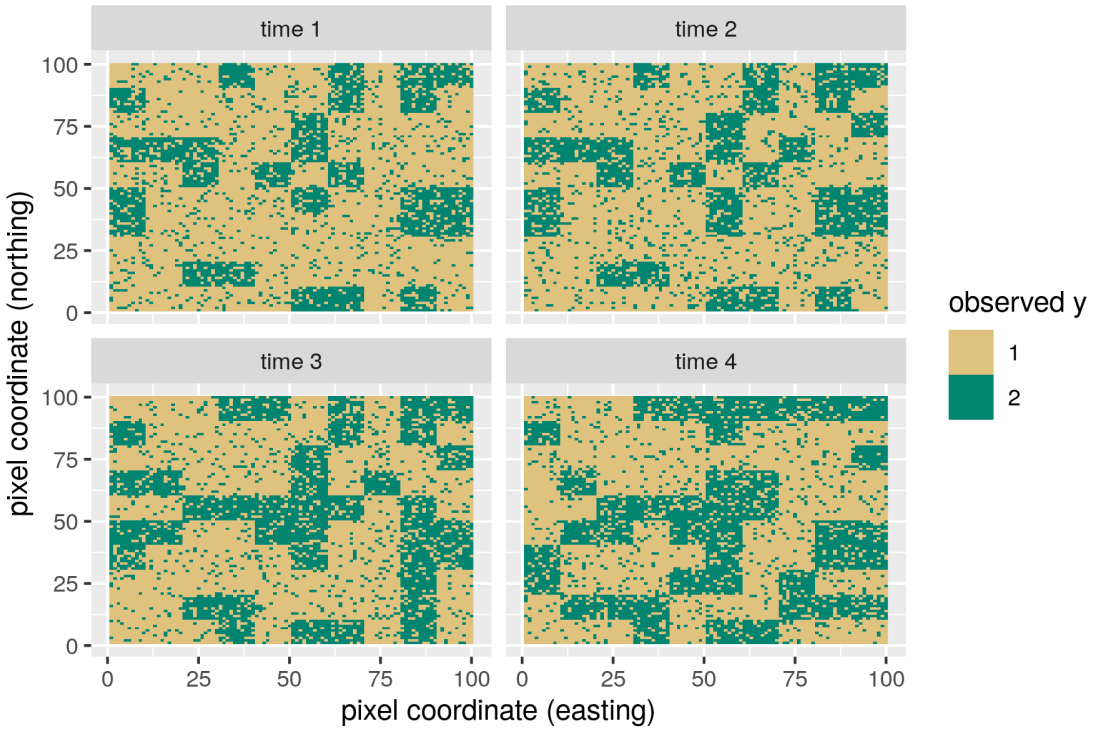


(b) Misclassification Probability

Figure E6: Monte Carlo Results for Varying Transition Probabilities

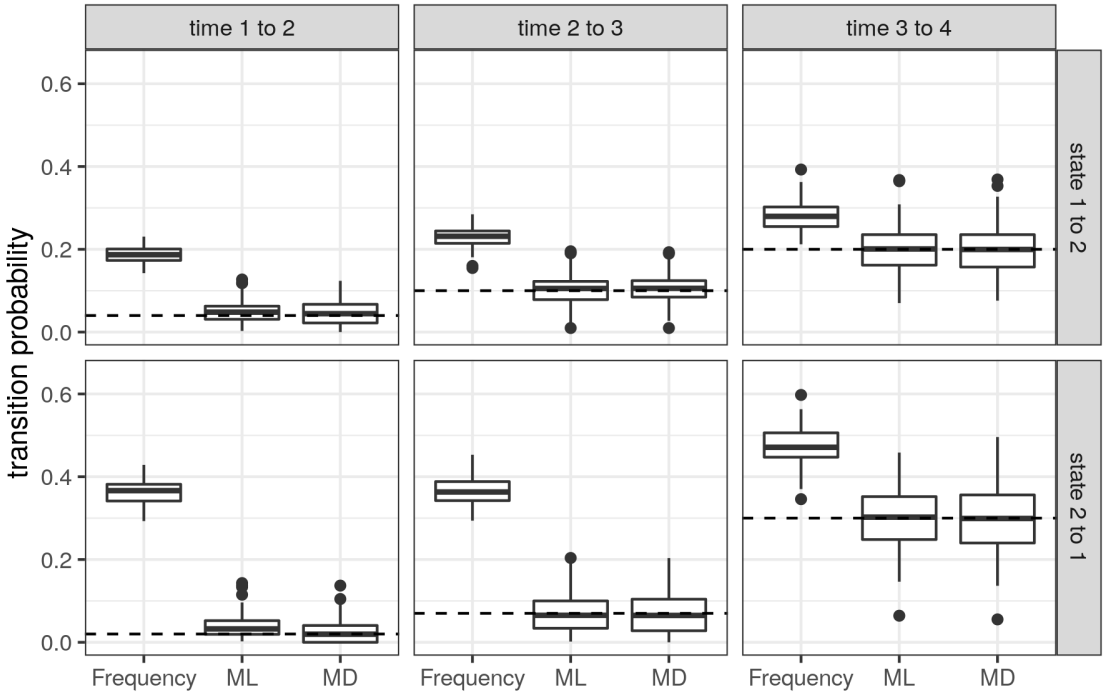


(a) Distribution of True Land Use, S_{it}

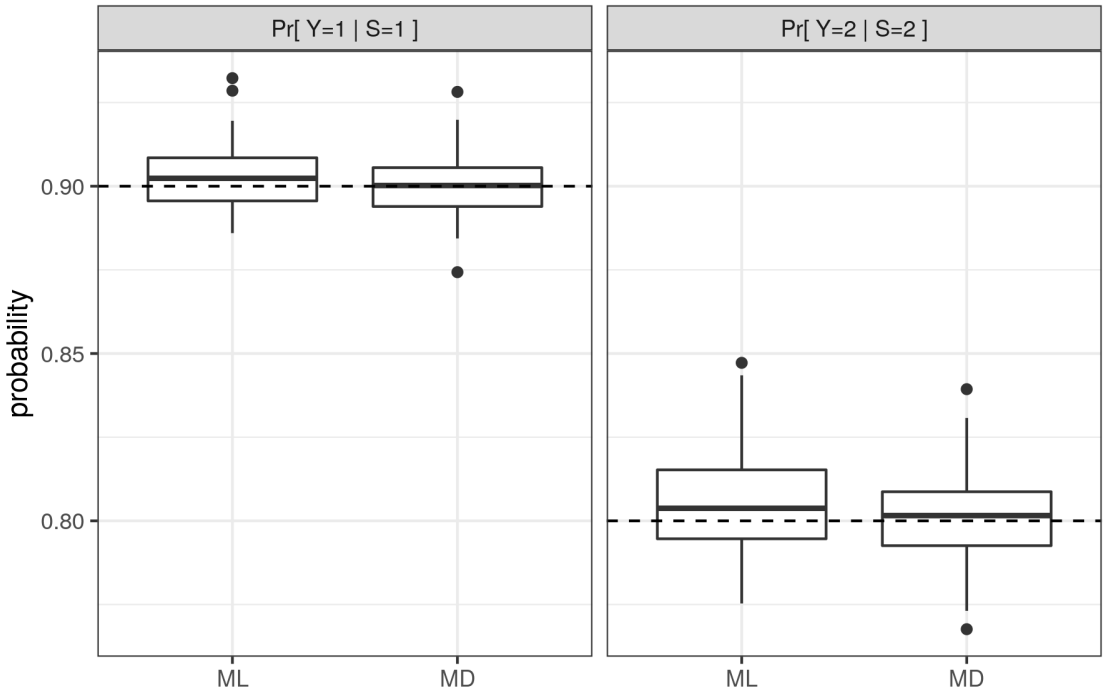


(b) Distribution of Observed Land Use, Y_{it}

Figure E7: Monte Carlo: Spatially Correlated Land Use, an Example

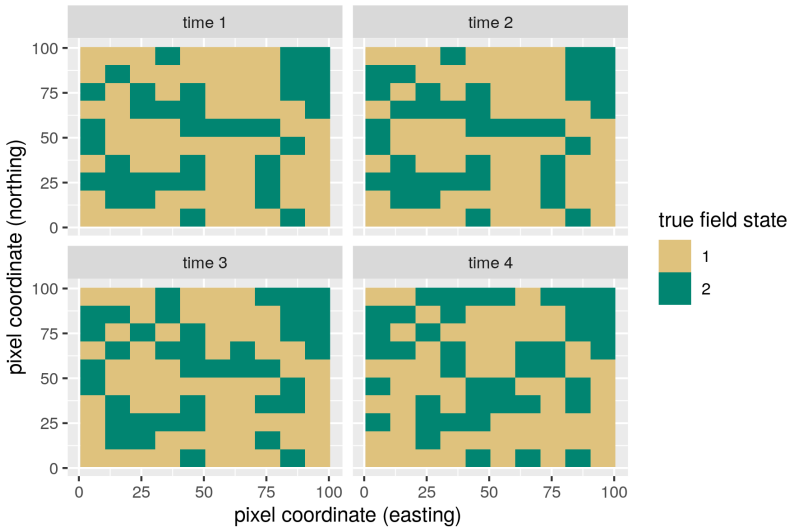


(a) Transition Probabilities

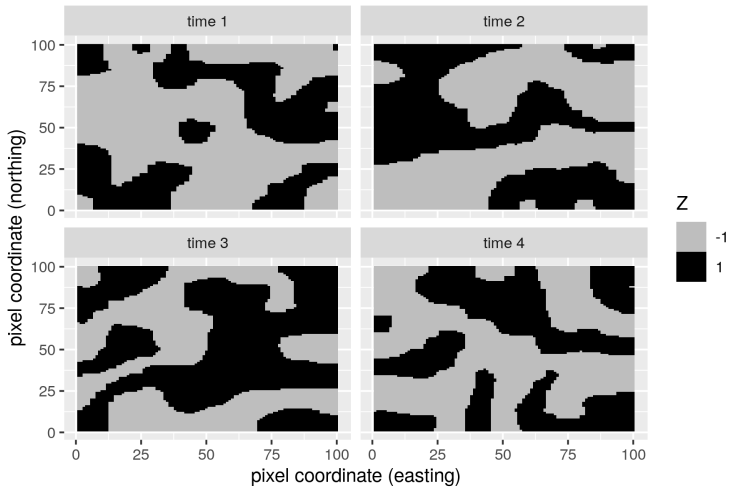


(b) Misclassification Probabilities

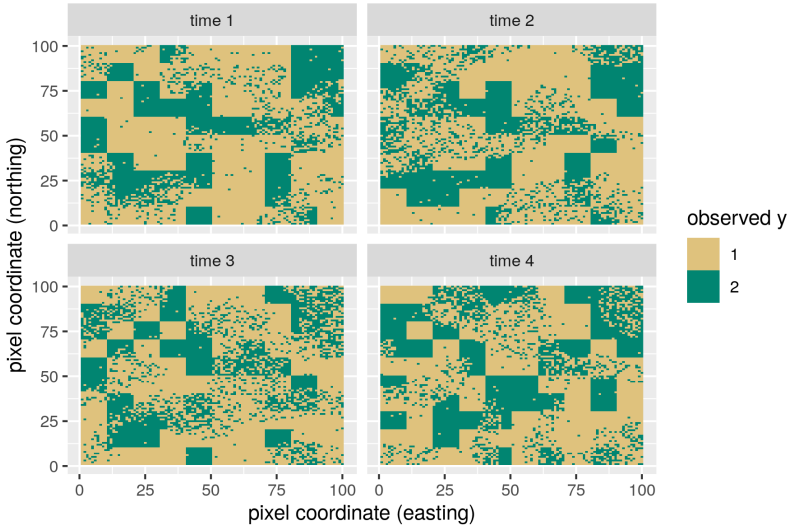
Figure E8: Monte Carlo: Spatially Correlated Land Use Results



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

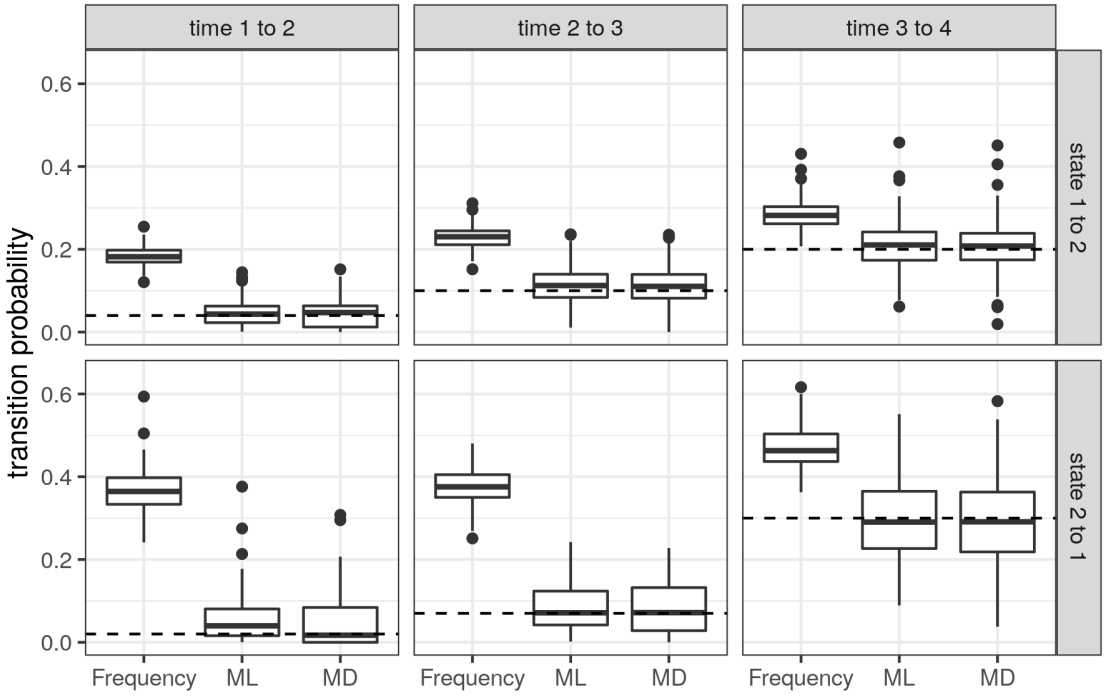


(b) Distribution of Z_{it}

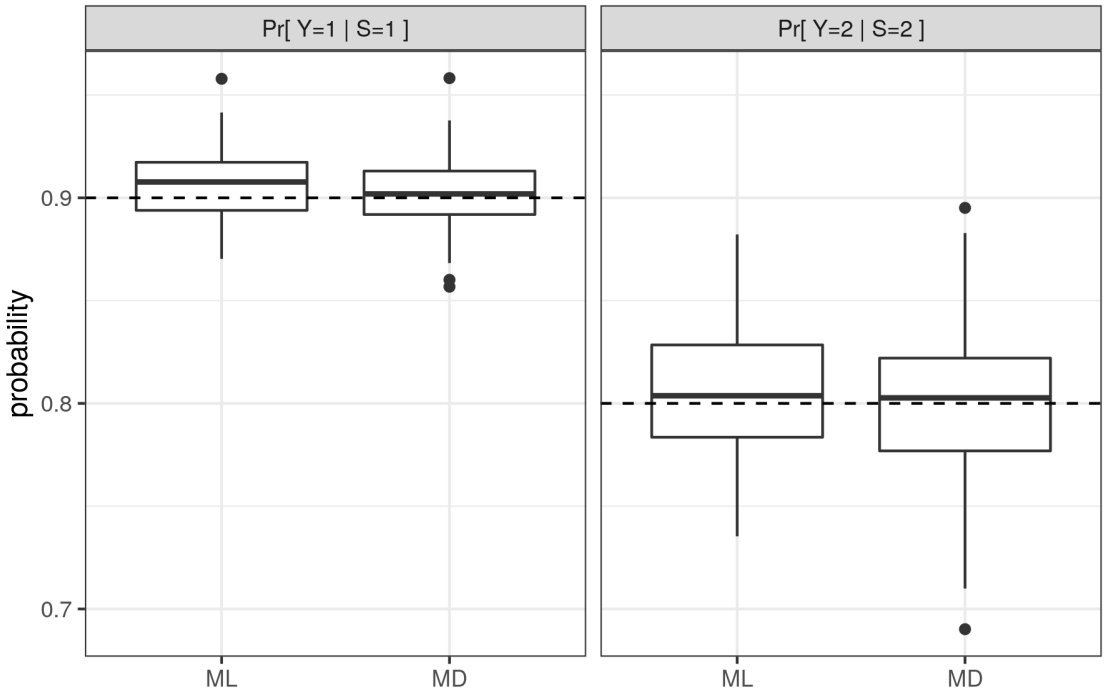


(c) Distribution of Observed Land Use, Y_{it}

Figure E9: Monte Carlo: Spatially Correlated Land Use and Misclassifications, an Example

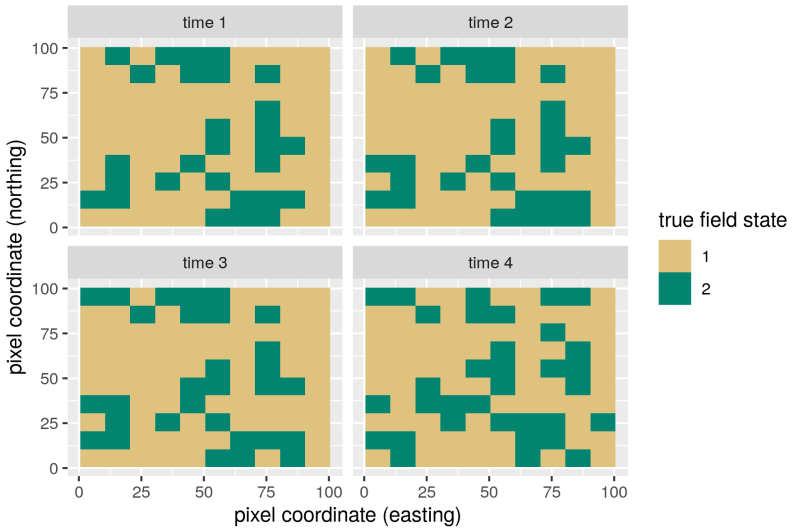


(a) Transition Probabilities

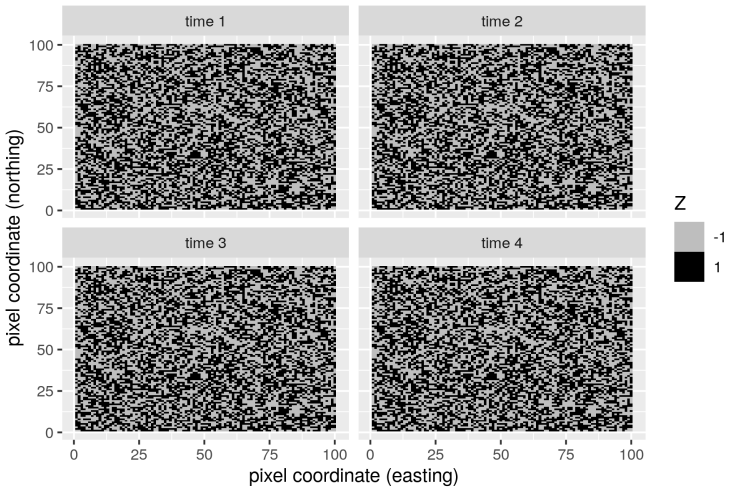


(b) Misclassification Probabilities

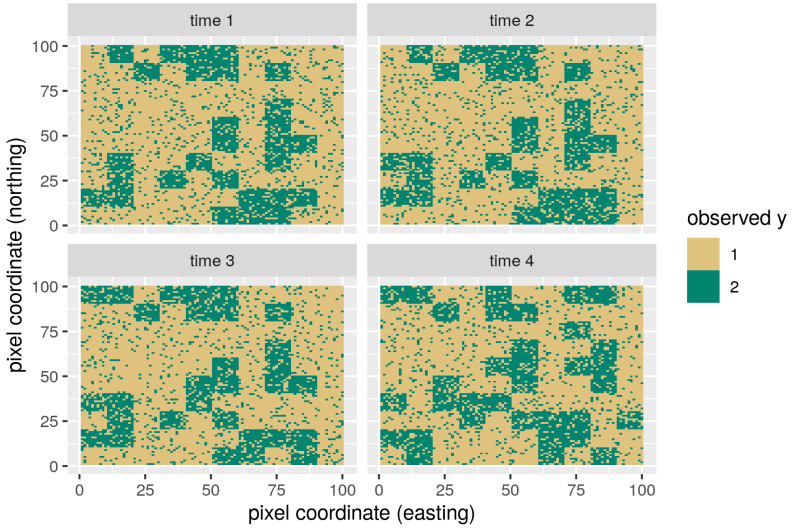
Figure E10: Monte Carlo: Spatially Correlated Land Use and Misclassifications Results



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

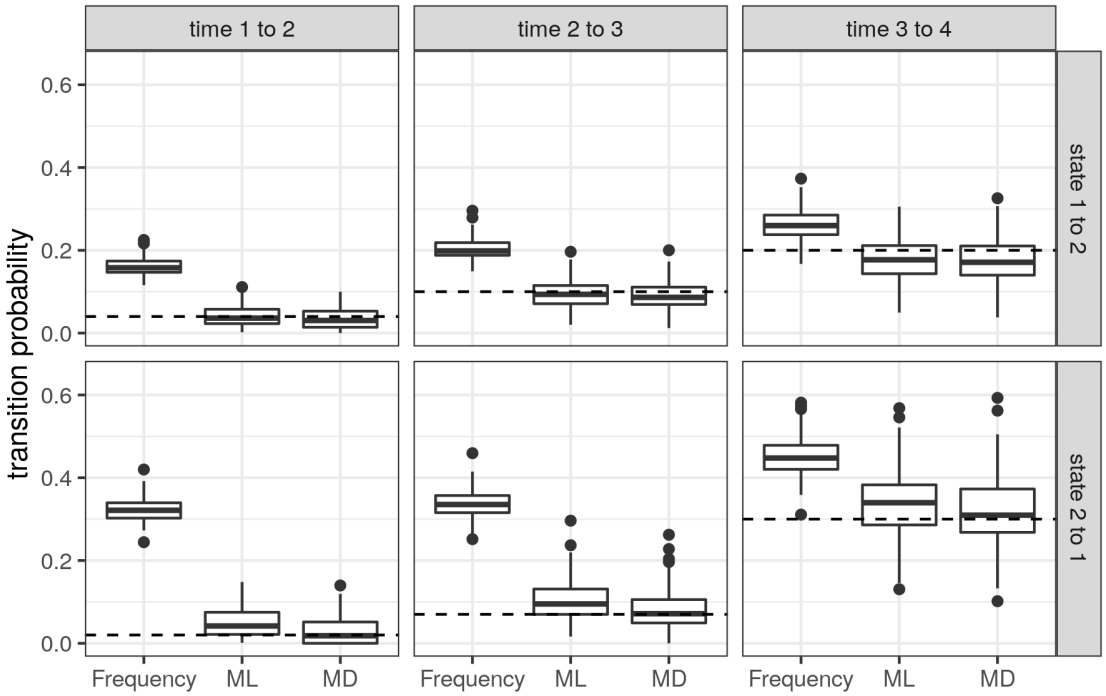


(b) Distribution of Z_{it}

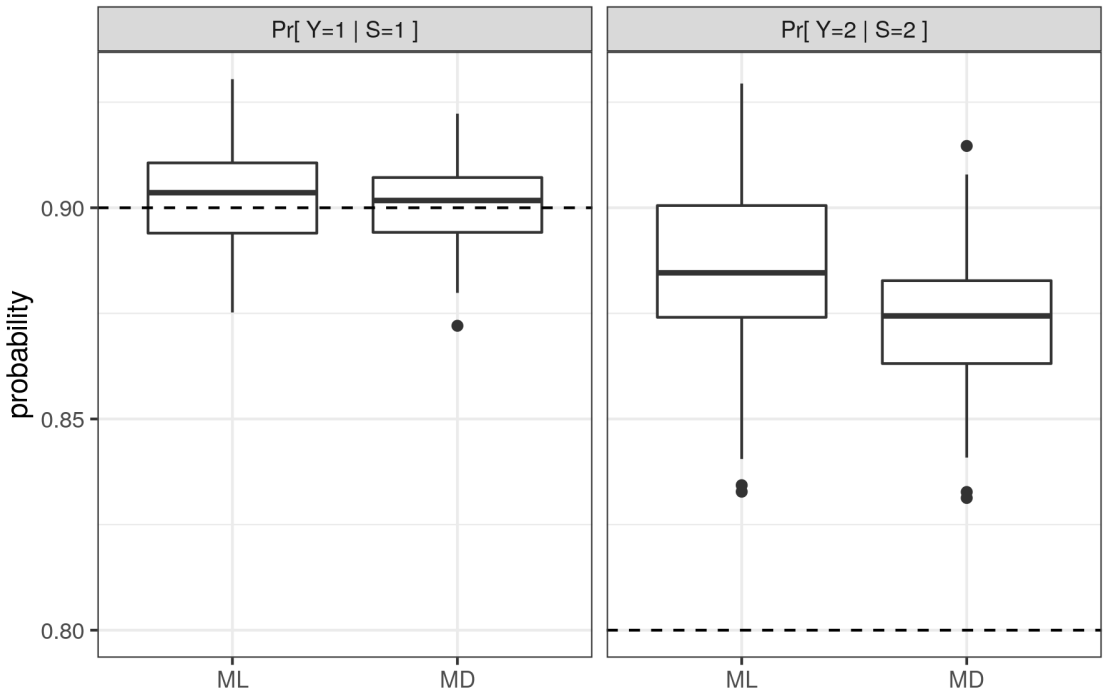


(c) Distribution of Observed Land Use, Y_{it}

Figure E11: Monte Carlo: Spatially Correlated Land Use and Serially Correlated Misclassifications, an Example

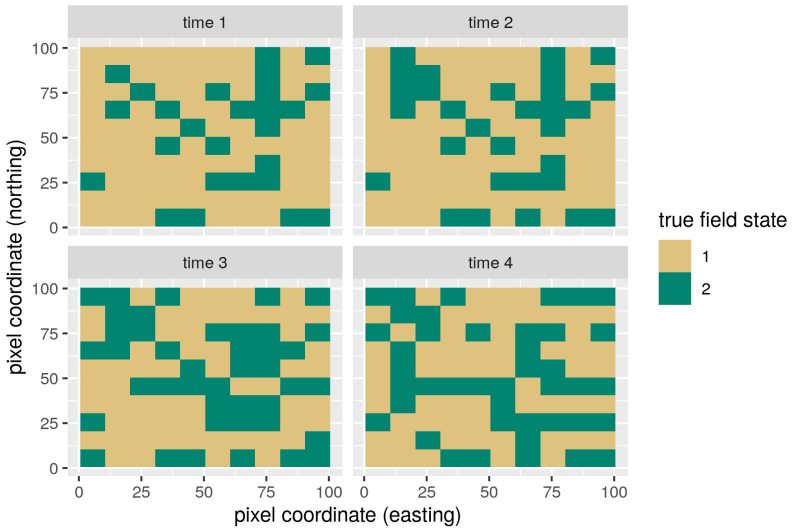


(a) Transition Probabilities

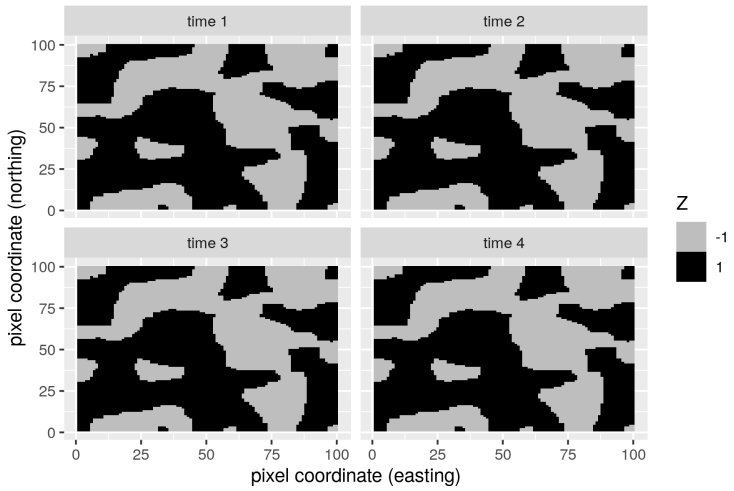


(b) Misclassification Probabilities

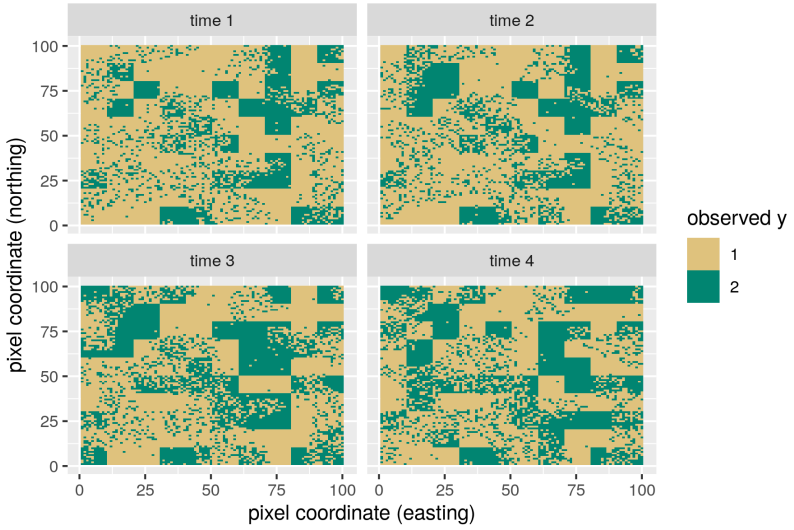
Figure E12: Monte Carlo: Spatially Correlated Land Use and Serially Correlated Misclassifications Results



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

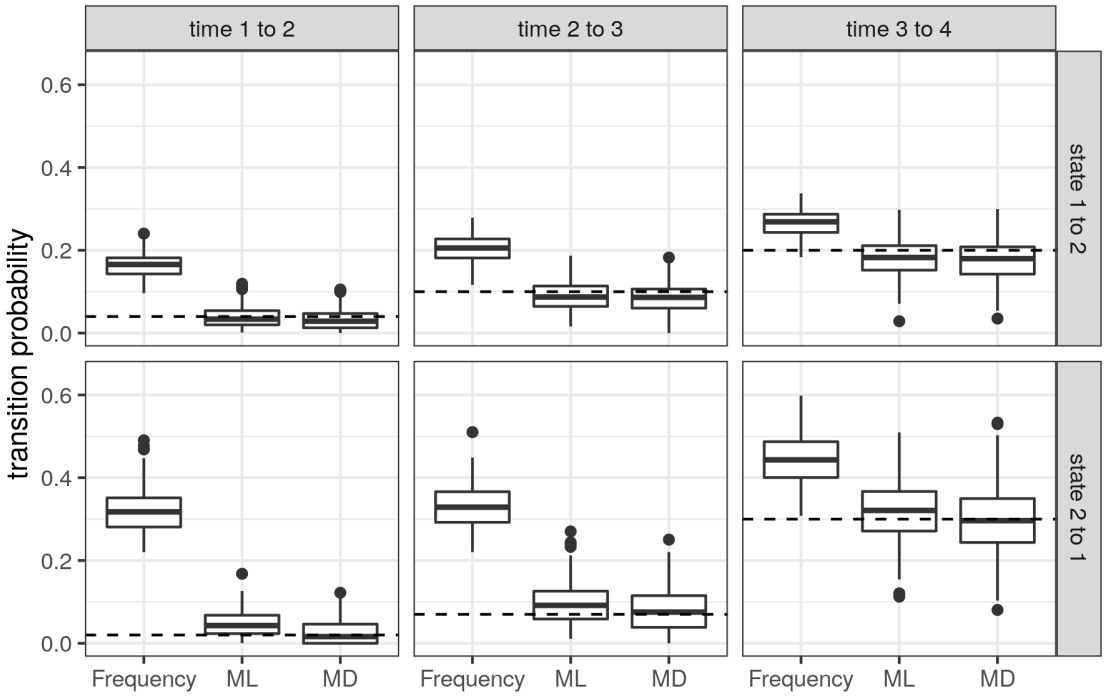


(b) Distribution of Z_{it}

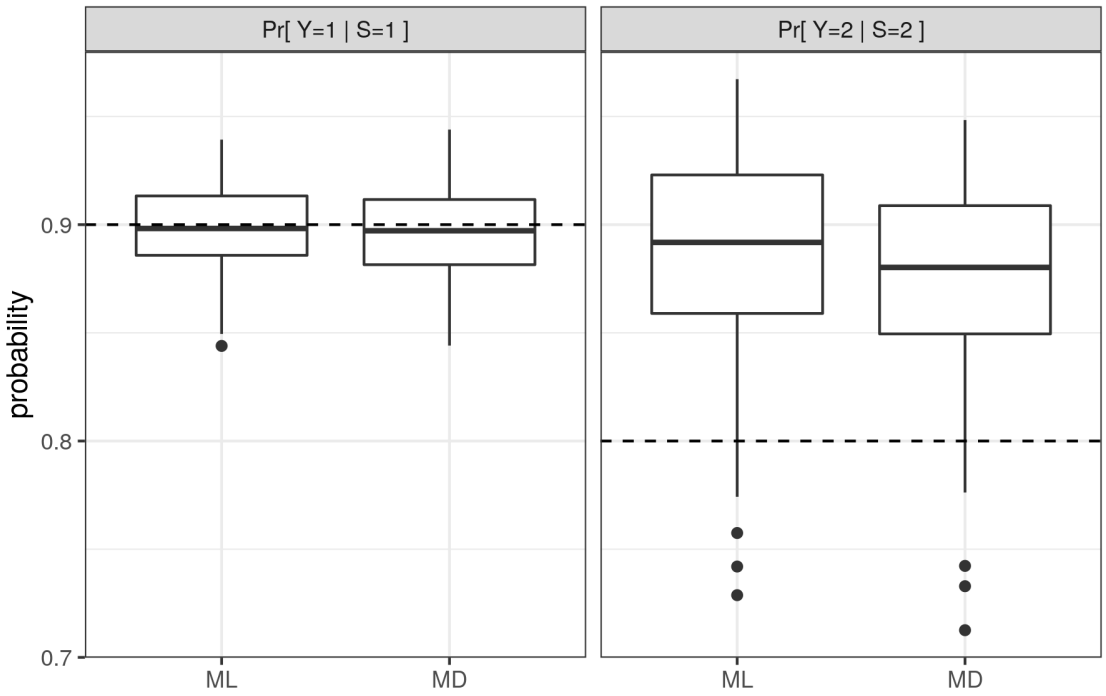


(c) Distribution of Observed Land Use, Y_{it}

Figure E13: Monte Carlo: Spatially Correlated Land Use, and Spatially and Serially Correlated Misclassifications, an Example



(a) Transition Probabilities



(b) Misclassification Probabilities

Figure E14: Monte Carlo: Spatially Correlated Land Use, and Spatially and Serially Correlated Misclassifications Results

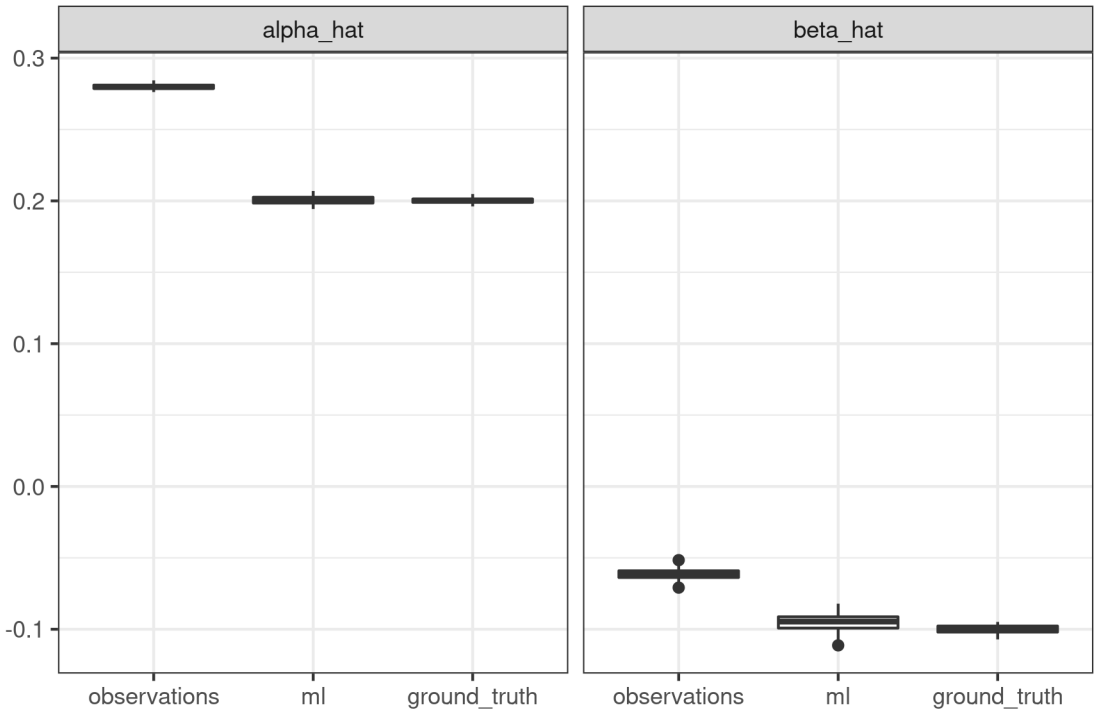


Figure E15: Regression Monte Carlo Results

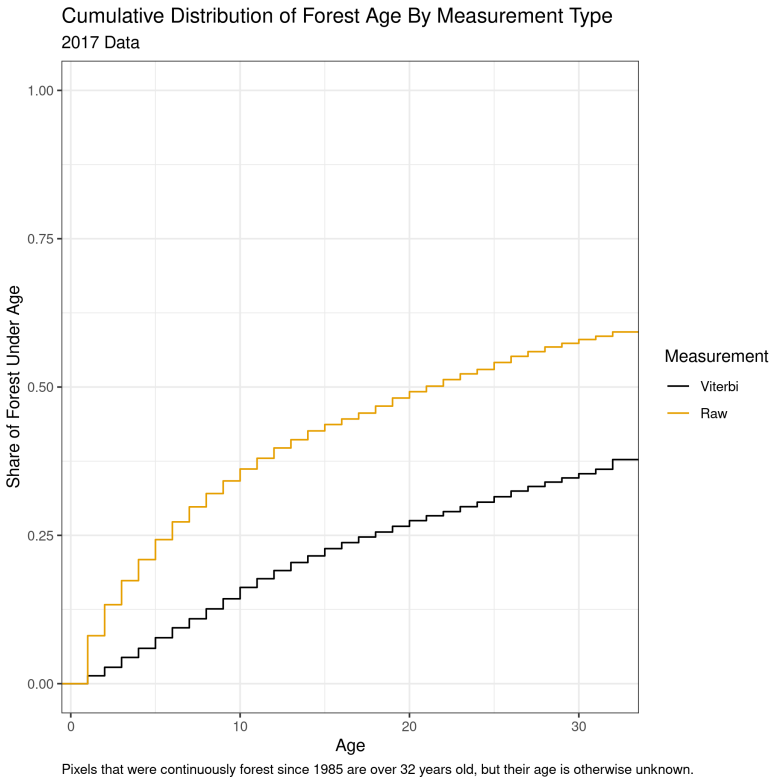


Figure E16: Forest Age Distribution

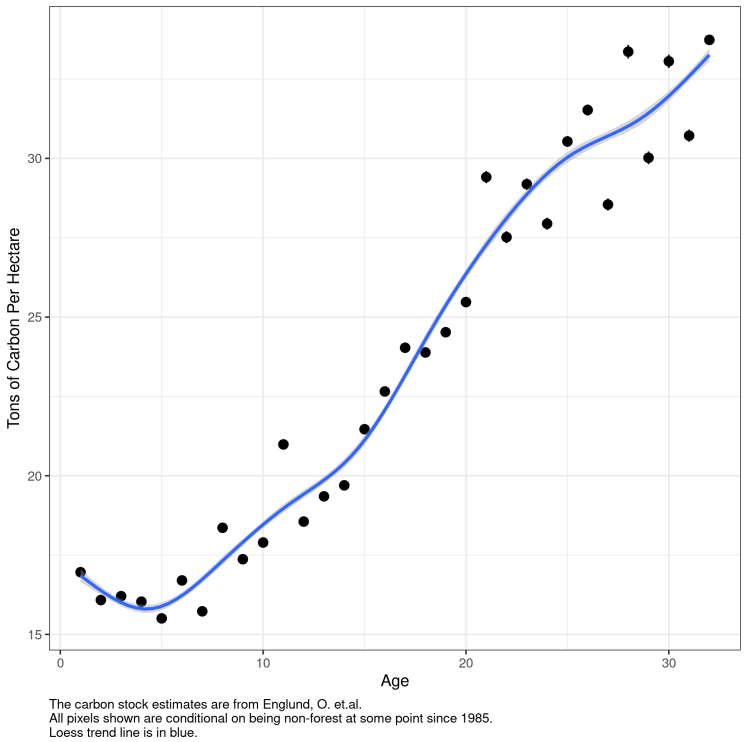


Figure E17: Carbon Stock by Age of Forest