Title: Quantifying uncertainty in land-use land-cover classification using conformal statistics

Denis Valle[1*], Rafael Izbicki[2], Rodrigo Leite[3]

[1] School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, Florida, USA.

[2] Department of Statistics, Federal University of Sao Carlos, Sao Paulo, Brazil.

[3] Department of Forestry, Federal University of Vicosa, Vicosa, Brazil

*Corresponding author: Tel: (352) 392-3806; Email: drvalle@ufl.edu

Target Journal: Full article for Remote Sensing of Environment (max. 15,000 words)

Abstract

Land-use land-cover (LULC) change is one of the most important anthropogenic threats to biodiversity and ecosystems integrity. As a result, the systematic generation of annual regional, national, and global LULC map products derived from the classification of satellite imagery data have become critical inputs for multiple scientific disciplines. The importance of quantifying pixel-level uncertainty to improve the robustness of downstream analyses has long been acknowledged but this practice is still not widely adopted in the generation of these LULC products. The lack of uncertainty quantification is likely due to the fact that most approaches that have been put forward for this task are too computationally intensive for large-scale analysis (e.g., bootstrapping). In this article, we describe how conformal statistics can be used to quantify pixel-level uncertainty in a way that is not computationally intensive, is statistically rigorous despite relying on few assumptions, and can be used together with any classification algorithm that produces class probabilities. Our simulation results show how the size of the predictive sets created by conformal statistics can be used as an indicator of classification uncertainty at the pixel level. Our analysis based on data from the Brazilian Amazon reveals that both forest and water have high certainty whereas pasture and the "natural (other)" category have substantial uncertainty. This information can guide additional ground-truth data collection and the resulting raster combining the LULC classification with the uncertainty results can be used to communicate in a transparent way to downstream users which classified pixels have high or low uncertainty. Given the importance of systematic LULC maps and uncertainty quantification, we believe that this approach will find wide use in the remote sensing community.

Keywords: conformal statistics, land-use land-cover, LULC, uncertainty quantification, image classification

1. Introduction

There has been an increasing trend towards the systematic creation of large-scale (regional, national, and global) land-use land-cover (LULC) map products (Stehman and Foody 2019). Prominent products include annual national (e.g., Mapbiomas (mapbiomas.org; Souza et al. 2020)) and global LULC maps (e.g., Buchhorn et al. 2020; Potapov et al. 2022). Because LULC change is the main driver of terrestrial and freshwater biodiversity and ecosystems integrity loss across the world (Díaz et al. 2019; Tilman et al. 2017), these LULC products have become increasingly foundational for a wide range of downstream scientific applications (Canibe et al. 2022; Jain 2020; Lyons et al. 2018; Stehman and Foody 2019). For example, these maps have been used to generate global estimates of market and non-market value of ecosystem service (Sutton and Costanza 2002), to assess LULC changes associated with large-scale projects such as hydroelectric dams (Guerrero et al. 2020), the potential large-scale benefits as well as opportunity costs of conservation initiatives (e.g., integrated crop-livestock systems with soybeans and Amazon soy moratorium) (Nepstad et al. 2019; Rausch and Gibbs 2021), and to determine how the expanding footprint of human activities (greatly influenced by LULC changes) have impacted wildlife movement (Tucker et al. 2018).

The need for accuracy and uncertainty quantification for these LULC maps has long been acknowledged (Congalton et al. 2014; Foody 2002; Gao et al. 2020; Khatami et al. 2017; Stehman and Foody 2019). For example, it has been shown that the spatial and temporal predictions of species distribution models are strongly affected by uncertainty in LULC maps (Canibe et al. 2022). Similarly, a recent review has highlighted the impact that errors in LULC classification can have for causal inference in the field of environmental economics (Jain 2020). Nevertheless, uncertainty quantification associated with these large-scale LULC products is rare (Lyons et al. 2018). We believe that one of the main reasons for this

refers to the fact that the great majority of the approaches put forth to quantify classification uncertainty have relied on re-sampling approaches such as bootstrapping (Cheng et al. 2021; Hsiao and Cheng 2016; Lyons et al. 2018; Weber and Langille 2007). This is unfortunate because these resampling approaches are often too computationally intensive for large-scale products and bootstrapping is only able to capture the uncertainty associated with variability in the input dataset whereas LULC classification includes numerous other sources of uncertainty (e.g., the ability of these classifiers to make accurate predictions) (Canibe et al. 2022).

Earlier approaches have focused on quantifying uncertainty for population-level parameters such as parameters within, or derived from, confusion matrices (e.g., kappa coefficient, user and producer accuracy statistics; Cheng et al. 2021; Foody 2004; Stehman 1997; Weber and Langille 2007). However, more emphasis has recently been placed on creating spatially explicit uncertainty estimates as these are likely to be more useful to end users (Foody 2002; Gao et al. 2020; Khatami et al. 2017; Stehman and Foody 2019). In this regard, multiple studies have proposed the use of summary statistics based on class probabilities (e.g., maximum class probability, difference between the largest and the second largest probabilities, and Shannon entropy) to quantify classification uncertainty at the pixel level (D'Urso and Menenti 1996; Hsiao and Cheng 2016; Park et al. 2016). Unfortunately, determining which pixels are too uncertain to be used for downstream analysis based on these summary statistics is challenging. For instance, a pixel with a maximum class probability of 0.5 might have very little uncertainty if all the remaining classes have substantially smaller probabilities or might indicate greater uncertainty if another class has probability of 0.5. It is even more challenging to make this type of determination based on more abstract summary statistics such as Shannon entropy. Importantly, some increasingly common classification algorithms (e.g., random forest and deep learning models) can sometimes produce poorly calibrated probability estimates (i.e., the estimated probabilities associated with each

class label overestimate the likelihood that those class labels are actually correct) (Guo et al. 2017; Mukhoti et al. 2020; Niculescu-Mizil and Caruana 2005), potentially resulting in classifiers confidently mis-classifying certain cases.

In this article, we introduce conformal statistics as a straight-forward but rigorous approach to quantify uncertainty in LULC classification that holds great promise for three main reasons. First, it does not rely on resampling approaches and, as a result, it is not as computationally intensive as bootstrapping. Second, it can work with any classification algorithm (e.g., machine learning algorithms) that can generate estimates of the probability associated with each class, an important characteristic given that machine learning classifiers tend to outperform their parametric counterparts (Maxwell et al. 2018). Third, conformal statistics rely on minimal assumptions to generate predictive sets with the desired empirical coverage. Given that conformal statistics is, to our knowledge, new to the field of remote sensing, we start this article by providing a brief overview of conformal statistics. We then show how this approach compares favorably against other alternative approaches using simulated data and illustrate the insights that can be obtained using empirical data from the Amazon region. Finally, we end this article with a discussion on remaining challenges and future research.

2. Conformal statistics

Conformal statistics (also known as distribution-free uncertainty quantification) is focused on creating predictive sets or regions Γ_C that have the desired empirical coverage *C* (Angelopoulos and Bates 2021; Shafer and Vovk 2008). For example, a 95% predictive set/region is valid if it contains the truth 95% of the time. While conformal approaches have also been developed for continuous response variables (i.e., regression setting) (Izbicki et al. 2022; Lei et al. 2018; Romano et al. 2019), here we focus on its use for

classification problems. In this case, the response variable y_i consists of the class label (y_i=1,2,…, K, where K is the overall number of classes) for unit i (e.g., pixel, image, or individual tree) and therefore the predictive set generated by conformal statistics refers to a subset of class labels (e.g., $\Gamma_C = \{1,2,4\}$ in a 5-class classification problem). More formally, conformal statistics is focused in creating Γ_C such that:

$$p(y_i \in \Gamma_C) \geq C$$

where p() stands for probability. As a result, the number of classes in the subset Γ_C (the cardinality of Γ_C) is a natural measure of the amount of uncertainty. For example, if the subset Γ_C only contains class label 2, this is an indication of very low uncertainty in our predictions. On the other hand, if the subset Γ_C contains class labels 1, 2, and 4, this indicates higher uncertainty because it suggests that any of these classes are likely to be the true class. Interestingly, it is also possible for the subset Γ_C to be empty. This can happen if none of the classes are likely to be the true one according to our predictive algorithm. Therefore, empty subsets also indicate high uncertainty.

Note that conformal statistics is focused on creating predictive sets that include the true (reference) class $C \times 100$ percent of the times. As a result, differently from earlier approaches that only took into account accuracy metrics (i.e., how well the map labels matched the true (reference) classes; Cheng et al. 2021; Foody 2004; Stehman 1997; Weber and Langille 2007) or approaches that only took into account the distribution of the estimated class probabilities (D'Urso and Menenti 1996; Hsiao and Cheng 2016; Park et al. 2016), the uncertainty quantification from conformal statistics relies on both the distribution of class probabilities and how likely the map labels agree with the true classes.

There are a variety of conformal statistics approaches. In this article, we focus on inductive (also known as split) conformal prediction, one of the most widely used conformal approaches. In this approach, we start by splitting the data into two sets called the training and the calibration datasets. We fit the classification algorithm to the training data and use this algorithm to output the probability for each class $\hat{f}(\boldsymbol{x_i}) \in [0,1]^K$, where K is the total number of classes and $\boldsymbol{x_i}$ is the vector containing the predictor variables associated with the observation within the calibration set that we wish to classify. In the next step, we define the score s_i as the class probability associated with the true class y_i (i.e., $s_i = \hat{f}(\boldsymbol{x_i})_{y_i}$). Assuming that the calibration dataset contains *n* observations, we calculate the score for all observations within this dataset (i.e., $s_1, \ldots, s_n$). Table 1 illustrates these calculations for 4 observations in the calibration dataset.

Table 1. Example of the calculation of the scores for 4 hypothetical observations in the calibration dataset. Cells with bold numbers correspond to the probabilities associated with the true classes.

Observations	True class y_i	Class probabilities $\hat{f}(\boldsymbol{x_i})$					Score s_i
		1	2	3	4	5	
1	3	0.10	0.10	**0.80**	0.00	0.00	0.80
2	3	0.00	0.10	**0.30**	0.30	0.30	0.30
3	1	**0.25**	0.25	0.10	0.40	0.00	0.25
4	4	0.00	0.00	0.05	**0.90**	0.05	0.90

The next step consists of using the conformal scores $s_1, \ldots, s_n$ in the calibration dataset (i.e., rightmost column in Table 1) to calculate $\hat{q}_{1-C}$, the $1 - C^{th}$ empirical quantile of these scores. For instance, if predictive sets that have 90% coverage are desired, then C=0.9 and $\hat{q}_{1-C}$ is the 10% quantile. This quantity can be readily calculated by ordering the conformal scores $s_1, \ldots, s_n$ and picking the value for

which 10% of the $s_1, \ldots, s_n$ scores are below it. Finally, we create a predictive set for each pixel in the area of interest. This predictive set is defined as the subset containing all the classes for which $\hat{f}(\boldsymbol{x_{test}})_y \geq \hat{q}_{1-C}$, where $\boldsymbol{x_{test}}$ is a vector containing the predictor variables for the test pixel. Table 2 illustrates the resulting 90% predictive sets for five new observations, assuming that $\hat{q}_{0.1}$ (the 10% quantile of the conformal scores $s_1, \ldots, s_n$ from the calibration dataset) was calculated to be equal to 0.21. Notice that the predictive set for observation 2 contains 4 labels because all of these labels have probability higher than our $\hat{q}_{0.1}$ threshold of 0.21. On the other hand, observation 5 has an empty predictive set because none of the labels have high enough probability. In other words, none of the 5 labels is likely to be the true label.

Uncertainty is quantified by assessing the size of these predictive sets. More specifically, uncertainty is smallest when the predictive set contains only a single class and increases with the number of classes within the predictive set. For example, observations 3, 4, and 2 in Table 2 have increasingly higher classification uncertainty because their predictive sets are increasingly larger. However, uncertainty is also high when the predictive set is empty (e.g., observation 5 in Table 2).

Table 2. Example of creating 90% predictive sets for new observations, assuming that $\hat{q}_{0.1} = 0.21$. Cells with bold numbers correspond to labels y that satisfy the inequality $\hat{f}(\boldsymbol{x_{test}})_y \geq \hat{q}_{0.1}$.

Observations	Class probabilities $\hat{f}(\boldsymbol{x_{test}})$					90% Predictive sets
	1	2	3	4	5	
1	**0.85**	0.05	0.00	0.10	0.00	{1}
2	**0.25**	**0.25**	**0.25**	**0.24**	0.01	{1,2,3,4}
3	0.10	0.10	**0.40**	**0.40**	0.00	{3,4}

4	**0.30**	0.00	**0.25**	0.05	**0.40**	{1,3,5}
5	0.20	0.20	0.20	0.20	0.20	{}

168

169 One of the most attractive features of conformal statistics is that it does not rely on parametric
170 assumptions. The only required assumption is that observations are exchangeable (or the slightly stricter
171 assumption that the observations are independent and identically distributed), a common assumption
172 across the great majority of the machine learning methods (Shafer and Vovk 2008). Although we refrain
173 from providing a mathematical proof, an intuitive explanation of why this approach works is that,
174 because of the exchangeability assumption, the coverage result for the calibration dataset should be the
175 same as the coverage result for the dataset for which we want predictions (i.e., test data) if the same
176 method is applied to both datasets. In our example, if 90% of the true classes have an estimated
177 probability greater than 0.21 in the calibration dataset, then predictive sets defined by all classes with
178 probability greater than 0.21 for the test data should also encompass the true class 90% of the times.

179

180 The inductive/split conformal prediction should feel familiar to remote sensing experts that have
181 performed cross-validation due to the process of splitting the data into multiple sets. The major
182 difference is that conformal statistics fits the model only once to the training dataset and this approach
183 "learns" how to generate predictive sets with the desired coverage using the left-out calibration data. A
184 very well written introduction to conformal statistics can be found in Angelopoulos and Bates (2021)
185 while a comprehensive treatment of this topic can be found in Vovk et al. (2005). Note that the
186 conformal score s_i is typically defined in such a way that higher scores imply lower confidence in
187 predictions. However, we have defined the conformal score differently (i.e., higher score values indicate
188 greater confidence in the prediction) because we believe that it is easier to understand conformal
189 predictions this way. We also note that other conformal scores can be used and these may lead to

different predictive sets. However, this article focuses on a single score to simplify the presentation of the methodology. Readers interested in learning about other scores can find additional information in the articles from Angelopoulos and Bates (2021), Chernozhukov et al. (2021), Izbicki et al. (2020), Izbicki et al. (2022), and Romano et al. (2020) and the references therein. Finally, to facilitate the adoption of this methodology, we provide a short tutorial to illustrate how to generate predictive sets in R using conformal statistics (Appendix 1).

3. Simulations

While any classification method could have been used, we rely on the random forest classifier for all simulated data examples as this is a very popular method for LULC classification (see review in Belgiu and Dragut 2016). We relied on the 'randomforest' R package (Liaw and Wiener 2002) with its default settings. Furthermore, we rely on external test datasets (i.e., datasets not used to fit the model or used to learn how to generate predictive sets) to assess the performance of this method. More specifically, we determine empirical coverage by calculating the proportion of times that the true classes in the test dataset were contained in the corresponding predictive sets.

3.1. Simulation set 1.

The goal of simulation set 1 is to show how conformal prediction can generate predictive sets that vary in size according to how challenging the prediction is. To this end, we created two simulated data sets. The first one contained 3 groups while the second one had 5 groups. The probabilities associated with these groups varied smoothly as a function of a single predictor variable x. This variable x was generated by creating a sequence of evenly spaced numbers between -3 and 3. We simulated a total of 50,000

observations and 88%, 2%, and 10% of these observations were randomly allocated to the training, calibration, and test datasets, respectively. We set the desired coverage level of the predictive sets to C=0.8.

3.2. Simulation set 2.

The goal of simulation set 2 is to compare conformal methods to other methods used to quantify uncertainty. We start by simulating the predictor variables x_1, x_2, and x_3 in the following way:

$$x_p \sim Unif(-3,3) \text{ for } p \in \{1,2,3\}$$

We assume that the response variable y can only belong to one of three classes, with probability for classes 1, 2, and 3, given by $\frac{\exp(x_1+x_2)}{D}$, $\frac{\exp(x_1-x_2)}{D}$, and $\frac{\exp(x_2+x_3)}{D}$, respectively. In this expression, D is the normalizing constant that ensures that these numbers sum to one (i.e., $D = \exp(x_1 + x_2) + \exp(x_1 - x_2) + \exp(x_2 + x_3)$).

We simulated 10 datasets. Each dataset contained 5,000 observations; 4,000 were used for training, 500 were used for calibration, and 500 were used for test purposes. Our classification algorithm was trained under two different scenarios. In the first scenario, all three predictor variables (i.e., $x_1, x_2, \text{and } x_3$) were available to the classifier. In the second scenario, we make classification more challenging by excluding the predictor variable x_3 (i.e., the model was trained using only predictor variables x_1 and x_2). Similar to simulation set 1, we train a random forest classifier and set the desired coverage level of the predictive sets to C=0.8.

We compare the conformal statistics approach to two other methods used to quantify uncertainty. The first method (onwards the conventional approach) starts by predicting the class probabilities for each observation in the test dataset. We create the smallest set that encompasses the true class 80% of the time in the following way. We first order these probabilities from greatest to smallest and then we include classes until the sum of their probabilities just exceeds 80%. For example, if the predicted probabilities are $[0.1,0.7,0.2]$ for classes 1, 2, and 3, respectively, then only classes 2 and 3 would be part of our 80% predictive set. The second method to quantifying uncertainty is similar to the approach adopted by Hsiao and Cheng (2016). In this bootstrap approach, we resample the training data with replacement 100 times and make 100 predictions for each observation in the test dataset. Then, we calculate the proportion of times that the different classes are predicted for each observation, yielding a vector of proportions. Finally, we use the same approach as the conventional approach to calculate predictive sets that encompass the true class 80% of the time.

4. Case study

To illustrate the use of conformal statistics based on a real example, we train a random forest classifier to the data used by Mapbiomas to validate their annual LULC classification products for Brazil (freely available at https://mapbiomas.org/pontos-de-validacao). These data were created by visually inspecting satellite imagery for each year between 1985 and 2018. Pixels were selected for inspection based on stratified random sampling and each pixel was evaluated by 3 independent analysts (Souza et al. 2020). For our purposes, we only used pixels for which the 3 analysts agreed on the LULC class to avoid introducing additional uncertainty associated with inconsistent reference class labels.

Our study region consisted of an area of approximately 80,000 km^2 in the Amazon region in Brazil that is traversed by the Transamazon highway (Fig. 1). To avoid using observations from very different ecosystems within Brazil, we selected Mapbiomas observations that were within approximately 200 km of this highway. We focused solely on 2018 LULC classes and we dropped LULC classes that were not observed (labeled as "non-observed" by Mapbiomas) or were too infrequently observed (i.e., temporary crops, urban area, and other non-vegetated areas). We also combined 3 natural classes that are likely to have similar spectral signatures and that were also relatively rare (e.g., savanna, grassland, other non-forest formations (natural)). Ultimately, this process resulted in 4,346 observations with 4 LULC classes (forest, pasture, water, and natural (others)).

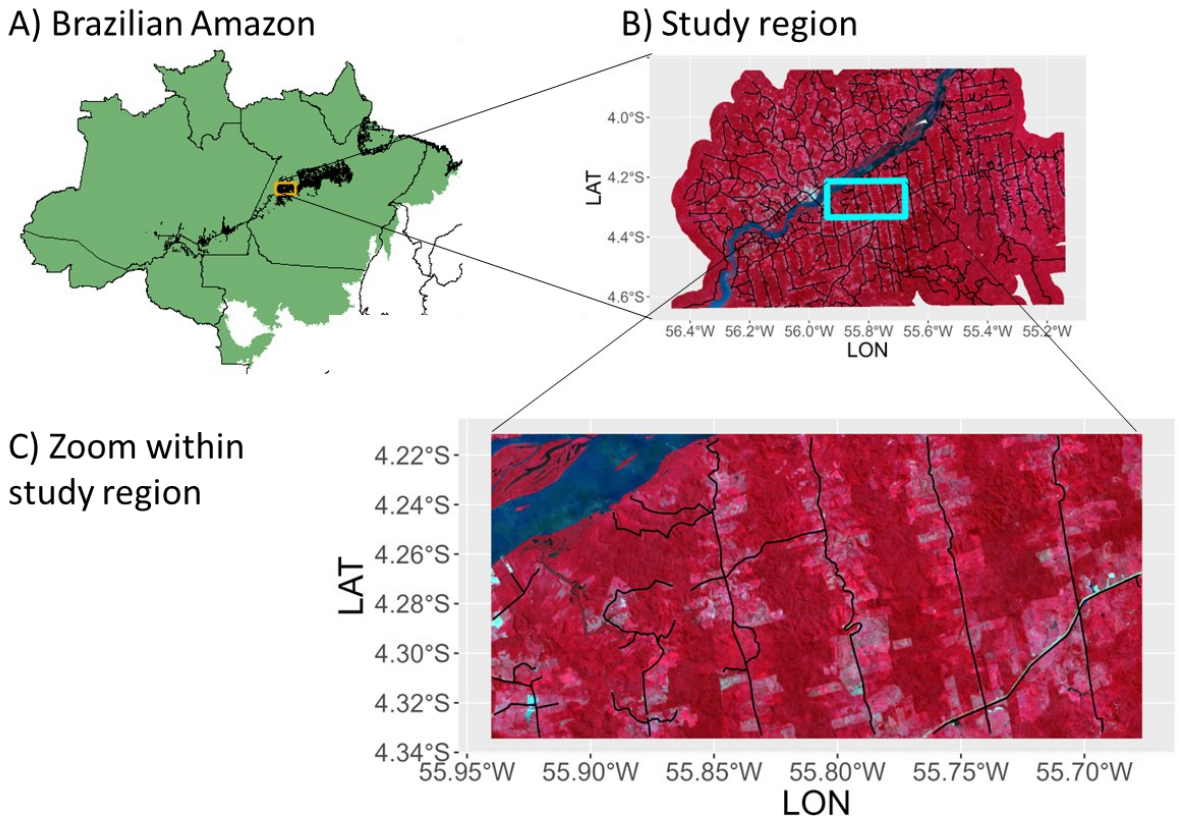

Fig. 1. Study region. Panel A displays the Brazilian Amazon (green polygon) and the study region (red rectangle). Panel B shows a false-color Landsat 8 mosaic (bands 4,3, and 2 were assigned to red, green, and blue, respectively) of the study region, created by calculating the median value per band of several 2018 images after removing pixels classified as cloud or shadow. Panel C zooms to a portion of the study region, shown with cyan rectangle in panel B. In all panels, the road network (obtained from Carrero 2022) is shown with black lines.

Spectral information for each ground-truth observation from Mapbiomas and for the overall study region was acquired from 2018 Landsat 8 imagery using Google Earth Engine (Gorelick et al. 2017). We trained a random forest algorithm on 80% of the observations, chosen completely at random, reserving 20% for calibration of our conformal statistics procedure. We used this classifier to make point predictions of LULC and to calculate the size of predictive sets with C=0.9 for the entire study region.

To evaluate the accuracy of our classification and how well this conformal methodology was able to quantify uncertainty, we relied on a 10-fold spatial cross-validation. More specifically, we divided our ground-truth observations into 10 non-overlapping spatial blocks. For each cross-validation fold, we use observations from 9 of these blocks to train and calibrate the model (with the training/calibration split being 80%/20%) to then predict the classes and create predictive sets for the observations in the left-out block. We use these out-of-sample predictions and predictive sets to create a confusion matrix and to determine empirical coverage.

5. Results

5.1. Simulation set 1

We find an empirical coverage of 80.64% and 79.92% for the simulated test datasets with 3 and 5 classes, respectively, values close to the desired coverage of C=0.8. Furthermore, we find that the mean predictive set size tends to be close to 1 whenever the class probabilities peak, indicating higher certainty in the predicted label (Fig. 2). On the other hand, for the dataset with 3 classes, the mean predictive set size tends to dip below 1 (indicating the presence of several empty predictive sets and therefore higher uncertainty) whenever the true class probabilities are small and prediction is more challenging (Fig. 2C). For the dataset with 5 classes, the mean predictive set size tends to rise above 1 (indicating the presence of several predictive sets with more than one class and therefore higher uncertainty) whenever the true class probabilities are small and prediction is more challenging (Fig. 2D).

Note that a mean predictive set size close to one is not the result of averaging predictive sets of size 0 (i.e., empty sets) and size 2 because the analysis of the dataset with 3 classes yielded only empty sets and sets of size 1 whereas the analysis for the dataset with 5 classes did not yield any empty predictive set. Ultimately, these simulated data results indicate that the predictive set size can be used as a measure of prediction difficulty and uncertainty as long as one remembers that both empty and large predictive sets correspond to high uncertainty.

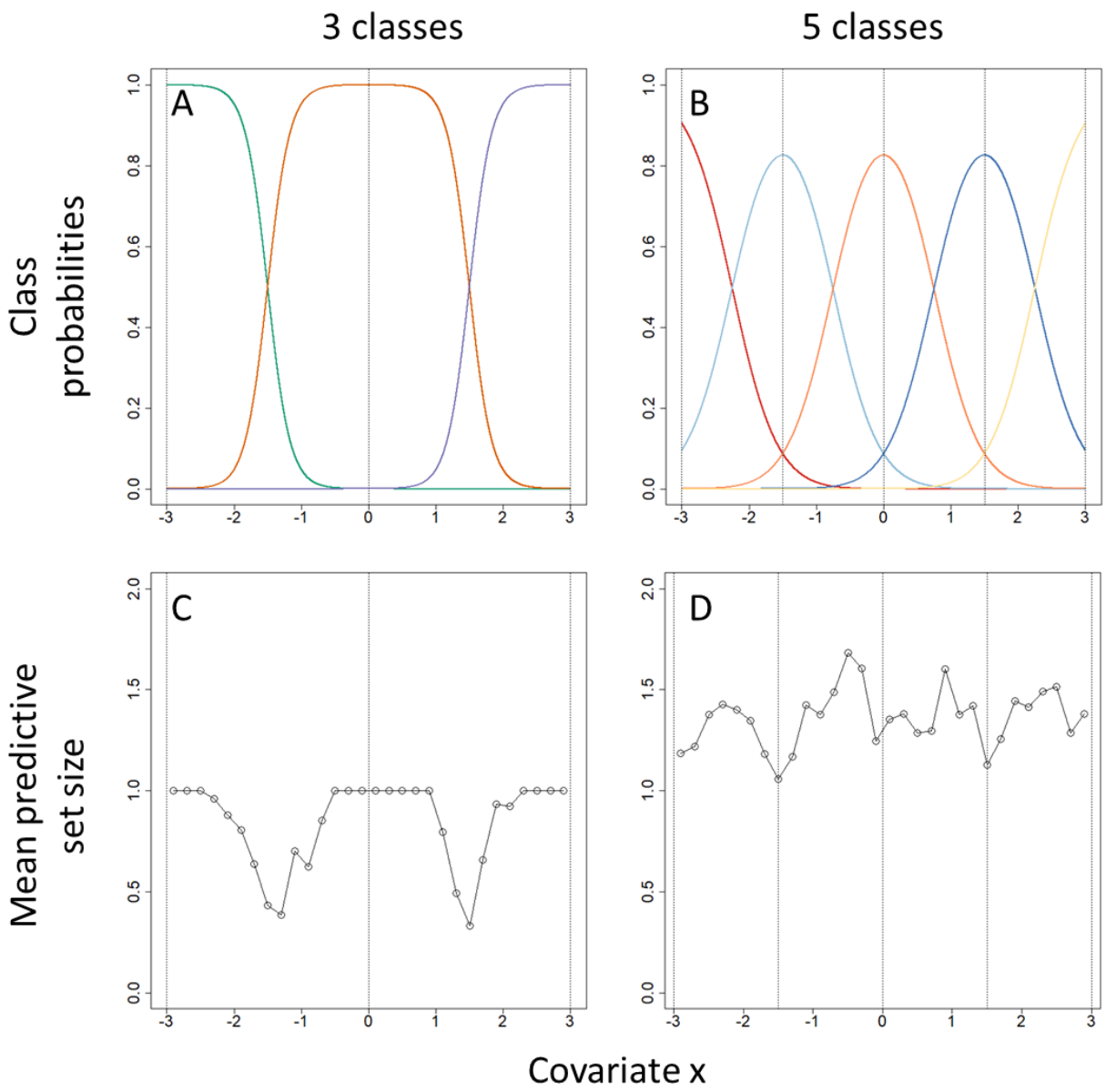

Fig. 2. Mean predictive set sizes that are greater or smaller than one indicate greater classification uncertainty. Panels A and B show how the probability associated with each class (used to simulate the data) changes as a function of the covariate x. Each line represents a different class. Panels C and D show the conformal statistics results, revealing how the mean size of the predictive sets (calculated by discretizing the covariate x into bins of width of 0.5) changes as a function of x. In all panels, the vertical

grey lines show where the probability for each class peaks. Left and right panels show class probabilities and conformal statistics results for datasets with 3 and 5 classes, respectively.

5.2. Simulation set 2

Recall that we rely on a classification problem with three predictor variables x_1, x_2, and x_3. While this is a straight-forward problem when these three variables are known, there is much greater uncertainty if x_3 is not available for the classification algorithm. Our simulation results reveal that the conformal approach was able to retain the desired coverage of C=0.8 for the test dataset regardless of the variable x_3 being available for the classifier or not (Fig. 3). In contrast, the conventional approach showed a larger empirical coverage than desired regardless of the presence or absence of x_3. The bootstrap approach performed better than the conventional approach but still suffered from larger empirical coverage when x_3 was present.

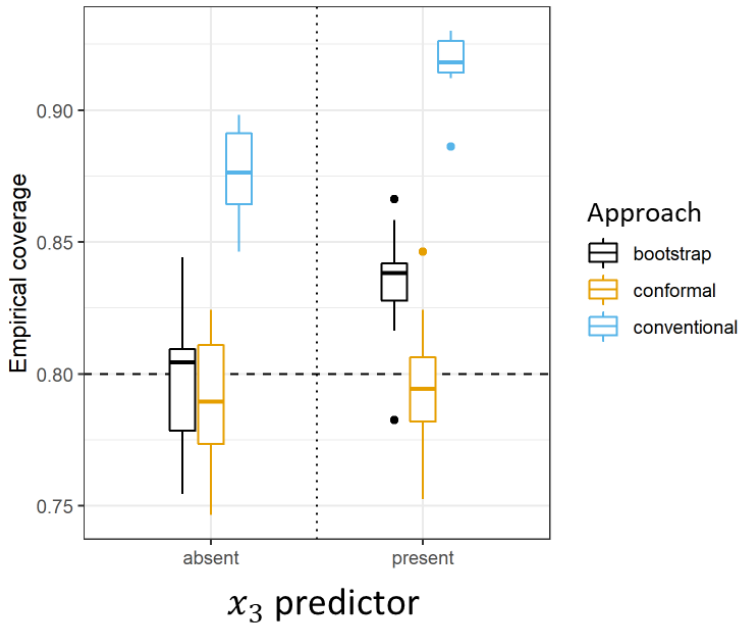

Fig. 3. The conformal approach outperforms the conventional and the bootstrapping approaches. Empirical coverage is shown for 2 classification scenarios in which the x_3 predictor variable was either absent or present. The horizontal dashed line is the 80% desired coverage.

We emphasize that the conformal approach can generate predictive sets with any desired coverage. To illustrate this, we use the same datasets as before but now we systematically vary the desired coverage from 0.5 to 0.9. Our results show that the conformal predictive sets generally have the desired coverage (Figs. 4). A comparison of these conformal results to those from the conventional and bootstrap approaches reveals that the predictive sets created by these latter approaches in general had empirical coverage that did not match the desired coverage. Part of the reason for this pattern is that both the conventional and the bootstrap approaches cannot generate empty sets. As a result, the smallest predictive set size is 1 for these methods and consequently empirical coverage for the bootstrap and conventional methods never declines below a given threshold. Finally, we note that the conformal approach yields better uncertainty quantification despite relying on fewer observations for model training (due to the data splitting procedure) when compared to the conventional and bootstrap approaches.

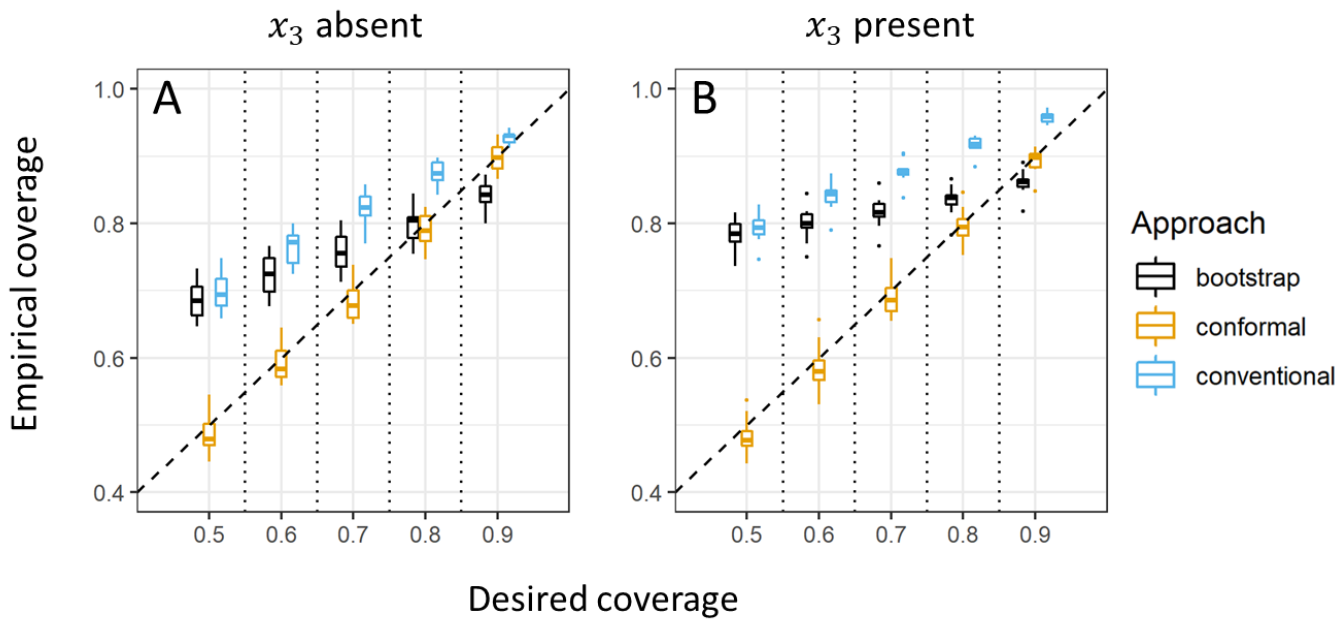

Fig. 4. The conformal approach performs well irrespective of the desired coverage (x-axes). These panels compare the conformal, conventional, and bootstrapping approaches regarding their empirical coverage (1:1 line is shown with diagonal dashed line). Panels A and B show the results for the classification scenario in which the predictor variable x_3 is absent and present, respectively.

5.3. Empirical results

We find that the LULC classes predicted by the random forest classifier display the expected spatial pattern of pastures close to the road network whereas forests are typically far away from roads (Fig. 5A). Furthermore, the river seems to be well delineated in this landscape.

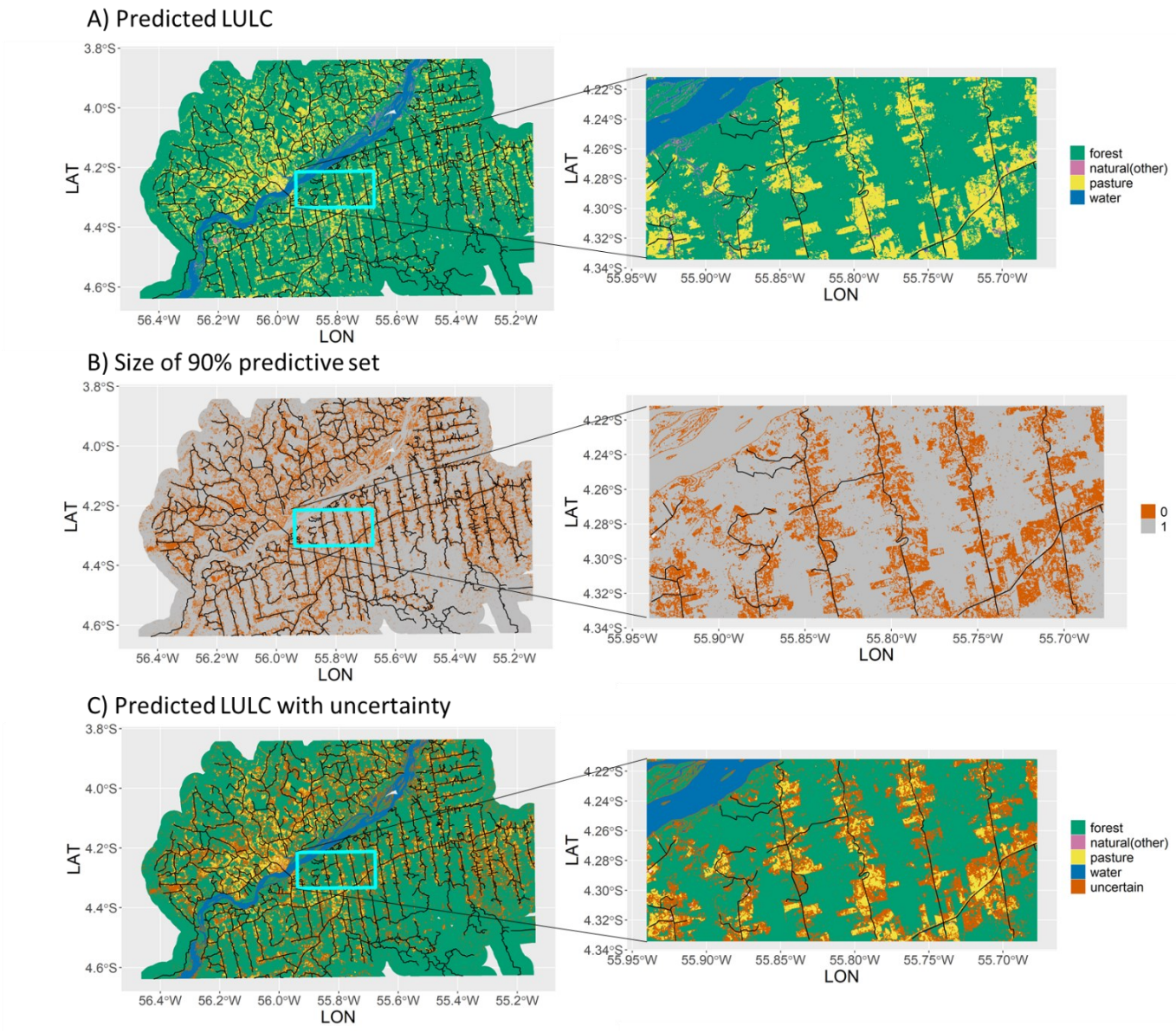

Fig. 5. LULC prediction for the study region ignoring uncertainty (panels A) or taking uncertainty into account (panels C). Panels C distinguish pixels with high level of uncertainty (i.e., empty predictive sets) as a separate “uncertain” class. Panels B show the size of the predictive set for each pixel. In all panels, the road network close to the Transamazon highway is displayed with black lines. Right panels show zoomed regions, depicted with cyan rectangles in left panels.

We calculate the confusion matrix based on our 10-fold spatial cross-validation (Table 3). These results reveal that the random forest algorithm resulted in 96% of correctly classified observations. Furthermore, we find that the classes, when ranked from best to worst (regardless if based on user or producer accuracy), were forest, water, pasture, and the natural (other) category. Our 10-fold spatial cross-validation also revealed that the empirical coverage of the 90% predictive sets generated by conformal statistics was on average equal to 95% whereas, when using 80% predictive sets, the empirical coverage was on average equal to 82% (Appendix 2). These results suggest that the conformal statistics approach is able to quantify uncertainty well when out-of-sample predictions are made.

		Actual LULC class					User acc. (%)
		Forest	Natural (other)	Pasture	Water	Total	
Predicted LULC class	Forest	3792	28	43	8	3871	98
	Natural (other)	6	43	20	4	73	59
	Pasture	18	20	230	0	268	86
	Water	2	4	0	125	131	95
	Total	3818	95	293	137	4343	

Table 3. Confusion matrix calculated based on the 10-fold spatial cross-validation exercise.

Interestingly, when applied to this dataset, conformal statistics generated either empty predictive sets or predictive sets with only a single LULC class (Fig. 5B). Furthermore, differently from the accuracy results in the confusion matrix, we find that water was the LULC class with the least classification uncertainty (i.e., the class with smallest proportion of pixels with empty predictive sets), followed by forest (Table 4). Importantly, although there is almost an equal number of uncertain pixels in forests and pastures, these pixels represent over half of the pasture pixels (Table 4 and Fig. 5C). These results suggest that there is substantial heterogeneity in pastures within the region, likely reflecting the gradient from well-maintained pastures without many trees or shrubs to abandoned pastures with over-

grown vegetation. Finally, both the confusion matrix and the conformal statistics results suggest that the class “natural (other)” had the most uncertainty.

LULC classes	Proportion of uncertain pixels	# uncertain pixels	# of pixels
Forest	0.13	1297073	9699045
Natural (other)	1.00	208573	208659
Pasture	0.54	1368957	2515110
Water	0.06	28350	458336

Table 4. Summary of the conformal statistics results across the landscape for each LULC class. Pixels deemed to be uncertain are those for which the predictive set is empty.

6. Discussion

In this article, we have introduced conformal statistics as a straightforward yet powerful approach to quantify pixel-level uncertainty in LULC classification. Using simulated data, we have shown that the size of the predictive set can be used as a measure of uncertainty. More specifically, pixels for which the predictive set is either empty or large (i.e., sets containing multiple classes) are uncertain pixels. We have also shown that this approach works better than bootstrapping both in terms of its simplicity (i.e., it does not require multiple model fits) and the ability to create predictive sets that have the desired coverage. Using an empirical dataset from the Amazon region, we show how this approach can generate insights regarding which LULC classes have low or high uncertainty (e.g., water and “natural (other)” class, respectively). These insights may or may not match those from a standard confusion matrix given that conformal statistics relies on both how likely the map labels agree with the true class and the estimated class probabilities. Finally, making these results available as a raster (either by directly incorporating into the LULC map as in Fig. 5C or as an additional uncertainty “band”) can enable

downstream users of LULC map products to consider which pixels to discard due to the presence of too much uncertainty.

Several approaches already exist to determine pixel-level uncertainty. For example, Hsiao and Cheng (2016) proposed a bootstrapping methodology to identify pixels with high classification uncertainty. In their approach, the bootstrap approach generates probability vectors for each pixel and pixels are deemed as unclassified if the maximum of these probabilities is below a user-determined threshold. This approach is computationally intensive and, as a result, might be challenging to implement for large-scale LULC products. Furthermore, the bootstrap approach only accounts for uncertainty in the input data (i.e., the fact that different training samples can potentially yield different predictions), failing to consider other sources of uncertainty (e.g., how well the model is able to predict individual LULC classes). The conformal statistics approach, on the other hand, is not computationally intensive and, despite not taking into account uncertainty in the input data, captures well the proportion of times that the predictive sets encompass the true (reference) classes.

The most similar approaches that we found in the literature were proposed by Park et al. (2016) and Khatami et al. (2017). They proposed to create an accuracy map by first labeling each pixel in the calibration dataset as 0 (if misclassified) and 1 (if correctly classified) and then using spatial and/or spectral information to interpolate these results in order to generate an accuracy map. One limitation of this approach relative to conformal statistics is that it does not provide information regarding which other classes are likely for the pixels with high probability of misclassification. Another limitation is that, because these approaches rely on models trained on the calibration data to create the accuracy map, it is possible that these models might fail to generalize well for out-of-sample data due to under or over-fitting. The conformal approach, on the other hand, has theoretical guarantees regarding the coverage

of the generated predictive sets (if the exchangeability assumption holds) because the model is only fitted to the training data whereas the calibration dataset is comprised of truly out-of-sample data.

A major benefit of conformal statistics is that it can be used to quantify uncertainty associated with any algorithm. By algorithm, we mean not only machine learning black-box classifiers (e.g., deep learning, random forests, and support vector machines) but also algorithms that rely on these classifiers and post-hoc rules. For example, post-hoc rules used in Mapbiomas include taking into account the local neighborhood of a pixel to avoid isolated pixels and pixel-specific time-series of LULC classes to ensure temporal consistency and eliminate prohibited LULC class transitions (Souza et al. 2020). Indeed, Manandhar et al. (2009) have shown that post-classification corrections like these can improve LULC classification accuracy. Because conformal statistics quantifies uncertainty after the full algorithm is applied, its results should be valid regardless of the exact details of the classifier and post-hoc rules, ultimately accounting for many of the different sources of uncertainty in LULC mapping described in the literature (Canibe et al. 2022).

A key parameter in conformal statistics is the desired coverage *C*. How should remote sensing practitioners choose *C*? The greater the coverage, the smaller the threshold for including labels in the predictive set, generally resulting in larger predictive sets and a higher number of pixels deemed to be uncertain. For example, it is easy to ensure that 100% coverage is achieved by creating predictive sets that contain all possible labels. We do not have specific guidelines for how to choose *C* because this decision fundamentally depends on the purpose of the analysis and inherent tradeoffs. Analyses that require pixels with little uncertainty could set C to a high value (e.g., 95%) and just use pixels for which the predictive set contains a single class. However, this procedure might also result in much fewer pixels

being available for analysis when compared to adopting the same procedure with C set to a lower value (e.g., 80%).

We believe that conformal statistics is likely to be even more useful if class-specific probabilities from large-scale LULC products are made available to users (e.g., as in the Google's Dynamic World LULC classification product; Venter et al. 2022). Aside from allowing better estimation of overall LULC area (Sales et al. 2022), maps with class-specific probabilities enable users to decide which coverage *C* to adopt to create their own customized uncertainty maps assuming they have access to the ground-truth data. Importantly, users can also create tailor-made uncertainty maps to their particular area using conformal statistics without requiring additional modeling or remote sensing work as long as local validation data are available. The ability to customize the uncertainty maps is important because users often have different needs and some classification errors might be more severe than others depending on these needs (Foody 2002; Stehman and Foody 2019). For example, users might wish to create their own definition of what constitutes a pixel that is too uncertain to be used. For example, for wildlife studies, a pixel that includes two very different vegetation types (e.g., forest and grassland) in its predictive set might be too uncertain to use. On the other hand, if many pixels have predictive sets containing grassland and pasture, users might judge these LULC classes to be sufficiently similar in terms of their vegetation cover and choose to lump these two classes into a single one. In this case, a pixel with these two LULC classes in its predictive set would not be considered too uncertain to be of use. Finally, it is possible that locally derived uncertainty maps might be more accurate for the region being studied than global uncertainty maps, ultimately improving map relevance (Stehman and Foody 2019).

Despite its promise, conformal statistics also has some important limitations. First, although conformal statistics does not have many assumptions, it nevertheless does rely on the key assumption of data

exchangeability (an assumption that is shared with most statistical and machine learning methods used for LULC classification). As a result, it is possible that if large spatial correlations are present, then the coverage guarantees of conformal statistics may not be accurate. One potential way to indirectly test this assumption is to quantify differences between the desired coverage C and the empirical coverage arising from a spatial cross-validation exercise. If large discrepancies arise, that could be an indication that the exchangeability assumption is being violated. Other approaches to more formally test the exchangeability assumption exist but they can be quite technical and are beyond the scope of this article (e.g., Fedorova et al. 2012; Ramdas et al. 2022). Additional research is clearly needed to determine the degree to which spatial correlation impacts the validity of conformal statistics results and to develop alternative conformal approaches to circumvent this problem.

Second, conformal statistics requires the splitting of data into a training and a calibration dataset but determining the best way to split the data remains to be determined and is an active area of research. Having more training data is critical to estimate well the class probabilities but having more calibration data is also important to generate well calibrated uncertainty estimates. Third, conformal statistics does not quantify the uncertainty associated with the training data and how the data are split into training and calibration data. Fourth, because of the need to split the data, an important limitation is that this approach is likely to only be suitable for situations with relatively large datasets (i.e., >1,000 observations). Finally, conformal statistics is an area of rapid development, with a wide range of conformal algorithms still being proposed in the literature. In particular, the conformal approach described here ensures marginal coverage (i.e., the true classes will lie within the predictive sets *C* proportion of the times across all observations in the validation dataset) but modelers increasingly want approaches that can provide conditional coverage (i.e., the true classes will lie within the predictive sets *C* proportion of the times for all observations that have a particular combination of predictor variables).

Developing conformal approaches that can ensure conditional coverage and that can take into account the variability in the training and calibration datasets is an important area of research.

We have focused on using conformal statistics for LULC classification but we note that this methodology is likely to be very useful for other remote sensing classification problems as well such as tree or wildlife species classification (e.g., Besson et al. 2022; Christin et al. 2019; Marconi et al. 2022; Oswald et al. 2022). The conformal approach that we have described is surprisingly simple (i.e., it does not require multiple model fits and can be implemented with just a few lines of code) and yet can generate predictive sets with the desired coverage (assuming the exchangeability assumption is met and that a large dataset is available) irrespective of the classification algorithm that is employed. For these reasons, we believe that conformal statistics has the potential to become a key approach in the toolkit of remote sensing scientists.

7. Acknowledgements

We would like to thank Dr. Gabriel Carrero for making the shapefile containing the road network surrounding the Transamazon highway available to us. This work was partly supported by the US Department of Agriculture National Institute of Food and Agriculture McIntire–Stennis project 1005163 and US National Science Foundation award 2040819 to DV. RI is grateful for the financial support of FAPESP (grant 2019/11321-9) and CNPq (grants 309607/2020-5 and 422705/2021-7).

8. References

Angelopoulos, A.N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*

Belgiu, M., & Dragut, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 24-31

Besson, M., Alison, J., Bjerge, K., Gorochowski, T.E., Hoye, T.T., Jucker, T., Mann, H.M.R., & Clements, C.F. (2022). Towards the fully automated monitoring of ecological communities. *Ecology Letters*, 1-23

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N.E., Linlin, L., & Tarko, A. (2020). *Copernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015-2019: Product User Manual*. Geneve, Switzerland

Canibe, M., Titeux, N., Dominguez, J., & Regos, A. (2022). Assessing the uncertainty arising from standard land-cover mapping procedures when modelling species distributions. *Diversity and Distributions, 28*, 636-648

Carrero, G.C. (2022). Frontier Heterogeneity : Development Processes in the Brazilian Amazon. In, *Geography*. Gainesville, Florida: University of Florida

Cheng, K.-S., Ling, J.-Y., Lin, T.-W., Liu, Y.-T., Shen, Y.-C., & Kono, Y. (2021). Quantifying uncertainty in land-use/land-cover classification accuracy: a stochastic simulation approach. *Frontiers in Environmental Science, 9*

Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Science, 118*, e2107794118

Christin, S., Hervet, E., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution, 10*, 1632-1644

Congalton, R.G., Gu, J., Yadav, K., Thenkabail, P., & Ozdogan, M. (2014). Global Land Cover Mapping: A Review and Uncertainty Analysis. *Remote Sensing, 6*, 12070-12093

D'Urso, G., & Menenti, M. (1996). Performance indicators for the statistical evaluation of digital image classifications. *ISPRS Journal of Photogrammetry and Remote Sensing, 51*, 78-90

Díaz, S., Settele, J., Brondízio, E.S., Ngo, H.T., Agard, J., Arneth, A., Balvanera, P., Brauman, K.A., Butchart, S.H.M., Chan, K.M.A., Garibaldi, L.A., Ichii, K., Liu, J., Subramanian, S.M., Midgley, G.F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R.R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K.J., & Zayas, C.N. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science, 366*, 1327

Fedorova, V., Gammerman, A., Nouretdinov, I., & Vovk, V. (2012). Plug-in martingales for testing exchangeability on-line. In, *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh, Scotland

561 Foody, G.M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of*
562 *Environment, 80*, 185-201

563 Foody, G.M. (2004). Thematic Map Comparison: Evaluating the Statistical Significance of Differences in
564 Classification Accuracy. *Photogrammetric Engineering & Remote Sensing, 70*, 627-633

565 Gao, H., Jia, G., & Fu, Y. (2020). Identifying and Quantifying Pixel-Level Uncertainty among Major
566 Satellite Derived Global Land Cover Products. *Journal of Meteorological Research, 34*, 806-821

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

569 Guerrero, J.V.R., Escobar-Silva, E.V., Chaves, M.E.D., Mataveli, G.A.V., Bourscheidt, V., de Oliveira, G.,
570 Picoli, M.C.A., Shimabukuro, Y.E., & Moschini, L.E. (2020). Assessing land use and land cover changes in
571 the direct influence zone of the Braco Norte hydropower complex, Brazilian Amazonia. *Forests, 11*

572 Guo, C., Pleiss, G., Sun, Y., & Weinberger, K.Q. (2017). On Calibration of Modern Neural Networks. In, *34*
573 *th International Conference on Machine Learning*. Sydney, Australia

574 Hsiao, L.-H., & Cheng, K.-S. (2016). Assessing uncertainty in LULC classification accuracy by using
575 bootstrap resampling. *Remote Sensing, 8*

576 Izbicki, R., Shimizu, G., & Stern, R.B. (2020). Flexible distribution-free conditional predictive bands using
577 density estimators. In, *Proceedings of the 23rdInternational Conference on Artificial Intelligence and*
578 *Statistics (AISTATS)*. Palermo, Italy

579 Izbicki, R., Shimizu, G., & Stern, R.B. (2022). CD-split and HPD-split: efficient conformal regions in high
580 dimensions. *Journal of Machine Learning Research, 23*, 1-32

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

583 Khatami, R., Mountrakis, G., & Stehman, S.V. (2017). Mapping per-pixel predicted accuracy of classified
584 remote sensing images. *Remote Sensing of Environment, 191*, 156-167

585 Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., & Wasserman, L. (2018). Distribution-free predictive
586 inference for regression. *Journal of the American Statistical Association, 113*, 1094-1111

587 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*, 18-22

588 Lyons, M.B., Keith, D.A., Phinn, S.R., Mason, T.J., & Elith, J. (2018). A comparison of resampling methods
589 for remote sensing classification and accuracy assessment. *Remote Sensing of Environment, 208*, 145-
590 153

Manandhar, R., Odeh, I.O.A., & Ancev, T. (2009). Improving the Accuracy of Land Use and Land Cover Classification of Landsat Data Using Post-Classification Enhancement. *Remote Sensing, 1*, 330-344

Marconi, S., Weinstein, B.G., Zou, S., Bohlman, S.A., Zare, A., Singh, A., Stewart, D., Harmon, I., Steinkraus, A., & White, E.P. (2022). Continental-scale hyperspectral tree species classification in the United States National Ecological Observatory Network. *Remote Sensing of Environment, 113264*

Maxwell, A.E., Warner, T.A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing, 39*, 2784-2817

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P.H.S., & Dokania, P.K. (2020). Calibrating Deep Neural Networks using Focal Loss. In, *34th Conference on Neural Information Processing Systems*. Vancouver, Canada

Nepstad, L.S., Gerber, J.S., Hill, J.D., Dias, L.C.P., Costa, M.H., & West, P.C. (2019). Pathways for recent Cerrado soybean expansion: extending the soy moratorium and implementing integrated crop livestock systems with soybeans. *Environmental Research Letters, 14*, 044029

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities With Supervised Learning. In, *Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany

Oswald, J.N., Erbe, C., Gannon, W.L., Madhusudhana, S., & Thomas, J.A. (2022). Detection and Classification Methods for Animal Sounds. In C. Erbe, & J.A. Thomas (Eds.), *Exploring animal behavior through sound: volumne 1*. Cham: Springer

Park, N.-W., Kyriakidis, P.C., & Hong, S.-Y. (2016). Spatial Estimation of Classification Accuracy Using Indicator Kriging with an Image-Derived Ambiguity Index. *Remote Sensing, 8*

Potapov, P., Hansen, M.C., Pickens, A., Hernandez-Serna, A., Tyukavina, A., Turubanova, S.A., Zalles, V., Li, X., Khan, A., Stolle, F., Harris, N., Song, X.-P., Baggett, A., Kommareddy, I., & Kommareddy, A. (2022). The global 2000-2020 land cover and land use change dataset derived from the ladsat archive: first results. *Frontiers in Remote Sensing, 3*, 856903

Ramdas, A., Ruf, J., Larsson, M., & Koolen, W.M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning, 141*, 83-109

Rausch, L.L., & Gibbs, H.K. (2021). The low opportunity costs of the Amazon soy moratorium. *Frontiers in Forests and Global Change, 4*, 621685

Romano, Y., Patterson, E., & Candes, E.J. (2019). Conformalized quantile regression. In, *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada

Romano, Y., Sesia, M., & Candes, E.J. (2020). Classification with valid and adaptive coverage. In, *34th Conference on Neural Information Processing Systems*. Vancouver, Canada

Sales, M.H.R., de Bruin, S., Souza Jr, C., & Herold, M. (2022). Land Use and Land Cover Area Estimates From Class Membership Probability of a Random Forest Classification. *IEEE Transactions on Geoscience and Remote Sensing, 60*

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research, 9*, 371-421

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

Stehman, S.V. (1997). Estimating standard errors of accuracy assessment statistics under cluster sampling. *Remote Sensing of Environment, 60*, 258-269

Stehman, S.V., & Foody, G.M. (2019). Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment, 231*

Sutton, P.C., & Costanza, R. (2002). Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. *Ecological Economics, 41*, 509-527

Tilman, D., Clark, M., Williams, D.R., Kimmel, K., Polasky, S., & Packer, C. (2017). Future threats to biodiversity and pathways to their prevention. *Nature, 546*, 73-81

Tucker, M.A., Bohning-Gaese, K., Fagan, W.F., Fryxell, J.M., Van Moorter, B., Alberts, S.C., Ali, A.H., Allen, A.M., Attias, N., Avgar, T., Bartlam-Brooks, H., Bayarbaatar, B., Belant, J.L., Bertassoni, A., Beyer, D., Bidner, L., van Beest, F.M., Blake, S., Blaum, N., Bracis, C., Brown, D., de Bruyn, P.J.N., Cagnacci, F., Calabrese, J.M., Camilo-Alves, C., Chamaille-Jammes, S., Chiaradia, A., Davidson, S.C., Dennis, T., DeStefano, S., Diefenbach, D., Douglas-Hamilton, I., Fennessy, J., Fichtel, C., Fiedler, W., Fischer, C., Fischhoff, I., Fleming, C.H., Ford, A.T., Fritz, S.A., Gehr, B., Goheen, J.R., Gurarie, E., Hebblewhite, M., Heurich, M., Hewison, A.J.M., Hof, C., Hurme, E., Isbell, L.A., Janssen, R., Jeltsch, F., Kaczensky, P., Kane, A., Kappeler, P.M., Kauffman, M., Kays, R., Kimuyu, D., Koch, F., Kranstauber, B., LaPoint, S., Leimgruber, P., Linnell, J.D.C., Lopez-Lopez, P., Markham, A.C., Mattisson, J., Medici, E.P., Mellone, U., Merrill, E., Mourao, G.D., Morato, R.G., Morellet, N., Morrison, T.A., Diaz-Munoz, S.L., Mysterud, A., Nandintsetseg, D., Nathan, R., Niamir, A., Odden, J., O'Hara, R.B., Oliveira-Santos, L.G.R., Olson, K.A., Patterson, B.D., de Paula, R.C., Pedrotti, L., Reineking, B., Rimmler, M., Rogers, T.L., Rolandsen, C.M., Rosenberry, C.S., Rubenstein, D.I., Safi, K., Said, S., Sapir, N., Sawyer, H., Schmidt, N.M., Selva, N., Sergiel, A., Shiilegdamba, E., Silva, J.P., Singh, N., Solberg, E.J., Spiegel, O., Strand, O., Sundaresan, S., Ullmann, W., Voigt, U., Wall, J., Wattles, D., Wikelski, M., Wilmers, C.C., Wilson, J.W., Wittemyer, G., Zieba, F., Zwijacz-Kozica, T., & Mueller, T. (2018). Moving in the Anthropocene: Global reductions in terrestrial mammalian movements. *Science, 359*, 466-469

661 Venter, Z.S., Barton, D.N., Chakraborty, T., Simensen, T., & Singh, G. (2022). Global 10 m Land Use Land
662 Cover Datasets: A Comparison of DynamicWorld, World Cover and Esri Land Cover. *Remote Sensing, 14*

663 Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer

664 Weber, K.T., & Langille, J. (2007). Improving Classification Accuracy Assessments with Statistical
665 Bootstrap Resampling Techniques. *GIScience & Remote Sensing, 44*, 237-250

666

667

668

List of Figure Captions

Fig. 1. Study region. Panel A displays the Brazilian Amazon (green polygon) and the study region (red rectangle). Panel B shows a false-color Landsat 8 mosaic (bands 4,3, and 2 were assigned to red, green, and blue, respectively) of the study region, created by calculating the median value per band of several 2018 images after removing pixels classified as cloud or shadow. Panel C zooms to a portion of the study region, shown with cyan rectangle in panel B. In all panels, the road network (obtained from Carrero 2022) is shown with black lines.

Fig. 2. Mean predictive set sizes that are greater or smaller than one indicate greater classification uncertainty. Panels A and B show how the probability associated with each class (used to simulate the data) changes as a function of the covariate x. Each line represents a different class. Panels C and D show the conformal statistics results, revealing how the mean size of the predictive sets (calculated by discretizing the covariate x into bins of width of 0.5) changes as a function of x. In all panels, the vertical grey lines show where the probability for each class peaks. Left and right panels show class probabilities and conformal statistics results for datasets with 3 and 5 classes, respectively.

Fig. 3. The conformal approach outperforms the conventional and the bootstrapping approaches. Empirical coverage is shown for 2 classification scenarios in which the x_3 predictor variable was either absent or present. The horizontal dashed line is the 80% desired coverage.

Fig. 4. The conformal approach performs well irrespective of the desired coverage (x-axes). These panels compare the conformal, conventional, and bootstrapping approaches regarding their empirical coverage

692 (1:1 line is shown with diagonal dashed line). Panels A and B show the results for the classification
693 scenario in which the predictor variable x_3 is absent and present, respectively.
694
695 Fig. 5. LULC prediction for the study region ignoring uncertainty (panels A) or taking uncertainty into
696 account (panels C). Panels C distinguish pixels with high level of uncertainty (i.e., empty predictive sets)
697 as a separate "uncertain" class. Panels B show the size of the predictive set for each pixel. In all panels,
698 the road network close to the Transamazon highway is displayed with black lines. Right panels show
699 zoomed regions, depicted with cyan rectangles in left panels.
700