

## Article

# How Response Designs and Class Proportions Affect the Accuracy of Validation Data

Julien Radoux <sup>1,\*†</sup>, François Waldner <sup>2,†</sup>  and Patrick Bogaert <sup>1</sup>

<sup>1</sup> Earth and Life Institute–Environment, Université catholique de Louvain, Croix du Sud 2, Louvain-la-Neuve B-1348, Belgium; patrick.bogaert@uclouvain.be

<sup>2</sup> CSIRO Agriculture & Food, Queensland Bioscience Precinct, 306 Carmody Road, St Lucia 4067 QLD, Australia; franz.waldner@csiro.au

\* Correspondence: julien.radoux@uclouvain.be

† These authors contributed equally to this work.

Received: 28 November 2019; Accepted: 7 January 2020; Published: 11 January 2020



**Abstract:** Reference data collected to validate land-cover maps are generally considered free of errors. In practice, however, they contain errors despite best efforts to minimize them. These errors propagate during accuracy assessment and tweak the validation results. For photo-interpreted reference data, the two most widely studied sources of error are systematic incorrect labeling and vigilance drops. How estimation errors, i.e., errors intrinsic to the response design, affect the accuracy of reference data is far less understood. In this paper, we analyzed the impact of estimation errors for two types of classification systems (binary and multiclass) as well as for two common response designs (point-based and partition-based) with a range of sub-sample sizes. Our quantitative results indicate that labeling errors due to proportion estimations should not be neglected. They further confirm that the accuracy of response designs depends on the class proportions within the sampling units, with complex landscapes being more prone to errors. As a result, response designs where the number of sub-samples is predefined and fixed are inefficient. To guarantee high accuracy standards of validation data with minimum data collection effort, we propose a new method to adapt the number of sub-samples for each sample during the validation process. In practice, sub-samples are incrementally selected and labeled until the estimated class proportions reach the desired level of confidence. As a result, less effort is spent on labeling univocal cases and the spared effort can be allocated to more ambiguous cases. This increases the reliability of reference data and of subsequent accuracy assessment. Across our study site, we demonstrated that such an approach could reduce the labeling effort by 50% to 75%, with greater gains in homogeneous landscapes. We contend that adopting this optimization approach will not only increase the efficiency of reference data collection, but will also help deliver more reliable accuracy estimates to the user community.

**Keywords:** validation; reference data; remote sensing; resolution; Accuracy assessment; response design; quality control; sub-sampling; overall accuracy

## 1. Introduction

Over recent decades, the number of openly available geographic data sets has tremendously increased along with their use in policy making, environmental monitoring, hazard prevention and scientific studies. It is of paramount importance that their quality is rigorously evaluated to inform users about their limitations and to limit contradicting results. Good practices in accuracy assessment include recommendations about (1) the sampling design that determine how many sampling units should be collected along with their locations; (2) the response design that defines the protocol for labeling each sampling unit; and (3) a rigorous estimation of accuracy using specific metrics [1–5].

A statistically rigorous assessment is thus a combination of a probability sampling design, appropriate accuracy estimators, and a response design chosen in accordance with features of the mapping and classification process.

Uncertainties linked to the sampling design and variance of the estimators are usually well quantified, but validation methods typically assume that reference data are error-free. In fact, the process of determining the so-called “ground truth” is seldom discussed in the literature and is often considered a straightforward—yet costly—task. Nonetheless, generating authoritative reference data sets remains a major challenge in accuracy assessment and it merits greater consideration in accuracy assessment [6]. Errors can indeed alter the process of generating reference data and even a small amount of errors can propagate and significantly impact the accuracy assessment [7–9]. There is thus a need for new methods that offer better control over the quality of reference data.

Practical constraints, such as poor accessibility to the sample locations on the ground, often affect the implementation of an ideal accuracy assessment protocol. For instance, this prompted Stehman [4] to propose criteria for quality and statistical rigor while taking, at the same time, practical utility (accessibility and reduced costs) into account. Another alternative is to replace ground observations with photo-interpreted very high-resolution images. Photo-interpretation by a group of experts with regional knowledge is often seen as the gold standard for reference data collection when dealing with large-area thematic products. However, photo-interpretation is not perfect—it typically reaches 80% accuracy—and it varies considerably among operators with accuracy levels ranging from 11% up to 100% [10]. Thematic errors ( $e$ ) affecting photo-interpreted samples can be divided in three categories: vigilance, systematic, and estimation errors:

$$e = e_{\text{vigilance}} + e_{\text{systematic}} + e_{\text{estimation}} \quad (1)$$

Vigilance errors, i.e., loss of performance after performing the same monotonous task over a long period, has been highlighted for a wide range of visual interpretation tasks [11]. Attitude, either optimistic or pessimistic, may determine how an operator will respond to training for vigilance [12]. Drops of vigilance, which are difficult to predict and manage for an individual interpreter, can be reduced by relying on more than one operator [13] either with consensual labeling or automated cross-validation.

Systematic errors ( $e_{\text{systematic}}$ ) occur when a photo-interpreter is incorrectly reading images. Reading of images and maps belongs to cartographic and visual literacy [14], a skill that changes over time and can be improved with the development of geospatial thinking. Image interpretation is a process that combines perception and cognition, both of which tend to facilitate identification (the cognitive task of identifying a pattern) and signification (the assignment of a meaning to a particular pattern [15,16]). The types of insight derived from imagery are strongly influenced by the interpreter’s expertise. Experts bring specialized knowledge, highly attuned perceptual skills and flexible reasoning abilities that novices lack [17]. There is however not always a strong relationship between the field work experience of operators and their photo-interpretation accuracy [18]. This might be explained by the dissimilarities between, on one side, air- and spaceborne images and, on the other side, panoramic images in at least three important aspects: (1) the portrayal of features from a downwards—often unfamiliar—perspective; (2) the use of wavelengths outside the visual portion of the spectrum; and (3) the depiction of the Earth’s surface at unfamiliar scales and resolutions [19]. The most capable interpreters have keen powers of observation, coupled with imagination and a great deal of patience [19]. Another individual factor potentially influencing image interpretation accuracy is search strategy. Compared to random search, training in systematic inspection caused higher performance [20]. Maruff et al. [21] have suggested that behavioral goals constrain the selection of visual information more than the physical characteristics of the information. This suggests that photo-interpreters with a search strategy based on previous experiences would be more successful at extracting relevant information than someone randomly searching for this information. Geographers would therefore be more successful than non-geographers during a single categorization round of aerial

photos [22]. Accordingly, crowdsourcing (i.e., when photo-interpreters are replaced by volunteers) is particularly prone to errors as it is open to anyone, regardless of the level of expertise of the volunteers. Systematic errors can thus largely be avoided by providing training to photo-interpreters, by selecting operators with local knowledge and by relying on multiple contributors [13].

Estimation errors ( $e_{\text{estimation}}$ ) arise when the class proportions within sampling units are imprecise even when all labels related to the sampling units are correct. These errors stem from three main factors: the number of sub-samples to label per sampling unit, the landscape structure, and the classification system. Imprecise estimates of the proportion of different classifiers for mixed or transitional classes reportedly account for most disagreements among photo-interpreters [23]. It has also been shown that the accuracy of the labeling as well as the accuracy of the image-based classification generally decrease when the sub-pixel heterogeneity increases [24]. Contrary to the systematic and vigilance errors, there is currently no mechanism to control estimation errors. Therefore, even when best practices in quality control are implemented (i.e.,  $e_{\text{systematic}} = 0$  and  $e_{\text{vigilance}} = 0$ ), uncertainties in the photo-interpreted labels remain due to estimation errors. If left unchecked, these estimation errors can bias reference data and subsequent accuracy assessments for they are intrinsically linked to the complexity of the sub-pixel landscape structure. Here, we propose that estimation errors need to be managed in the response design.

The objective of this paper is to untangle the intricate interplay between classification systems, response designs and landscape fragmentation with regards to the estimation errors. Specifically, we (1) quantify the impact of imprecise estimation of land-cover proportions on the accuracy of reference data; and (2) propose a response design that optimizes the labeling effort. We particularly focused on two aspects of response designs for binary and multiclass majority classification systems: their structure (point-based vs. partition-based designs) and the labeling effort (the number of sub-samples to be labeled per sampling unit). Because the three components of the errors (Equation (1)) are strongly intertwined in real case studies, we relied on synthetic ground truth data which gave full control over the sampling strategy and allowed us to isolate estimation errors. Synthetic ground truth is indeed necessary to ensure that the actual properties of the data were known and to exclude effects due to other sources of ground reference data error [25–27]. Our main contributions can be summarized as follows:

- We provide an in-depth review of the different types of response designs and their applications;
- We analyze the performance of response designs for different types of classification systems;
- We generalize case-specific results using indices of landscape composition;
- We optimize the sampling effort with adaptive response designs that leverage the confidence intervals of the estimated proportions.

## 2. Background

### 2.1. Types of Classification System

There is no single ideal classification system and it is unlikely that one could be developed [28]. Many classification systems have therefore been developed depending on the purpose of the map and the scale of the analysis. Congalton et al. [29] suggest that all classes of a classification system should be clearly determined at the start of mapping projects, which is not always the case [30]. Those classification systems can be divided into three categories: semantic, aggregative and continuous.

A semantic classification system provides a formal description of the classes, including properties and relationships with their sub-parts. They are usually preferred for describing spatial entities at large scales like e.g., trees or buildings. Spatial entities unambiguously described by a semantic classification system can in turn be used as diagnostic criteria also referred to as classifiers in the context of classification systems [31] to build aggregative classification systems. In some cases, semantic classification systems can be used to describe spatial regions instead of spatial entities [32]. In this case, the semantic refers to a specific meaning that encompasses a large set of properties and relationships,

e.g., a city (land use type) or a savanna (ecosystem type). To avoid ambiguity, all classes should be precisely described by an ontology that includes a representation, a formal naming, and a definition of their properties and relations [33].

When the spatial resolution of the image becomes coarser than the size of the spatial entities, pixels can either represent a continuous field with the proportions of all (or a subset of) the semantic classifiers [34], or can be described by an aggregative categorical classification system. For the latter, the proportion of spatial entities is computed, and decision rules are applied to define the classes at the coarser resolution [35]. These decision rules are usually based on a majority rule or on fixed threshold values for class proportions.

Majority classification systems are the most prevalent type of aggregative classification systems in land-cover mapping. For instance, all the first global land-cover maps used the International Geosphere-Biosphere Programme (IGBP) classification system [30]. It is often implicitly assumed that labels correspond to the class of the prevalent spatial entity within each mapping unit. The boundary between purely semantic and majority classification systems is therefore fuzzy. The drawbacks are (1) that majority is undefined when multiple classes are equally dominant; and (2) that no information about the actual class proportions within the mapping units is conveyed to users. For example, a majority label in a 10-class classification system could be assigned to a class covering between 10% (all classes present in the same proportion) and 100% (“pure” class) of the area of the mapping unit.

Threshold-based classification systems are also a widely used type of aggregative classification systems. They rely on a set of rules to partition the feature space of the classifiers’ proportions. Those rules introduce sharp boundaries to the continuous field of the classifiers’ proportions to obtain a limited number of categorical classes. The number of those classes usually exceeds the number of classifiers in the mapping area. Threshold-based classification systems should be defined with consistent classification systems such as the Land-Cover Classification System LCCS [31]. The LCCS guarantees no overlap between classes and a full description of the possible combinations of classifiers and is considered by Grekousis et al. [30] as the only universally applicable classification system. LCCS is used for coarse to medium-resolution global land-cover maps [36,37], but also for high-resolution object-based classification [32]. They provide a larger thematic precision (more classes from the same set of classifiers) than majority-based classification systems, but they require definition of (often arbitrary) thresholds and conditional statements to draw the boundaries between the classes. This may lead to difficult naming conventions when the feature space of classifiers is large, and the rules become complex.

Binary classification systems are a particular case of aggregative classification systems that indicates whether a given classifier is present or absent inside a given pixel. Several examples include global crop/non-crop [38], water/non-water [39] or forest/non-forest [40] maps. The labeling is then defined by a proportion threshold. Most of the time, this threshold is set to 50%, which is then nearly equivalent to a majority classification system for two classes, but can be unambiguously defined at the proportion of 50% (unlike the majority). There are however examples where the threshold is not 50%, e.g., 10% for a forest/non-forest map in arid regions [41].

## 2.2. Structures of Response Designs

The values of the accuracy parameters are strongly affected by the protocols implemented for the response design. This includes the choice of spatial units and how within-unit homogeneity is addressed when assigning class labels [42]. Generally, the sampling unit is the elementary unit that is labeled according to the classification system of the map. In this study, the sampling unit is a pixel and there is an agreement when the map classification and the response design converge to the same label. The studied response design is therefore crisp, as opposed to alternative response designs where the proportions of spatial entities are used to validate categorical variables, e.g., fuzzy validation. Best practices in accuracy assessment suggest that photo-interpretation should rely on images of finer

resolution than the map being validated. With finer resolution data, the sampling unit could appear as heterogeneous. Three protocols can then be used to assign a reference value to those sampling units:

**Direct assignment:** a single label is directly assigned to the sampling unit; this is the most common response design (see for instance Perger et al. [43]);

**Point-based sub-sampling:** the sampling unit is sub-sampled by a set of points that are individually labeled by the interpreter before automated aggregation with decision rules (see for instance Bastin et al. [41], Bey et al. [44]);

**Partition-based sub-sampling:** the sampling unit is partitioned into sub-parts that are labeled individually by the interpreter before automated aggregation with decision rules (see for instance Bayas et al. [45], Waldner et al. [46]).

Direct assignment is the fastest method because a single class is assigned for each sampling unit by looking at it as a whole. However, it is strongly dependent on the operators' skills and on their level of concentration, it is poorly inter-operable because there is no information about classifiers proportion that would help to apply other labeling rules, and the confidence of the labeling must be provided by the operator. In this study, we focused on the two designs that involve sub-samples, namely point-based and partition-based designs (Figure 1) because our primary assumption was that a large part of the labeling errors could be quantified from the selected response design, i.e., independently from the operators.



**Figure 1.** Two types of response designs compared in this study: point-based and partition-based designs. Grassland are in green, impervious surfaces are in red and ploughed bare soil are in yellow. Both methods would agree on most of the grassland, as well as the direct assignment of the majority class without looking at sub-samples, but it is worth noting that the proportions of impervious soil and grassland are not the same.

### 2.2.1. Point-Based Response Designs

In point-based response designs, photo-interpreters label a set of points within every sampling unit. The final class is then assigned based on the proportions of the number of points for each observed category. By definition, points are dimensionless but, in practice, the photo-interpretation is limited by the spatial resolution of the reference image. Nevertheless, even if the vicinity of points provides contextual information as a part of the photo-interpretation process, the label is defined at the precise location of the point.

Randomly selecting a set of points inside each sampling unit inherits the same properties as for the sampling designs of a map as a whole. Systematic sampling is therefore often the most efficient [47]. However, if strong periodicity in the spatial pattern of the landscape is suspected, systematic sampling should be avoided unless sufficient information is available to avoid the phasing between spatial pattern and sampling interval [48].



### 2.2.2. Partition-Based Response Designs

Contrary to their point-based counterparts, partition-based response designs provide exhaustive coverage of the sampling unit (Figure 1). The counterpart is that it implies a discretization of the landscape, which could result in inaccurate labels. In practice, photo-interpreters are tasked with estimating the proportion of each class within each sub-sample, based on their entire content. The final label is automatically attributed from the estimated class proportions following a set of rules that are specific to the classification system. These rules are applied in a two-step process: the first step is performed by a photo-interpreter who assigns a label to each sub-part, and the second step consists of aggregating the labels of the sub-parts to attribute a final label, which can be automated. While square sub-parts are the most widespread type of partition [45], irregular polygons can also be used [46]. In the latter case, accurate delineation of the polygons plays a major role in the reliability of the response design.

In the case of binary classification, there are two approaches to define the sampling unit labels. The first approach, hereafter referred to as Threshold-Then-Majority (TTM), applies the labeling rule for each sub-sample level, then determines the majority label among sub-samples of the whole sampling unit. The second protocol, hereafter referred to as Majority-Then-Threshold (MTT), starts by determining the majority class inside the sub-samples, then applies the labeling rule with threshold at the level of the sampling unit. These methods are identical when the binary threshold is equal to 50% but differ otherwise. These two frameworks could be applied to LCCS-like multiclass classification systems as well. In the case of majority classification system, they both simplify into a two-stages application of the majority rule. TTM is expected to work best in fragmented landscapes and MTT for large homogeneous patches.

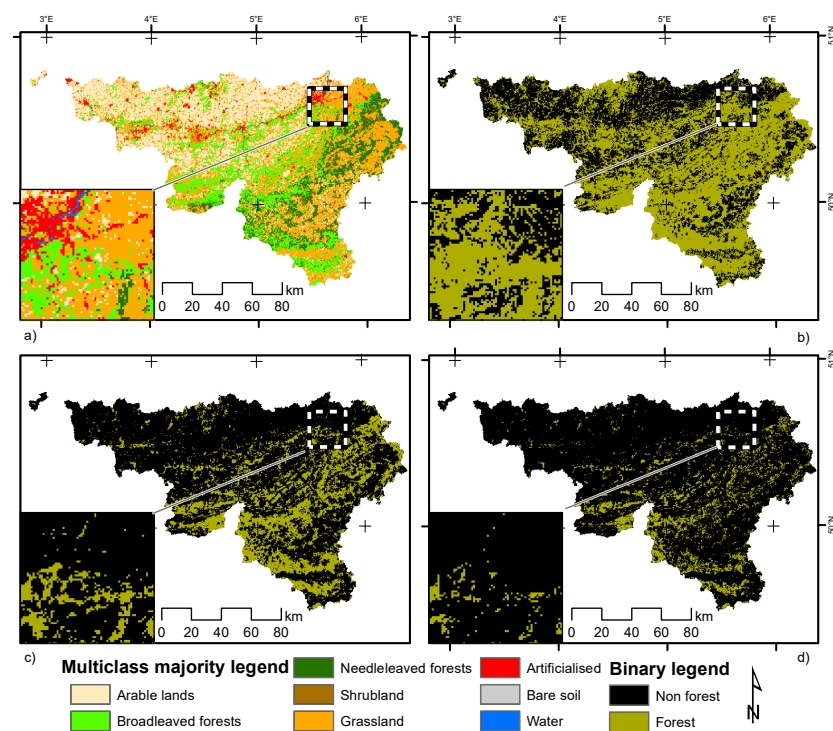
## 3. Data

### 3.1. Generating Sampling Units and Sub-Samples from a Reference Land-Cover Map

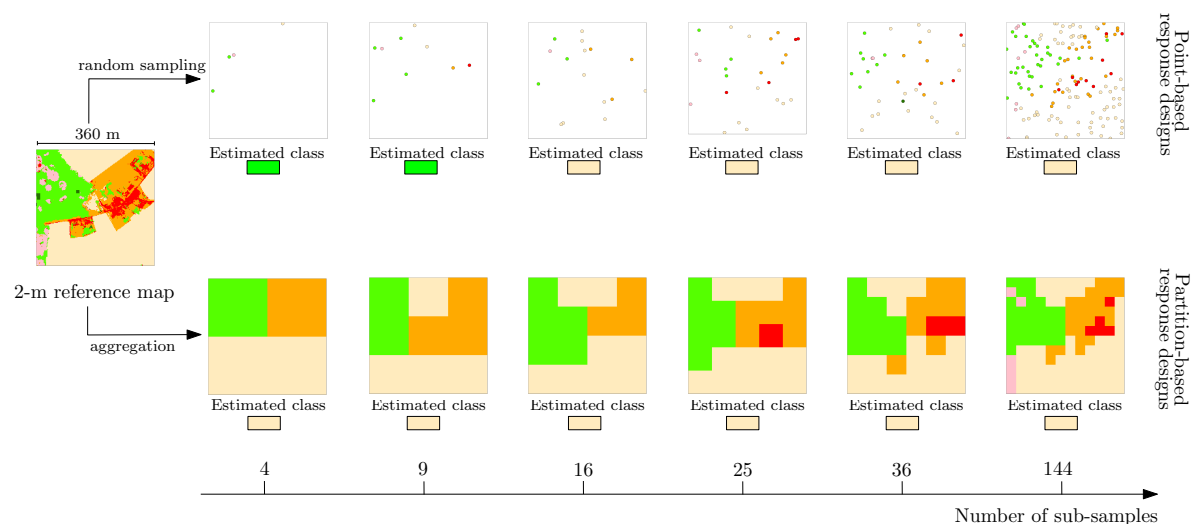
The main contribution of this study is a theoretical framework to maximize the accuracy of photo-interpreted reference sampling units while minimizing the photo-interpretation effort. Testing the performance of response designs for different types of classification systems and optimizing the sampling effort requires a wall-to-wall ground truth data set. Because wall-to-wall ground truth data are unavailable in practice, we used a fine resolution map from which we derived synthetic maps at different spatial resolution, thereby retaining full control over the labeling process and on the sampling protocols.

Without loss of generality, we considered a 2-m land-cover map as ground truth (Figure 2). The original map covered the Walloon Region (approximately 16,500 km<sup>2</sup>), Belgium, and includes ten land-cover classes with an estimated overall accuracy of 93% [49]. The two marginal grassland classes were merged with the agricultural grassland because of their scarcity. Using a 2-m resolution map provides unambiguous labeling of classifiers because the size of the pixels is inferior to the size of standard objects in this landscape. In case of mixed pixels located at the boundary of two spatial objects, the label was chosen based on the pixel's centroid.

In this experiment, pixels in the map at 360-m resolution corresponded to sampling units and their labels were considered to be the target values for the different response designs. The high-resolution data set was aggregated at the resolution of 360 m using the different labeling rules based on the pixel counts (with several 32,400 2-m pixels per 360-m cell; Figure 3). While 360 m does not correspond to the spatial resolution of any current satellites, it is similar to medium-resolution satellite such as PROBA-V, MODIS, Sentinel-3 or VIIRS and it has the advantage of being factorizable by a diversity of integers (2, 3, 4, 5, 6, 9, 12, 15, 20, ...), thereby allowing partitions of a wide variety of size of squares used as sub-samples.



**Figure 2.** Maps of the study area based on majority classification system (a) and the different binary classification systems for the forest classes: 10% for top right (b); 50 % for bottom left (c) and 75% for bottom right (d).



**Figure 3.** Procedure to generate sampling units and sub-samples from the 2-m reference map. The colors correspond to the multiclass legend of Figure 2.

### 3.2. Types of Classification Systems

Two types of aggregative classification systems were compared in this study: a multiclass majority classification system and a binary classification system for three thresholds (Figure 2). For the majority classification system, the label of the most frequent class within the mapping unit was selected. There are eight classes in the 2-m map so there are also eight classes at the aggregated level. For the binary classification system, the map represents the presence or absence of a specific class. When pixels are larger or similar in size to the spatial objects or the spatial regions of interest, an arbitrary threshold on

the proportion of the class becomes necessary to handle mixed pixels. In this study, we considered maps of forests (broad-leaved and needle-leaved forests all together). Three different thresholds of crown cover have been chosen, namely 10% (FAO's forest definition) [50]; 50% (the most commonly used threshold); and 75% (threshold used for closed canopy forests).

#### 4. Methods

We sought to answer the following questions:

1. What is the accuracy of point-based and partition-based response designs for different number of sub-samples in a realistic case study?
2. How can the accuracy of response designs be predicted based on landscape structure indices?
3. How to optimize the number of sub-samples per sampling unit?

We addressed these questions in three successive steps for the four classification systems described in Section 3.2. First, the accuracy of the various response designs was compared with simulated sampling across the study site (Section 4.1). Second, we generalized the relationship between the error rate and the underlying landscape of sampling units (Section 4.2). We finally proposed a method that iteratively adds sub-samples to label until the estimated class proportions driving the labeling process reach the desired confidence level (Section 4.3).

This optimization method is formulated for point-based designs only, as theoretical confidence intervals are not available for partition-based designs and labels cannot be reused when the number of partitions is increased. In fact, optimizing the partition-based designs depends on the ability of the operator to decide the appropriate number of sub-samples. As we assumed perfect operators throughout this paper, the question of optimizing partition-based designs falls beyond the scope of this paper. Nonetheless, the impact of photo-interpretation errors on the response design is discussed in Section 6.

##### 4.1. Accuracy of Point-Based and Partition-Based Response Designs

Our approach to empirically quantify the accuracy of response designs was based on a Monte Carlo framework. For every sampling unit, we repeatedly estimated the class proportions of ground truth for a range of sub-sampling efforts. The labels of the randomly simulated response designs were then assigned using the decision rules of the four classification systems. The same decision rules were applied on the true proportions (i.e., computed from the 2-m reference map) to derive the true label. For each iteration, the error rate was computed by dividing the number of disagreements between the true and the simulated labels of the sampling units, with

$$\text{Error rate} = \frac{\text{Number of erroneous labels}}{\text{Number of sampling units}} \quad (2)$$

For the point-based response designs, the sub-sample selection was repeated 36 times. Sub-samples were selected by simple probabilistic sampling of the 2-m pixels located inside each sampling unit.

For partition-based response designs, the 36 realizations were generated by shifting the origin of the grid by six multiples of 11 pixels (22, 33, 44, 55, 66, 77) in both the  $x$  and  $y$  directions (sampling units that are not completely inside the study area were discarded). Spatial resampling of the 2-m reference map was performed at intermediate spatial resolutions of 180, 120, 90, 72, 60, 36 and 30 m, which correspond to a partitioning in 4, 9, 16, 25, 36, 100 and 144 squares, respectively. These resolutions were constrained by the availability of integer divisors of 180. For the sake of comparison, the same numbers of sub-samples were used for the point-based approach. In the TTM case, the proportion of the forest class was computed within every intermediate resolution pixel, which were then labeled as forest or non-forest according to the threshold value. Those pixels were then resampled at the spatial resolution of 360 m with a majority rule to select the final label. In the MTT case, the forest label was assigned to each sub-sample where forest was the majority class. The proportion of forest pixels was



then computed for each 360 m pixel and the final forest/non-forest label was assigned based on the selected threshold. For the majority classification system, the majority class was first identified for each sub-sample, then the majority of the sub-sample labels was assigned to the sampling unit.

#### 4.2. Impact of Landscape Fragmentation

We sought to evidence the link between landscape structure and response design to predict the response design accuracy in other landscapes where *prior* structure knowledge is available. We therefore selected two landscape metrics, one per type of classification system, which can be easily computed for any areal sampling unit and any scale.

For binary classification systems, we characterized landscapes by reporting the proportion of our main class within each sampling unit. Because there is only one degree of freedom with two classes, the choice of the main class does not influence the reasoning, hence the forest class was arbitrarily selected, with:

$$p_m = \frac{S_{\text{forest}}}{S_{\text{total}}} \quad (3)$$

where  $S_{\text{forest}}$  is the area of forest (more precisely in this case, tree crown cover) inside the sampling unit, and  $S_{\text{total}}$  is the area of the sampling unit.

For multiclass classification systems, we opted for the Equivalent Reference Probability ( $\epsilon$ ) [51]. Rooted in information theory, the equivalent reference probability is particularly interesting because it accounts for the full set of probabilities and remains consistent with the maximum probability, unlike entropy. Given  $\mathbf{p} = (p_1, \dots, p_k)$  the vector of the class proportions in the landscape,  $k$  the number of classes and  $i^*$  the index of the dominant class, the equivalent reference probability is

$$\epsilon = \frac{\exp(E[D(i|i^*)])}{\exp(E[D(i|i^*)]) + k - 1} \quad (4)$$

where  $E[D(i|i^*)]$  is the expected difference of information as described in equation :

$$E[D(i|i^*)] = \ln p_{i^*} - \frac{1}{1 - p_{i^*}} \sum_{i \neq i^*} p_i \ln p_i \quad (5)$$

with  $p_{i^*}$ , the proportion of the majority class. Class purity and  $\epsilon$  were computed for each sampling unit based on the true proportions.

Average error rates of the response designs were estimated for the full range of possible  $p_m$  and  $\epsilon$  values with a step of 0.05. For visualization purposes, the error rate was smoothed by fitting local regressions LOESS [52].

#### 4.3. Local Optimization of the Number of Sub-Samples

When collecting validation data, the structure of the landscapes covered by the sample units is generally unknown, so that the optimal number of sub-sampling units cannot be estimated *a priori* from relationships between accuracy and landscape fragmentation. However, thanks to the interactivity of the Web 2.0, online validation platforms can be tailored to compute and update class proportion estimates as soon as sub-samples are labeled by photo-interpreters. This part of the study aimed to optimize the number of sub-samples needed for reaching a certain level of accuracy, resulting in an optimal response design that minimizes costs and/or time constraints. We propose to define an optimal number of sub-samples for each sampling unit based on the confidence intervals of the estimated sub-sample class proportions. Here, the confidence levels were set to 99.9% to illustrate the stringent requirements of building authoritative reference data sets, and to 90% to illustrate the required effort for collecting reference data under constrained conditions.

In practice, the local optimization process consisted of randomly selecting an initial set of nine sub-samples and assessing the corresponding confidence level. Sub-samples were then added one at

a time until the confidence on the estimated proportions reached the desired confidence. For binary classification systems, the confidence interval (for a given confidence level) around the estimated proportion must not include the threshold value that divides the study area in the two binary classes. For the multiclass majority classification system, the confidence interval around the estimated proportions of the majority class must not include the estimated proportion of the second most frequent class.

For binary classification systems, a given sampling unit is correctly labeled if the estimated proportion is on the same side of the threshold value as the true proportion. In practice, the proportion of the sampled area is unknown. However, the probability of assigning the correct label can be estimated based on the estimated value of the binomial distribution.

The confidence interval around estimated class proportions or accuracy indices is usually estimated using a Normal approximation

$$CI_{LB} = m/n - z_{\alpha} \sqrt{(m/n * (1 - m/n)) / n} \quad (6)$$

$$CI_{UB} = m/n + z_{\alpha} \sqrt{(m/n * (1 - m/n)) / n} \quad (7)$$

where  $CI_{LB}$  is the confidence interval lower bound,  $CI_{UB}$  is the confidence interval upper bound,  $n$  is the number of sub-sampling units,  $m$  is the number of points belonging to the class label selected by the decision rule,  $z$  is a percentile from the standard normal distribution and  $\alpha$  is the percent chance of making a Type I error (so that  $1 - \alpha$  is the confidence level).

However, the two main hypotheses of the Normal approximation are not respected in our incremental case: the number of points is small and proportions close or equal to 1 (pure pixels) are likely to be observed. The Clopper-Pearson exact confidence interval (CI) was therefore used instead of the Normal approximation [53], with

$$CI_{LB} = 1 - \text{BetaInv}\left(\frac{\alpha}{2}, n - m, m + 1\right) \quad (8)$$

$$CI_{UB} = 1 - \text{BetaInv}\left(1 - \frac{\alpha}{2}, n - m + 1, m\right) \quad (9)$$

where  $CI_{LB}$  is the confidence interval lower bound,  $CI_{UB}$  is the confidence interval upper bound,  $n$  is the number of sub-sampling units,  $m$  is the number of points belonging to the majority class,  $\alpha$  is the percent chance of making a Type I error, and  $1 - \alpha$  is the confidence level. Those parameters are taken by the BetaInv function, which computes the inverse of the beta cumulative distribution function.

Exact confidence intervals are not available for multinomial cases. Several approximations have been proposed [54–56]. Simultaneous confidence interval estimates from Goodman [56] were selected because preliminary tests revealed that in a binomial case, it provides a closer match to the Clopper-Pearson interval than other alternatives. For a multinomial distribution  $\mathbf{p}$ , Goodman's simultaneous confidence interval for the  $i^{th}$  class is given by

$$CI_i = \frac{b + 2p_i \pm \sqrt{b[b + 4p_i(n - p_i)/n]}}{2(n + b)}; i = 1, \dots, k \quad (10)$$

where  $p_i$  is the proportion of class  $i$ ,  $n$  is the total number of samples and  $b = \chi^2_{1-\alpha/k}(1)$ , the  $1 - \alpha/k$  quantile of the chi-square distribution with one degree of freedom.

In some cases, e.g., where the observed class proportion is equal to the arbitrary threshold in a binary classification or when several classes have the same proportion in the case of majority rule, the number of points to meet the required confidence could grow infinitely. Therefore, the maximum number of points was arbitrarily set to 144. This process was repeated 25 times to compare the theoretical confidence levels with the observed accuracy and to estimate the average number of sub-samples needed for each sampled area.

## 5. Results

### 5.1. Impact of Response Design and Sampling Effort on Accuracy of the Labels

Overall, our results highlight the relatively large uncertainty linked with the response designs for all types of classification systems in the study area. In addition, the average error rate is not only linked with the sampling effort, but also depends on the combination of the classification systems and the type of response design (Table 1).

**Table 1.** Average error rate in reference labels across the study area for point-based and partition-based designs. The binary maps on which the test was performed define the forest class based on the percentage of tree cover. The other type of map (called majority) includes 9 land-cover (LC) classes assigned based on the class that covers the largest area inside the sampling unit. The two types of partition-based designs are identical when the classification system is a majority rule.

| Classification System                                  | Average Error Rate (%)                  |      |      |      |      |      |      |      |
|--------------------------------------------------------|-----------------------------------------|------|------|------|------|------|------|------|
|                                                        | Number of Sub-Samples Per Sampling Unit |      |      |      |      |      |      |      |
|                                                        | 4                                       | 9    | 16   | 25   | 36   | 81   | 100  | 144  |
| <i>Point-based design</i>                              |                                         |      |      |      |      |      |      |      |
| Binary (forest > 10%)                                  | 13.7                                    | 10.6 | 7.8  | 6.5  | 5.5  | 3.8  | 3.5  | 2.9  |
| Binary (forest > 50%)                                  | 13.0                                    | 8.5  | 6.3  | 5.1  | 4.2  | 2.8  | 2.5  | 2.1  |
| Binary (forest > 75%)                                  | 17.1                                    | 11.5 | 9.3  | 7.6  | 6.5  | 4.4  | 4.0  | 3.4  |
| Majority (9 LC classes)                                | 27.2                                    | 18.7 | 14.1 | 11.5 | 9.5  | 6.4  | 5.7  | 4.8  |
| <i>Partition-based design: threshold-then-majority</i> |                                         |      |      |      |      |      |      |      |
| Binary (forest > 10%)                                  | 8.6                                     | 9.7  | 12.0 | 13.9 | 15.5 | 19.0 | 19.9 | 21.4 |
| Binary (forest > 50%)                                  | 4.4                                     | 2.2  | 2.1  | 1.6  | 1.3  | 1.3  | 1.3  | 1.2  |
| Binary (forest > 75%)                                  | 4.5                                     | 4.2  | 5.0  | 5.5  | 6.0  | 7.0  | 7.2  | 7.7  |
| <i>Partition-based design: majority-then-threshold</i> |                                         |      |      |      |      |      |      |      |
| Binary (forest > 10%)                                  | 20.3                                    | 12.5 | 12.9 | 12.1 | 10.9 | 8.3  | 8.0  | 7.0  |
| Binary (forest > 50%)                                  | 4.8                                     | 2.2  | 2.3  | 1.6  | 1.7  | 1.3  | 1.3  | 1.2  |
| Binary (forest > 75%)                                  | 10.0                                    | 6.5  | 6.3  | 5.2  | 5.1  | 4.3  | 4.2  | 4.0  |
| <i>Partition-based designs</i>                         |                                         |      |      |      |      |      |      |      |
| Majority (9 LC classes)                                | 10.6                                    | 6.8  | 5.8  | 4.9  | 4.4  | 3.5  | 3.4  | 3.1  |

For any sub-sample size, the most reliable labels are obtained for the binary classification system with a threshold at 50% for both partition-based and point-based response designs. For the other classification systems, the ranking of the ease of validation differs across response designs. For instance, the second most consistent labeling is obtained for the majority classification system with a partition-based design, while the binary classification system with 10% threshold ranks second for point-based validation. For the same classification system, the partition-based response design performs poorly, with an error rate of 12% for 25 sub-samples.

The average error rates of point-based designs markedly decrease between 4 and 100 sub-samples (Table 1). This trend is observed across the four classification systems. With only four points, error rates are >15%. The error rates then drop to <2.5% for sub-sample sizes larger than 100 in the case of threshold-based classification systems. The decreasing error rate with respect to the sampling effort is also observed for the majority classification system, but the improvement is smaller (6% error with 100 sub-samples). In comparison, the other binary classification systems provide more correct labels (4% error with 100 sub-samples for the 75% binary classification system), with the most accurate labeling obtained from the binary 50% classification system (3.5% error).

The two types of partition-based response designs exhibit an opposite behavior for the binary thresholds of 10% and 75%. In those two cases, the error rates of a perfect operator increase in TTM

(but decrease in MTT) for increasing numbers of sub-samples. The binary classification system at 50% yielded similar results for MTT and TTM, with slightly better results from the TTM approach. It reaches 98.4% accuracy with 25 sub-samples. The majority classification system fails to generate labels with less than 3% error when using less than 144 sub-samples, and achieves less than 5% errors starting from 25 sub-samples (Table 1).

Our results show that the most efficient response design depends on the classification system. Given the spatial resolution of the sampling units and the relatively fragmented landscape of the study area, the partition-based response design outperformed point-based response design for the majority classification system. With the latter, 25 sub-samples were necessary to achieve 95% of accuracy. The validation effort required for binary classification systems depends on the threshold value. The least effort was required with a threshold of 50% and a partition-based model (95.2% with only 4 sub-samples). On the contrary, point-based response design outperformed partition-based response designs for the 10% threshold. This classification system was the most difficult to validate in the study area—36 sub-samples were needed to reach at least 95% accuracy.

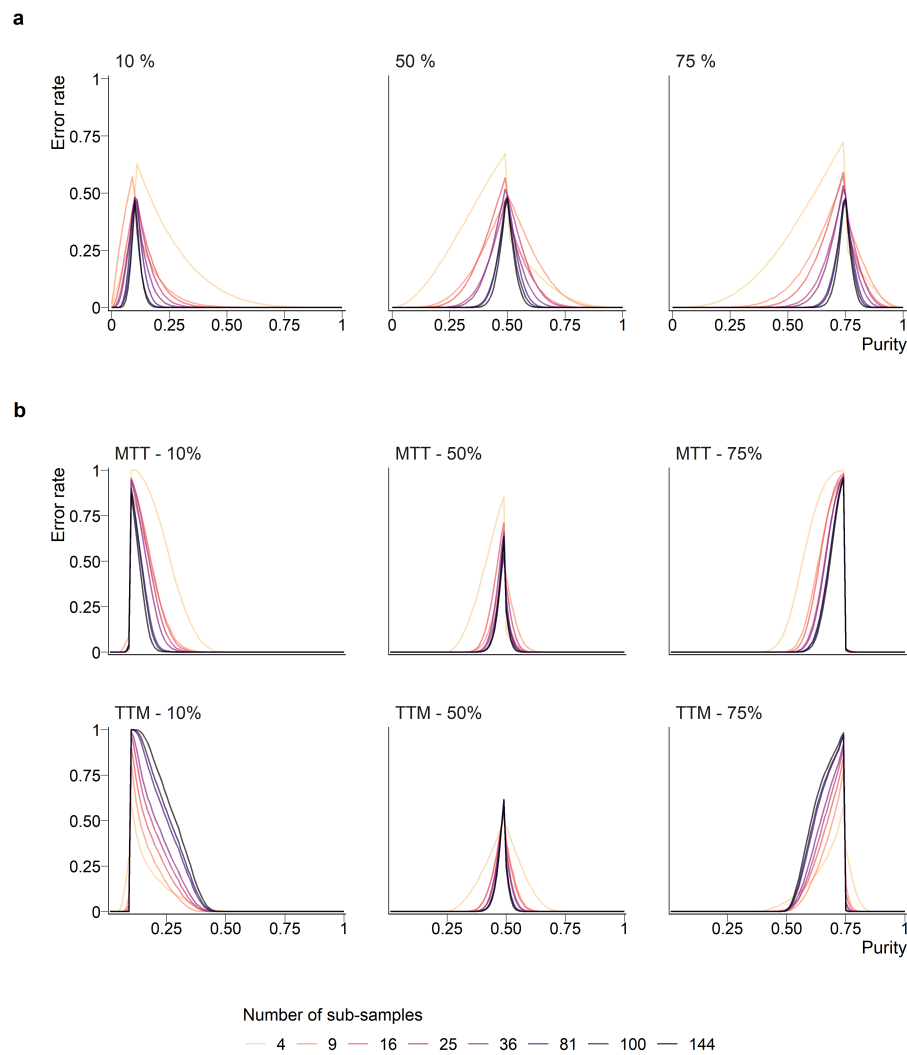
## 5.2. Relationship Between Sampling Unit Heterogeneity and Accuracy

Heterogeneity indices allow us to generalize the overall error rates estimated on the study area. The selected heterogeneity indices, which are independent of the landscape and spatial resolution, highlight the peaks of the labeling uncertainty and the sampling units where the label can be trusted. The error rates are strongly related to the heterogeneity indices of the sampling units for both binary ( $p_m$ , see Figure 4) and majority classification system ( $\epsilon$ , see Figure 5).

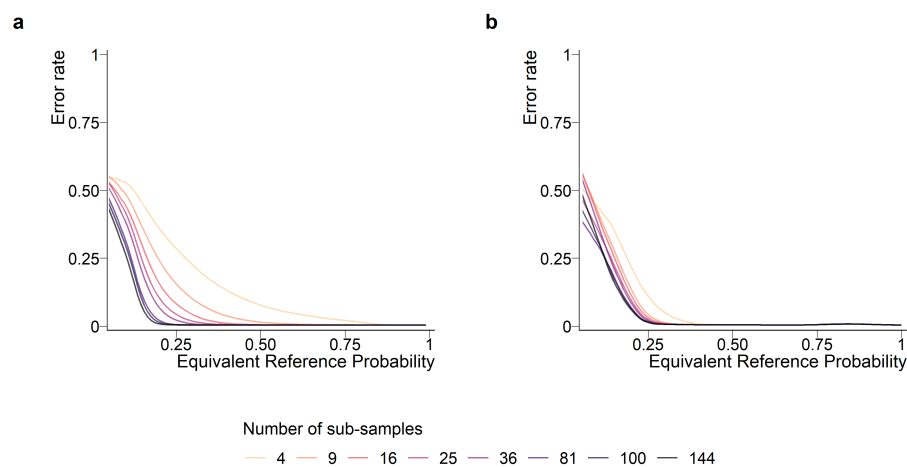
In point-based designs, the error rate is maximum for sampling units with forest proportions close to the class threshold. The error distribution is slightly asymmetric, especially with small sampling efforts, with the longest tail towards the proportion of 50%. Consistently with the expression of the variance for a binomial distribution, the largest variance of the distribution of errors is observed with the 50% threshold and decreased towards the extremities of the range. To achieve an error rate of  $<1\%$  on average with 16 points, the actual proportion needs to differ from the threshold value by circa 20%.

Thus, the error distribution of partition-based response designs also peaks near threshold values. Our results clearly indicate that the partition-based method are strongly biased when the threshold is not 50%: with thresholds at 10% and 75%, the errors rates increase from 50% towards the value of the threshold, where they are systematically wrong. This is due to the systematic omission of the class that contains the 50% interval when approaching the extremities of the range, which is the only type of error when using a partition-based response design with these threshold values. Indeed, the error rate drops to zero in terms of detection of the class that is located at one end of the interval.

For the majority classification system, the largest error rates are observed for sampling units with similar proportions (low  $\epsilon$  values). Labeling is 99% correct when the reference equivalent probability is  $\geq 0.5$ . Interestingly, point-based designs are more accurate than partition-based designs for complex landscapes. On the other hand, designs based on a limited number of partitions outperformed their point-based counterparts for simple landscapes ( $\epsilon > 0.25$ ). Overall, partition-based designs appear insensitive to an increase above 9 in the number of sub-samples, while larger numbers of sub-samples markedly improve the efficiency of point-based design within the range of values tested in this study. This corroborates the results of overall error rates in the case study (Table 1).



**Figure 4.** Error rates in reference labels as a function of landscape fragmentation for binary point-based (a) and partition-based (b) designs. The percentages in the titles of the graphs indicate the threshold that is used to determine the membership to one of the two binary classes.



**Figure 5.** Error rates in reference labels as a function of landscape fragmentation for majority point-based (a) and partition-based (b) designs. Low equivalent reference probabilities are caused by large fragmentation.



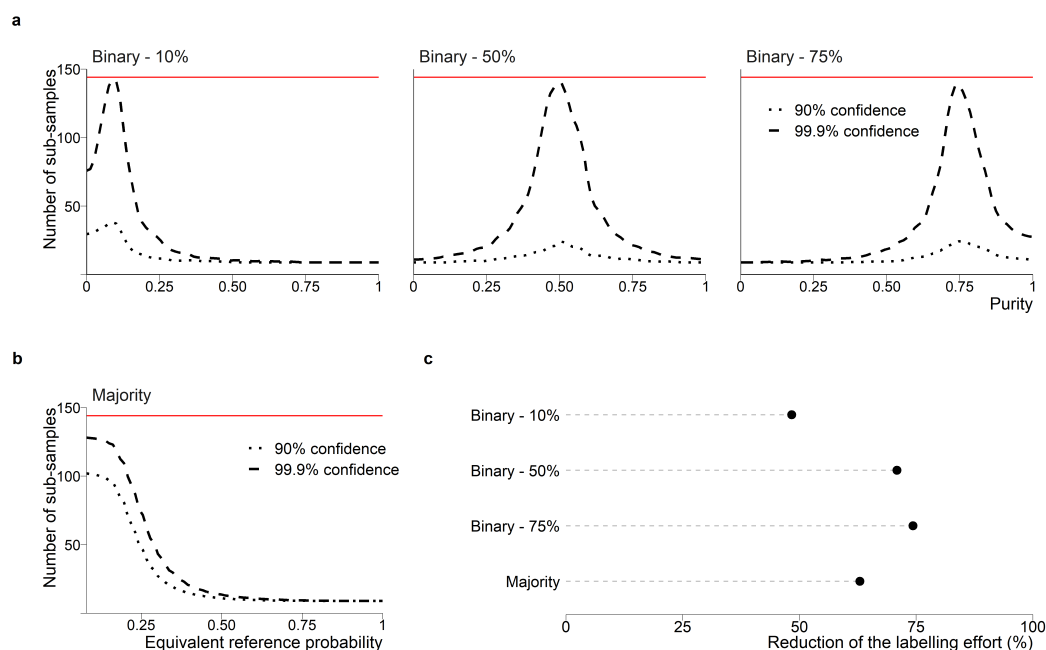
### 5.3. Optimization of the Number of Sub-samples

We optimized the number of sub-samples so that the resulting label reached either (1) a 90% or 99.9% confidence level for each sampling unit or (2) the maximum number of sub-samples (144) was attained. The difference between the error rates of the 99.9% optimized and the error rate with a fixed (144) number of sub-samples was  $<1\%$ .

The regions of high label uncertainty highlighted in Figure 4 and Table 1 are consistent with the regions that require more validation efforts (Figure 6a,b). More samples are needed when  $\epsilon$  is low or  $p_m$  is close to the threshold value. The mean number of sub-samples then decreases quickly, especially with the binary classification systems, so that less than 20 points are needed for most of the range of  $\epsilon$  or  $p_m$  when the confidence level is set to 99.9%. Furthermore, the 90% confidence level is achieved with low effort for the binary classification systems (Figure 6a). The sampling effort around the minimum  $\epsilon$  value for the majority classification system remains high in comparison (Figure 6b), which is consistent with the larger error rates observed for point-based validation.

The main difference between the shapes of the distribution of the error rates (Figure 4) compared with the mean optimized number of points (Figure 6a) occurs on the extreme values of  $p_m$  for the binary classification system. Indeed, the error rate for  $p_m$  value of 0% with 10% threshold (or 100% with threshold 75%) is close to zero, but the number of sub-samples needed to achieve 99.9% confidence is relatively high (75 for the 10% threshold and 40 for the 75% threshold).

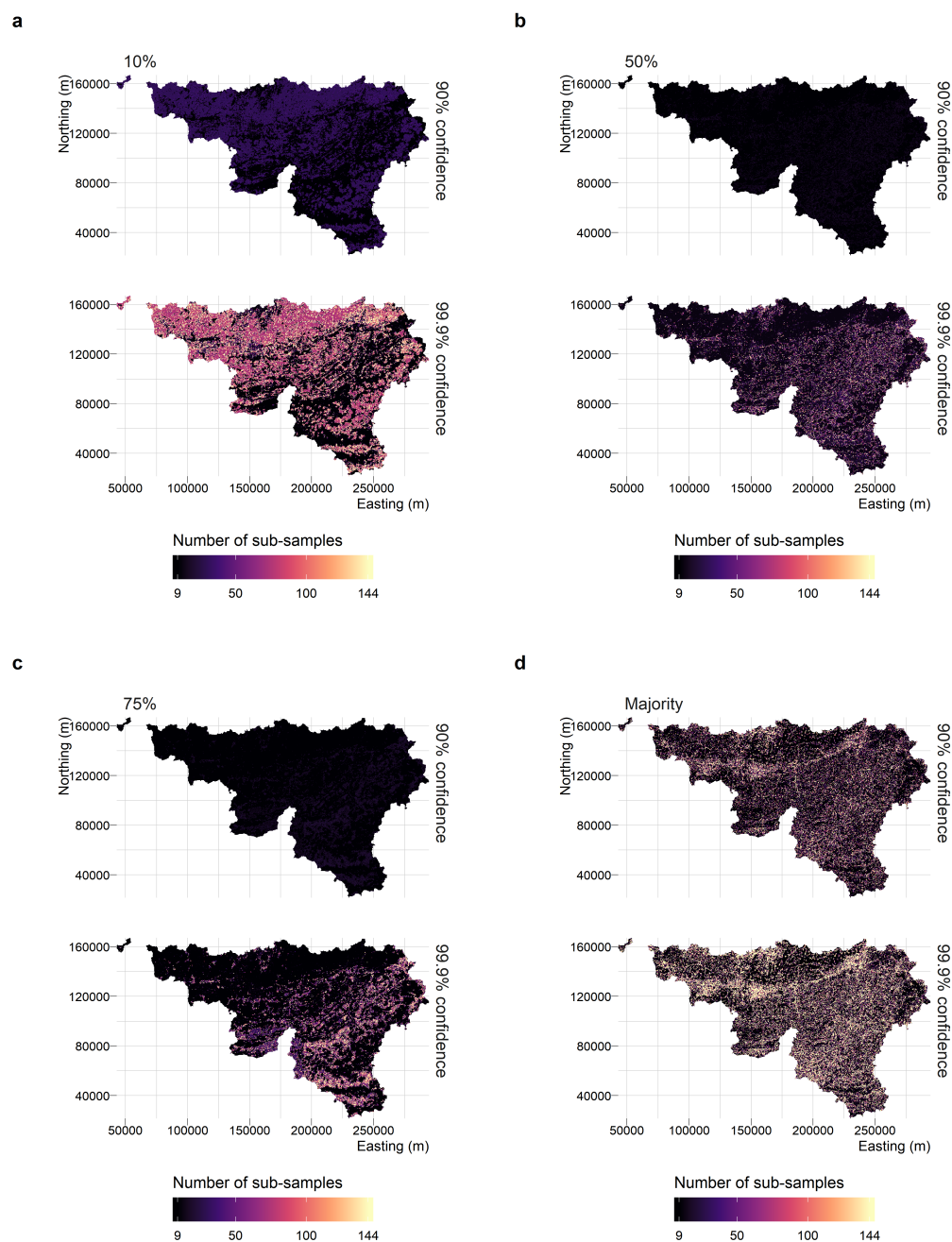
In the study area, the optimization method with a very high confidence (99.9%) could more than halve the labeling effort compared to a systematic sub-sampling of 144 sub-samples per sampling unit (Figure 6c). It required an average of 57, 27 and 26 sub-samples for binary thresholds of 10, 50 and 75%, respectively. The majority classification system was the most difficult to validate, with an average of 115 points needed and the maximum (144) number of sub-samples needed in 56.1% of the sampling units.



**Figure 6.** Effect of optimizing the number of sub-samples on the labeling effort and the error rate using different confidence levels for threshold-based (a) and majority-based classification system (b). The red line indicates the amount of sub-samples collected without optimisation (144 sub-samples). The proposed method markedly reduces the number of sub-samples in all cases (c).

The 90% confidence interval could be achieved for the binary classification systems with on average 21, 10 and 11 sub-samples for thresholds of 10, 50 and 75% respectively. For the majority classification system, it required an average of 99 points. Those results are due to the fact that the maximum number of sub-samples (that is 144 in this study) was reached 38% of the cases.

Collecting reference data with a confidence level of 90% can be achieved at a reasonable cost (Figure 7). The number of sub-samples that is needed to determine the binary class is indeed relatively low (between 9 and 40) over the entire study area, despite the fragmentation of the landscape. For the majority classification system, a spatial pattern of sampling units that need more than 100 sub-samples becomes visible along the main rivers (Meuse and Sambre) and the main cities, where the diversity of land-cover types is the highest.



**Figure 7.** Optimized number of sub-samples for binary classification systems at 10% (a); 50% (b); 75% (c); and for a majority classification system (d).

Reaching a confidence level of 99.9% on the reference data set is much more challenging (Figure 7). For the majority classification system, the required number of sub-samples is even larger in heterogeneous areas (such as the large urban areas) while only a few patches of homogeneous land-cover types can be validated with as little less than 20 sub-samples. Mapping the required number of sub-samples also highlights the particularities of the landscape in the study area. For instance, the binary classification system is more demanding in open landscapes (for the 10% threshold) or closed forests (for the 75% threshold), when the actual proportions are close to the threshold values. On the other hand, the benefits of the optimization of the number of sub-samples is substantial on large patches where the land-cover proportions are distinct from the threshold value of the classification system. On the study area, the threshold of 75% shows the biggest contrasts between the areas that need a large effort and the areas that can be easily validated. On the other hand, the 50% threshold is most of the time the less demanding in terms of validation effort and mainly requires extra efforts along the forest edges.

## 6. Discussion

We assessed the accuracy of two main types of quantitative response designs—a set of points and a grid of squares—based on a protocol that provided full control over the validation process. While it is well known that mixed pixels are more difficult to label than pure ones, we quantified how labeling uncertainty increased for class proportions close to the class boundaries. Our results highlight the underestimated difficulty of developing accurate reference data sets for any combination of response design and classification system. Indeed, the required number of sub-samples to reach 98% confidence level was often too high (more than 100 sub-samples per sampling unit) to be practically implemented. When factoring in the cost of response designs with large number of sub-samples, collecting error-free reference data seems barely feasible. Therefore, matching the data collection effort to the available resources appears critical. In other words, there is a necessary sacrifice of the confidence of reference data to achieve rigorous accuracy assessment at reasonable costs.

The efficiency of point-based and partition-based response designs differed depending on the classification system. Partition-based response designs ought to be preferred for majority classification systems and for binary classification systems with threshold values close to 50%. Point-based response designs become more efficient for thresholds that are close to 0% or 100%. The ability to directly determine the class proportions inside a sampling could also help to arbitrate between the two types of partition-based response designs, MTT and TTM, because TTM is much more dependent of the operator skills than the other response designs. Response designs that are solely based on the estimations of the proportions by an operator would however necessitate a specific quality control to evaluate their accuracy.

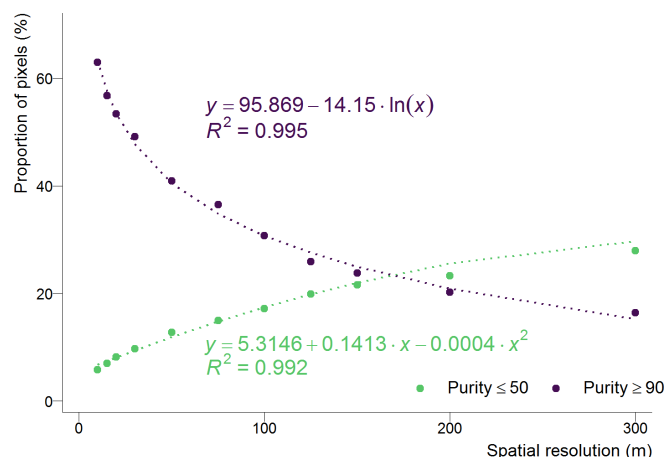
The main advantage of point-based validation is the possibility to estimate the reliability of the label from the points themselves, and hence to objectively optimize the sub-sampling process without prior knowledge about the sampling units. We demonstrated that relying on a fixed number of sub-samples is inefficient because the same amount of resources is spent to label both homogeneous and complex sampling units. An efficient approach would reduce the effort for those easy-to-interpret cases and allocate it to label complex cases to increase their confidence. To this aim, we propose to iteratively interpret sub-samples until the estimated class proportions reached the desired confidence level. Combined with advanced validation applications, such an approach computes the required number of sub-samples on the fly, thereby reducing the labeling cost as soon as there is no doubt (for a given confidence level) about the label of the sampling units. We showed that, in our study area, the labeling effort could be reduced by 50% to 75% without affecting the accuracy of the labels. As a result, the labeling effort was strongly reduced across the study site and concentrated in the fragmented and ambiguous areas. In some cases, i.e., close to the threshold value, the added value of labeling additional points plateaus because sampling units with proportions close to the classification

system definition are always uncertain. Limiting the maximum number of sub-samples to be labelled is thus recommended.

An iterative optimization approach for partition-based designs is impractical because labels could be contradictory when changing scale. Therefore, optimizing partition-based designs would rather depend on subjective operator decision about the proportions she/he estimates inside each sampling unit. Nonetheless, well trained operators could be granted the ability to select several partitions based on their impression of the complexity of the landscape. This method is likely to work well for threshold values near 50% and could avoid extra work in simple cases, but remains sensitive to the Modifiable Area Unit Problem (MAUP) –a statistical biasing effect that occurs when arbitrary units are used to collect data such as class proportions. As described in Jelinski and Wu [57], the MAUP applies to two types of problems which are relevant in partition-based designs. The first aspect of the MAUP is the “scale problem”, where the same set of areal data is aggregated into several sets of larger areal units, with each combination leading to different data values. The second is the “zoning problem”, where a given set of areal units is recombined into zones that are of the same size but located differently, again resulting in variation in data values. MAUP could be mitigated by generating partitions that correspond to actual image objects derived via segmentation [46]. Image segmentation has been used in response designs for coarse resolution image validation [58] and is mainly justified in landscapes that can be divided in a small number of homogeneous patches, not in areas that are very fragmented at a larger scale than the sampling unit. Because of the purity of image-objects derived by image segmentation is usually larger than square cells of the same area [49], this reduces the uncertainty of the labeling [59]. However, delineation errors frequently introduce variance and bias on the estimated surfaces of the patches extracted by automated image segmentation [60]. The accuracy of the response design should therefore be assessed with external data or again rely on an estimation provided by the operator. However, one may lose control over the number of sub-samples generated by the segmentation algorithm, leading to unpractical labeling effort.

In this study, the reported error rates resulted only from imprecise estimation of the class proportions. There are, however, additional errors that should still be considered for a complete understanding of the response design reliability: (1) the simplification of the pixel model, which is a simplified representation of the area observed by remote sensing [61–65]; (2) geolocation errors, which further increase the variance of the estimated proportions because the sub-samples may be matched with locations that are outside the sampling unit; and (3) photo-interpretation errors. Indeed, while we assumed that operators made no errors throughout the paper, their performance is in reality imperfect [23,46,66,67]. For instance, Powell et al. [23] concluded that five interpreters were required to agree upon a specific class. Human factors are responsible for no less than 20% of the inter-individual differences in operator performance [10]. To be more realistic, errors rates should account for errors of interpretation of the landscape and, in partition-based designs, errors in estimating the area of each class. As such, the error rates reported in this study are thus lower bounds.

Sampling units of  $360\text{ m} \times 360\text{ m}$  were used because these are divisible by a large number of integers and, therefore, allowed us to easily simulate a large set of regular partition-based designs. While this practical constraint has no direct impact on the generalization of our results, changing the size of sampling units would, however, indirectly impact the response design accuracy. Indeed, the average purity of the sampling units increases when the ratio between the ground sampling distance and the width of the object increases [62]. This general rule was also observed in our study area, which showed very strong relationship ( $R^2 > 0.99$ ) between the pixel purity and the spatial resolution (Figure 8). If the classification system remains identical, the accuracy of the response design will likely increase for sampling units of higher spatial resolution, and the estimation errors could be neglected when the sampling units become smaller than the spatial objects of interest.



**Figure 8.** Proportions of the purity index values for different spatial resolutions. The purple line indicates pixels with a purity above 90% and the green line shows pixels with a purity of less than 50%

Another solution to minimize the estimation error is to carefully select the classification system with threshold values as far as possible from the modes of the distributions of land-cover proportions. This explained why our results with the binary classification system at 50% threshold was more reliable than the other binary classification systems despite its larger variance on the estimated proportions around the threshold value. In practice, this is not always feasible because it might reduce the fitness for purpose of the map. Besides, the uncertainty of the classification system could become misleading when the study area becomes too large. The preliminary step of defining the classification system is therefore of paramount importance.

In consolidating good practices to collect gold-standard validation data, we demonstrated that the number of sub-samples required to meet stringent confidence levels is often too large to be realistically implemented. Therefore, this work suggests three main directions for future research. First, direct class assessment by operators should be compared with sub-sampling approaches to evaluate the overall level of confidence with real photo-interpretation in both cases. Second, unbiased confusion matrices could be built to account for uncertain reference data. While the errors affecting reference data cannot be predicted, we have however shown that the probability of estimation error could be estimated based on the sub-samples. This information could be used to quantify a large part of the uncertainty of a reference data set at no extra cost. Third, the recent advances in image recognition and computer vision suggest that computer-assisted labeling of sub-samples could help to increase the number of sub-samples at lower cost (see, for instance, [68]). However, algorithms would need to perform very accurately not to compromise the quality standards of reference data.

## 7. Conclusions

Photo-interpreted reference data sets are generally assumed error-free, but they are in practice affected by erroneous labeling due to inaccurate image interpretation, drops of vigilance and estimation errors. We argue that in contrast to interpretation errors and drops of vigilance that could be prevented (using for instance repeated labeling), estimation errors are intrinsic to the response design and cannot be avoided once the response design is defined. With the goal of improving good practices in reference data collection, we empirically assessed the relationship between the accuracy of reference data and the type of response design for binary and majority classification systems. Our results highlight the need for dense sub-sampling to collect error-free reference data. We further demonstrated that estimation errors are strongly linked to landscape composition, labeling errors being more prevalent when the class proportions are close to the class definition threshold (binary classification systems) or in areas with complex class compositions (majority classification systems). To leverage the relationship between landscape composition and labeling accuracy, we propose to iteratively interpret sub-samples until the class proportions are estimated with the desired confidence level. By quantifying the confidence



of photo-interpreted labels, this optimization method provides an efficient trade-off between the accuracy of the reference data and the labeling cost. Therefore, its uptake by the remote sensing community will likely improve the reliability of accuracy estimates and the assessment of the usability of thematic maps.

**Author Contributions:** Conceptualization, J.R., F.W. and P.B.; methodology, J.R., F.W. and P.B.; software, J.R. and F.W.; formal analysis, J.R., F.W. and P.B.; writing—original draft preparation, J.R. and F.W.; writing—review and editing, J.R., F.W. and P.B.; visualization, J.R. and F.W. All authors have read and agree to the published version of the manuscript.

**Funding:** This research was funded by the Fédération Wallonie–Bruxelles in the frame of the Lifewatch-WB project and by the CSIRO in the frame of the Digiscape Future Science Platform.

**Acknowledgments:** The authors thank the four reviewers for their detailed comments. The 2-m land-cover map used in this study is available for download on the Lifewatch-WB website with a CC-O license (<https://maps.elie.ucl.ac.be/lifewatch/ecotopes.html>).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Congalton, R.G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [\[CrossRef\]](#)
2. Stehman, S.V.; Czaplewski, R.L. Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sens. Environ.* **1998**, *64*, 331–344. [\[CrossRef\]](#)
3. Stehman, S.V. Basic probability sampling designs for thematic map accuracy assessment. *Int. J. Remote Sens.* **1999**, *20*, 2423–2441. [\[CrossRef\]](#)
4. Stehman, S.V. Statistical rigor and practical utility in thematic map accuracy assessment. *Photogramm. Eng. Remote Sens.* **2001**, *67*, 727–734.
5. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [\[CrossRef\]](#)
6. Stehman, S.V.; Foody, G.M. Key issues in rigorous accuracy assessment of land cover products. *Remote Sens. Environ.* **2019**, *231*, 111199. [\[CrossRef\]](#)
7. Woodcock, C.E.; Gopal, S. Fuzzy set theory and thematic maps: Accuracy assessment and area estimation. *Int. J. Geogr. Inf. Sci.* **2000**, *14*, 153–172. [\[CrossRef\]](#)
8. Foody, G.M. The impact of imperfect ground reference data on the accuracy of land cover change estimation. *Int. J. Remote Sens.* **2009**, *30*, 3275–3281. [\[CrossRef\]](#)
9. Foody, G.M. Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. *Remote Sens. Lett.* **2013**, *4*, 783–792. [\[CrossRef\]](#)
10. Van Coillie, F.M.; Gardin, S.; Anseel, F.; Duyck, W.; Verbeke, L.P.; De Wulf, R.R. Variability of operator performance in remote-sensing image interpretation: The importance of human and external factors. *Int. J. Remote Sens.* **2014**, *35*, 754–778. [\[CrossRef\]](#)
11. Parasuraman, R. *Vigilance, monitoring, and search*. In *Handbook of Perception and Human Performance*, Vol. 2, *Cognitive Processes and Performance*; Boff, K.R., Kaufman, L., Thomas, J.P., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 1986; pp. 1–39.
12. Szalma, J.L.; Hancock, P.A.; Dember, W.N.; Warm, J.S. Training for vigilance: The effect of knowledge of results format and dispositional optimism and pessimism on performance and stress. *Br. J. Psychol.* **2006**, *97*, 115–135. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Pengra, B.W.; Stehman, S.V.; Horton, J.A.; Dockter, D.J.; Schroeder, T.A.; Yang, Z.; Cohen, W.B.; Healey, S.P.; Loveland, T.R. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. *Remote Sens. Environ.* **2019**. [\[CrossRef\]](#)
14. Svatoňová, H. Reading Satellite Images, Aerial Photos and Maps: Development of Cartographic and Visual Literacy. In *Current Topics in Czech and Central European Geography Education*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 187–208.
15. Olson, C.E. Elements of photographic interpretation common to several sensors. *Photogramm. Eng.* **1960**, *26*, 651–656.



16. Colwell, R.N. The extraction of data from aerial photographs by human and mechanical means. *Photogrammetria* **1965**, *20*, 211–228. [[CrossRef](#)]
17. Klein, G.A.; Hoffman, R.R. Perceptual-cognitive aspects of expertise. In *Cognitive Science Foundations of Instruction*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, USA, 1993; pp. 203–226.
18. Strand, G.H.; Dramstad, W.; Engan, G. The effect of field experience on the accuracy of identifying land cover types in aerial photographs. *Int. J. Appl. Earth Obs. Geoinf.* **2002**, *4*, 137–146. [[CrossRef](#)]
19. Lillesand, T.; Kiefer, R.; Chipman, J. *Digital Image Interpretation and Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2008; Volume 6, pp. 545–581.
20. Wang, M.J.J.; Lin, S.C.; Drury, C.G. Training for strategy in visual search. *Int. J. Ind. Ergon.* **1997**, *20*, 101–108. [[CrossRef](#)]
21. Maruff, P.; Danckert, J.; Camplin, G.; Currie, J. Behavioral goals constrain the selection of visual information. *Psychol. Sci.* **1999**, *10*, 522–525. [[CrossRef](#)]
22. Lloyd, R.; Hodgson, M.E.; Stokes, A. Visual categorization with aerial photographs. *Ann. Assoc. Am. Geogr.* **2002**, *92*, 241–266. [[CrossRef](#)]
23. Powell, R.; Matzke, N.; De Souza, C.; Clark, M.; Numata, I.; Hess, L.; Roberts, D. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sens. Environ.* **2004**, *90*, 221–234. [[CrossRef](#)]
24. Tran, T.V.; Julian, J.P.; De Beurs, K.M. Land Cover Heterogeneity Effects on Sub-Pixel and Per-Pixel Classifications. *ISPRS Int. J. Geo-Inf.* **2014**, *3*, 540–553. [[CrossRef](#)]
25. Foody, G.M. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sens. Environ.* **2010**, *114*, 2271–2285. [[CrossRef](#)]
26. Radoux, J.; Bogaert, P.; Fasbender, D.; Defourny, P. Thematic accuracy assessment of geographic object-based image classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 895–911. [[CrossRef](#)]
27. Radoux, J.; Bogaert, P. Accounting for the area of polygon sampling units for the prediction of primary accuracy assessment indices. *Remote Sens. Environ.* **2014**, *142*, 9–19. [[CrossRef](#)]
28. Anderson, J.R. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*; US Government Printing Office: Washington, DC, USA, 1976; Volume 964.
29. Congalton, R.; Gu, J.; Yadav, K.; Thenkabail, P.; Ozdogan, M. Global land cover mapping: A review and uncertainty analysis. *Remote Sens.* **2014**, *6*, 12070–12093. [[CrossRef](#)]
30. Grekousis, G.; Mountrakis, G.; Kavouras, M. An overview of 21 global and 43 regional land-cover mapping products. *Int. J. Remote Sens.* **2015**, *36*, 5309–5335. [[CrossRef](#)]
31. Di Gregorio, A. *Land Cover Classification System: Classification Concepts and User Manual: LCCS*; FAO: Rome, Italy, 2005; Volume 8.
32. Radoux, J.; Bogaert, P. Good practices for object-based accuracy assessment. *Remote Sens.* **2017**, *9*, 646. [[CrossRef](#)]
33. Arvor, D.; Belgiu, M.; Falomir, Z.; Mougenot, I.; Durieux, L. Ontologies to interpret remote sensing images: Why do we need them? *GISci. Remote Sens.* **2019**, *56*, 1–29.
34. Fernandes, R.; Fraser, R.; Latifovic, R.; Cihlar, J.; Beaubien, J.; Du, Y. Approaches to fractional land cover and continuous field mapping: A comparative assessment over the BOREAS study region. *Remote Sens. Environ.* **2004**, *89*, 234–251. [[CrossRef](#)]
35. Petit, C.; Lambin, E. Integration of multi-source remote sensing data for land cover change detection. *Int. J. Geogr. Inf. Sci.* **2001**, *15*, 785–803. [[CrossRef](#)]
36. Bartholome, E.; Belward, A.S. GLC2000: A new approach to global land cover mapping from Earth observation data. *Int. J. Remote Sens.* **2005**, *26*, 1959–1977. [[CrossRef](#)]
37. Bontemps, S.; Defourny, P.; Radoux, J.; Van Bogaert, E.; Lamarche, C.; Achard, F.; Mayaux, P.; Boettcher, M.; Brockmann, C.; Kirches, G.; et al. Consistent global land cover maps for climate modelling communities: Current achievements of the ESA's land cover CCI. In *Proceedings of the ESA Living Planet Symposium, Edinburgh, UK, 9–13 September 2013*; pp. 9–13.
38. Waldner, F.; Fritz, S.; Di Gregorio, A.; Plotnikov, D.; Bartalev, S.; Kussul, N.; Gong, P.; Thenkabail, P.; Hazeu, G.; Klein, I.; et al. A unified cropland layer at 250 m for global agriculture monitoring. *Data* **2016**, *1*, 3. [[CrossRef](#)]

39. Lamarche, C.; Santoro, M.; Bontemps, S.; d'Andrimont, R.; Radoux, J.; Giustarini, L.; Brockmann, C.; Wevers, J.; Defourny, P.; Arino, O. Compilation and validation of SAR and optical data products for a complete and global map of inland/ocean water tailored to the climate modeling community. *Remote Sens.* **2017**, *9*, 36. [[CrossRef](#)]
40. Shimada, M.; Itoh, T.; Motooka, T.; Watanabe, M.; Shiraishi, T.; Thapa, R.; Lucas, R. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sens. Environ.* **2014**, *155*, 13–31. [[CrossRef](#)]
41. Bastin, J.F.; Berrahmouni, N.; Grainger, A.; Maniatis, D.; Mollicone, D.; Moore, R.; Patriarca, C.; Picard, N.; Sparrow, B.; Abraham, E.M.; et al. The extent of forest in dryland biomes. *Science* **2017**, *356*, 635–638. [[CrossRef](#)]
42. Stehman, S.V.; Wickham, J.D. Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sens. Environ.* **2011**, *115*, 3044–3055. [[CrossRef](#)]
43. Perger, C.; Fritz, S.; McCallum, I.; Salk, C. Cropland Capture: A gaming approach to improve global land cover. In Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellon, Spain, 3–6 June 2014; pp. 3–16.
44. Bey, A.; Sánchez-Paus Díaz, A.; Maniatis, D.; Marchi, G.; Mollicone, D.; Ricci, S.; Bastin, J.F.; Moore, R.; Federici, S.; Rezende, M.; et al. Collect earth: Land use and land cover assessment through augmented visual interpretation. *Remote Sens.* **2016**, *8*, 807. [[CrossRef](#)]
45. Bayas, J.C.L.; Lesiv, M.; Waldner, F.; Schucknecht, A.; Duerauer, M.; See, L.; Fritz, S.; Fraisl, D.; Moorthy, I.; McCallum, I.; et al. A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform. *Sci. Data* **2017**, *4*, 170136. [[CrossRef](#)]
46. Waldner, F.; Schucknecht, A.; Lesiv, M.; Gallego, J.; See, L.; Pérez-Hoyos, A.; d'Andrimont, R.; de Maet, T.; Bayas, J.C.L.; Fritz, S.; et al. Conflation of expert and crowd reference data to validate global binary thematic maps. *Remote Sens. Environ.* **2019**, *221*, 235–246. [[CrossRef](#)]
47. Stehman, S.V. Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* **1992**, *58*, 1343–1350.
48. Matérn, B. *Spatial Variation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 36.
49. Radoux, J.; Bourdouxhe, A.; Coos, W.; Dufrêne, M.; Defourny, P. Improving Ecotope Segmentation by Combining Topographic and Spectral Data. *Remote Sens.* **2019**, *11*, 354. [[CrossRef](#)]
50. Food and Agriculture Organization of the United Nations. *Global Forest Resources Assessment 2010—Main Report*; FAO: Rome, Italy, 2010.
51. Bogaert, P.; Waldner, F.; Defourny, P. Information-based criterion to measure pixel-level thematic uncertainty in land cover classifications. *Stoch. Environ. Res. Risk Assess.* **2016**, *31*, 2297–2312. [[CrossRef](#)]
52. Cleveland, W.; Grosse, E.; Shyu, W. Local Regression Models. In *Chapter 8 in Statistical Models*; Chambers, J.M., Hastie, T.J., Eds.; Wadsworth & Brooks/Cole: Pacific Grove, CA, USA, 1992.
53. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413. [[CrossRef](#)]
54. Gold, R.Z. Tests auxiliary to  $\chi^2$  tests in a Markov chain. *Ann. Math. Stat.* **1963**, *34*, 56–74. [[CrossRef](#)]
55. Quesenberry, C.P.; Hurst, D. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* **1964**, *6*, 191–195. [[CrossRef](#)]
56. Goodman, L.A. On simultaneous confidence intervals for multinomial proportions. *Technometrics* **1965**, *7*, 247–254. [[CrossRef](#)]
57. Jelinski, D.E.; Wu, J. The modifiable areal unit problem and implications for landscape ecology. *Landsc. Ecol.* **1996**, *11*, 129–140. [[CrossRef](#)]
58. Olofsson, P.; Stehman, S.V.; Woodcock, C.E.; Sulla-Menashe, D.; Sibley, A.M.; Newell, J.D.; Friedl, M.A.; Herold, M. A global land-cover validation data set, part I: Fundamental design principles. *Int. J. Remote Sens.* **2012**, *33*, 5768–5788. [[CrossRef](#)]
59. MacLean, M.G.; Campbell, M.J.; Maynard, D.S.; Ducey, M.J.; Congalton, R.G. Requirements for labelling forest polygons in an object-based image analysis classification. *Int. J. Remote Sens.* **2013**, *34*, 2531–2547. [[CrossRef](#)]
60. Radoux, J.; Defourny, P. A quantitative assessment of boundaries in automated forest stand delineation using very high resolution imagery. *Remote Sens. Environ.* **2007**, *110*, 468–475. [[CrossRef](#)]

61. Foody, G.M.; Arora, M.K. Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications. *Pattern Recognit. Lett.* **1996**, *17*, 1389–1398. [[CrossRef](#)]
62. Hsieh, P.F.; Lee, L.C.; Chen, N.Y. Effect of spatial resolution on classification errors of pure and mixed pixels in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2001**, *39*, 2657–2663. [[CrossRef](#)]
63. Radoux, J.; Chomé, G.; Jacques, D.; Waldner, F.; Bellemans, N.; Matton, N.; Lamarche, C.; d’Andrimont, R.; Defourny, P. Sentinel-2’s potential for sub-pixel landscape feature detection. *Remote Sens.* **2016**, *8*, 488. [[CrossRef](#)]
64. Fisher, P. The pixel: A snare and a delusion. *Int. J. Remote Sens.* **1997**, *18*, 679–685. [[CrossRef](#)]
65. Waldner, F.; Duveiller, G.; Defourny, P. Local adjustments of image spatial resolution to optimize large-area mapping in the era of big data. *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *73*, 374–385. [[CrossRef](#)]
66. Vancutsem, C.; Marinho, E.; Kayitakire, F.; See, L.; Fritz, S. Harmonizing and combining existing land cover/land use datasets for cropland area monitoring at the African continental scale. *Remote Sens.* **2012**, *5*, 19–41. [[CrossRef](#)]
67. See, L.; Comber, A.; Salk, C.; Fritz, S.; Van Der Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; Obersteiner, M. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PLoS ONE* **2013**, *8*, e69958. [[CrossRef](#)]
68. Xing, H.; Meng, Y.; Wang, Z.; Fan, K.; Hou, D. Exploring geo-tagged photos for land cover validation with deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 237–251. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).