Remote Sensing of Environment 221 (2019) 235–246

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Conflation of expert and crowd reference data to validate global binary thematic maps

François Waldner[a,b,*], Anne Schucknecht[c,d], Myroslava Lesiv[e], Javier Gallego[c], Linda See[e], Ana Pérez-Hoyos[c], Raphaël d'Andrimont[a,c], Thomas de Maet[a], Juan Carlos Laso Bayas[e], Steffen Fritz[e], Olivier Leo[c], Hervé Kerdiles[c], Mónica Díez[f], Kristof Van Tricht[g], Sven Gilliams[g], Andrii Shelestov[h], Mykola Lavreniuk[h], Margareth Simões[i,q], Rodrigo Ferraz[i], Beatriz Bellón[j], Agnès Bégué[j,r], Gerard Hazeu[k], Vaclav Stonacek[l], Jan Kolomaznik[l], Jan Misurec[l], Santiago R. Verón[m,n], Diego de Abelleyra[m], Dmitry Plotnikov[o], Li Mingyong[p], Mrinal Singha[p], Prashant Patil[p], Miao Zhang[p], Pierre Defourny[a]

[a] Université Catholique de Louvain, Earth and Life Institute, Louvain-la-Neuve, Belgium
[b] Commonwealth Scientific and Industrial Research Organisation, Agriculture and Food, St Lucia, Australia
[c] European Commission Joint Research Centre, Ispra, Italy
[d] Karlsruhe Institute of Technology, Garmisch-Partenkirchen, Germany
[e] International Institute for Applied Systems Analysis, Laxenburg, Austria
[f] DEIMOS IMAGING, Boecillo, Valladolid, Spain
[g] VITO Remote Sensing, Mol, Belgium
[h] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institue", Kyiv, Ukraine
[i] Embrapa Solos, Rio de Janeiro, Brazil
[j] CIRAD, UMR Tetis, Montpellier, France
[k] Wageningen Environmental Research (Alterra), Wageningen, the Netherlands
[l] Gisat s.r.o., Prague, Czech Republic
[m] Instituto Nacional de Tecnología Agropecuaria (INTA), Hurlingham, Argentina
[n] Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina
[o] Terrestrial Ecosystems Monitoring Laboratory, Space Research Institute of Russian Academy of Sciences (IKI), Moscow, Russia
[p] Key Laboratory of Digital Earth Science, Institude of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China
[q] Universidade do Estado do Rio de Janeiro – UERJ/FEN/DESC/PPGMA, Brazil
[r] Tetis, CIRAD, IRSTEA, AgroParisTech, CNRS, Univ Montpellier, Montpellier, France

ARTICLE INFO

Keywords:
Accuracy assessment
Crowdsourcing
Volunteered geographic information
Data quality
Stratified systematic sampling
Photo-interpretation

ABSTRACT

With the unprecedented availability of satellite data and the rise of global binary maps, the collection of shared reference data sets should be fostered to allow systematic product benchmarking and validation. Authoritative global reference data are generally collected by experts with regional knowledge through photo-interpretation. During the last decade, crowdsourcing has emerged as an attractive alternative for rapid and relatively cheap data collection, beckoning the increasingly relevant question: can these two data sources be combined to validate thematic maps? In this article, we compared expert and crowd data and assessed their relative agreement for cropland identification, a land cover class often reported as difficult to map. Results indicate that observations from experts and volunteers could be partially conflated provided that several consistency checks are performed. We propose that conflation, *i.e.*, replacement and augmentation of expert observations by crowdsourced observations, should be carried out both at the sampling and data analytics levels. The latter allows to evaluate the reliability of crowdsourced observations and to decide whether they should be conflated or discarded. We demonstrate that the standard deviation of crowdsourced contributions is a simple yet robust indicator of reliability which can effectively inform conflation. Following this criterion, we found that 70% of the expert observations could be crowdsourced with little to no effect on accuracy estimates, allowing a strategic reallocation of the spared expert effort to increase the reliability of the remaining 30% at no additional cost. Finally, we provide a collection of evidence-based recommendations for future hybrid reference data collection campaigns.

* Corresponding author at: Commonwealth Scientific and Industrial Research Organisation, Agriculture and Food, St Lucia, Australia.
E-mail address: franz.waldner@csiro.au (F. Waldner).

https://doi.org/10.1016/j.rse.2018.10.039
Received 25 April 2018; Received in revised form 30 October 2018; Accepted 31 October 2018

1. Introduction

With the increasing availability of remotely-sensed imagery acquired by satellites such as PROBA-V, Landsat-8, and Sentinel-1, -2 and -3, the number of available land cover maps will undoubtedly increase. In fact, such a trend is already visible for global binary thematic products, *e.g.*, Yu et al. (2013); Pesaresi et al. (2013); European Environment Agency (2015); Pekel et al. (2016); Lamarche et al. (2017). This development should foster the collection and sharing of reference databases by the user community to systematically assess and benchmark existing and future products, especially given the continued uncertainty in land cover products (Fritz et al., 2011). While some reference data sets can be sourced from the Global Observation of Forest and Land Cover Dynamics platform (GOFC-GOLD, 2015), they have several shortcomings for binary validation, mainly due to their small sample sizes. Moreover, it has been observed that map accuracy estimates can vary significantly depending upon the reference data set used, owing to differences in the sampling scheme and sampling density as well as to differences in the legend definition (Waldner et al., 2015). For instance, the accuracy estimates of cropland extent in global land cover maps range from 56% to 76% (Fritz et al., 2011).

The development of appropriate global reference data sets is a challenging task because of the lack of availability of *in situ* data over large areas and the cost associated with such collection efforts (Bastin et al., 2013). For these reasons, authoritative global reference data are, in the best case, collected by remote sensing experts with a strong knowledge and understanding of specific ecosystems *via* photo-interpretation of high spatial resolution imagery (*e.g.*, Mayaux et al., 2006; Defourny et al., 2012; Bontemps et al., 2013).

In recent years, the rise of new technologies and the free and open availability of very high resolution imagery such as Google Earth and Bing Maps have allowed vast amounts of land cover information to be collected (See et al., 2013). Citizens without professional expertise in remote sensing or geospatial sciences can become actively engaged in the creation and analysis of large data sets through what is known as crowdsourcing (Howe, 2015) or volunteered geographic information (VGI; Goodchild, 2007). The rise of user-created geospatial content has been of great benefit to the collection of large quantities of reference data (Iwao et al., 2006; Schepaschenko et al., 2015) and to improved product development (Clark and Aide, 2011). Geo-Wiki (Fritz et al., 2009) is a prime example of a land cover tool that has involved citizens in the collection of validation data in the past (Comber et al., 2013; See et al., 2015) while Collect Earth is a more recent alternative (Bey et al., 2016).

Several approaches to rigorously include crowdsourcing in design-based statistical inference for area estimation and accuracy assessment of land cover have been presented in Stehman et al. (2018). They include: 1) directing volunteers to obtain data at locations selected from a probability-based sampling scheme, 2) treating crowdsourced data as a certainty stratum and augmenting the crowdsourced data with reference data obtained from probability-based sampling, and 3) using crowdsourcing to create an auxiliary variable that is then used in a model-assisted estimator to reduce the standard error of an estimate produced from a probability-based sample.

Map accuracy assessment and area estimation are particularly sensitive to errors in the reference labels as many low-paid interpreters or volunteers are prone to giving noisy answers, thereby violating the basic assumption of error-free validation data. Errors that occur during data collection propagate through to the validation process (Woodcock and Gopal, 2000; Foody, 2011, 2013). Therefore the fundamental question is: how can we synergistically combine expert and crowd data to leverage the potential of crowdsourcing while maintaining the quality standards of accuracy assessment? Depending on the level of agreement between expert and crowd data, we foresee three possible outcomes:

1. **Exclusion**: differences between the crowdsourced and expert data are too large and the former must be discarded.
2. **Partial conflation**: the agreement between the two approaches is discontinuous, *i.e.*, they strongly correlate in certain cases and weakly in others. Conflation can be applied where correlation is expected to be high. Therefore, conflation requires an understanding of where the errors are likely to occur.
3. **Replacement**: the crowd and the experts strongly agree so that observations from the two groups are interchangeable, *i.e.*, experts and volunteers have a level of agreement similar to that of experts among themselves.

The objectives of this paper are two-fold. First, we seek to quantify the agreement between expert and crowd observations and determine the conflation outcome (exclusion, conflation, or replacement) accordingly. Secondly, we evaluate different variables that could serve as a proxy for crowd reliability that could therefore inform a conflation strategy. We illustrate the performance of informed and random conflation by validating two global cropland maps, as cropland is a land cover class that suffers from high uncertainty in global cropland maps (Fritz et al., 2011; Waldner et al., 2015). To meet these goals, we also introduce a probability-based sampling design based on systematic sampling that employs denser sampling in regions with higher uncertainty. Each sampling unit was interpreted by the two groups of interpreters (*i.e.*, experts and the crowd) using specific data collection tools. It is important to clarify here that our purpose is not to assess the respective interpretation capability of each groups. Rather we evaluate if they can be combined by acknowledging their intrinsic differences. We conclude this paper by proposing a set of guidelines for future hybrid (crowdsourced and expert-based) reference data collection.

2. Material and methods

2.1. Sampling scheme

2.1.1. Sampling design

The Committee on Earth Observing Satellites-Land Product Validation (CEOS-LPV) report for global land cover map validation (Strahler et al., 2006) has defined the recommended common standards for accuracy assessment but it is not explicit with regards to the implementation. A common sampling approach is stratified random sampling (*e.g.*, Arino et al., 2008; Bicheron et al., 2008; Bontemps et al., 2011; Clark and Aide, 2011). In an effort to improve the cost-efficiency of large-area land cover validation, Olofsson et al. (2012) and Stehman et al. (2012) introduced the fundamental design and estimation principles underlying stratified random sampling to enable a coordinated, comparable and regularly updated global land-cover validation database.

In this paper we opted for a stratified systematic sampling scheme with a common pattern of replicates. Stratified systematic sampling also provides unbiased estimators of accuracy and has other advantages over random sampling (Wolter, 1984; Gallego et al., 2016). For instance, it is more efficient than random sampling in terms of variance when the spatial correlation of the variable of interest decreases with distance (Bellhouse, 2014), which is generally valid for land cover data derived by remote sensing (Dunn and Harrison, 1993). However, stratified systematic sampling is achieved at the cost of a more complex implementation. It is often criticised for its lack of flexibility when the sample size needs to be modified (Stehman, 2009) and for the loss of spatial homogeneity of the sample distribution –which is the basis of its good performance–when strata are small and scattered. There is also no unbiased estimator of the variance as the usual estimator may strongly overestimate it. Some alternatives have been proposed to reduce overestimation, *e.g.*, on the basis of local variance (Gallego and Delincé, 2010; Wolter, 1984).

We sought to overcome the rigidity of stratified systematic sampling

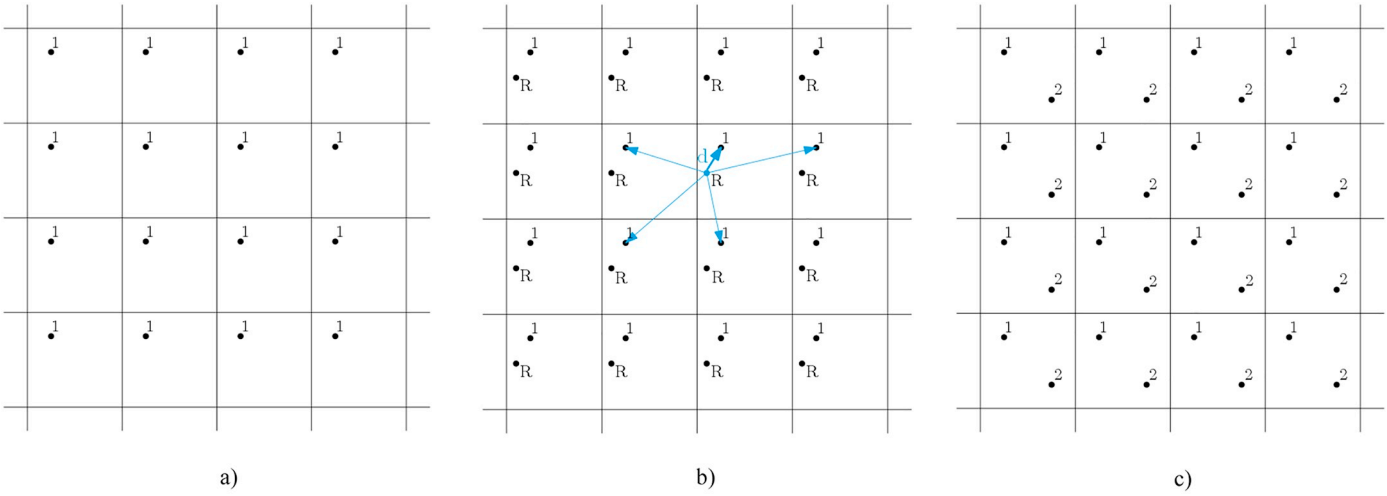

Fig. 1. Generation of a pattern of common replicates: a) a replicate is a set of points that occupy the same position in the grid cells, b) the distance between two replicates is the minimum distance between all pairs of the replicates, c) together replicates 1 and 2 make a new regular pattern.

by relying on a pattern of common replicates. The generation of the sample is a three-step process that 1) creates a systematic grid of replicates, 2) adjusts the number of sampling units by latitude for geographic coordinates, that do not preserve area, and 3) stratifies the area of interest to intensify samples in specific strata, *e.g.*, in error-prone strata. We describe the mechanics of the sampling in the next paragraphs.

2.1.1.1. Building a pattern of replicates. Let us start by defining a grid covering an area of interest. In practice, this grid can have any regular shape but for simplicity let us consider a regular square grid. We then select the location of the first *replicate* randomly (Fig. 1a). A *replicate* is a set of points that occupy the same position within each grid cell. The location of a second replicate is then defined as to maximise the minimum distance to all points of the first replicate regardless of the grid cell to which they belong (maximin criterion). For instance, the location of generic replicate R in Fig. 1b does not satisfy the maximin criterion because distance d (the smallest distance to a point of the first replicate; the bold arrow in the figure) can be increased without reducing the distances from all other points to the first replicate below d. As the sampling grid is square, there is only one location that satisfies the maximin distance criterion (Fig. 1c). Together replicates 1 and 2 constitute a new systematic pattern following diagonal lines. If the sampling needs to be intensified, additional replicates can be selected. Each additional replicate would be selected so as to maximise the minimum distance to all previously selected replicates. This prevents sampling units that are too close to be selected.

It may happen that the chosen sample size is not a multiple of a number of replicates. For instance, we may wish to select 500 points from a pattern of replicates that generates 400 or 800 points using one or two replicates, respectively. This may be solved by selecting a random subset of the last replicate, *e.g.*, 25% of the second replicate (100 points) in the previous example. A practical way to implement this is to attribute a random number ε with a uniform distribution between 0 and 1 to all points of the last replicate. We can then rank the points by increasing $R + \varepsilon$ values, where R is the replicate number assigned to each point. We finally select the points according to this ranking until the desired sample size is reached.

2.1.1.2. Correction for non-equal-area projections. Systematic sampling is usually based on an equal-area projection. However most maps are produced in geographic coordinates which requires further adaptations because the grid cell area diminishes when moving away from the equator. While this effect is minor in tropical areas, it becomes significant in temperate regions. Therefore we propose to downgrade a fraction of each replicate to account for the variability of the grid cell area as applied in Bontemps et al. (2011). This can be achieved by slightly modifying the ranking method introduced previously. Knowing that a parallel at latitude α has approximately a length of $\cos(\alpha)$ compared to the equator, the points are now ranked according to $(R + \varepsilon)/\cos(\alpha)$. Therefore, a proportion $1 - \cos(\alpha)$ of points belonging to replicate 1 is downgraded to replicate 2, a proportion $2 \times (1 - \cos(\alpha))$ is downgraded from replicate 2 to 3, and so on. This correction ensures a uniform sampling with marginal distortions owing to the Earth's oblateness.

2.1.1.3. Stratification. Points are assigned the stratum they fall in as attribute. Using the ranking method, they are then selected on a per stratum basis according to the sampling rates, *i.e.*, the number of sampling units to select per stratum. Consider a stratum for which p sampling units need to be selected. Replicates are generated until the number of points falling in this stratum surpasses the number of desired sampling points. The selected points for this stratum consist of 1) all the replicates belonging to the $R - 1$ replicates, and 2) a random sample of the last replicate so that the total number of selected points matches the desired sample size. If sampling units have to be added at at later stage, they can be selected from the last replicate or from new replicates if need be.

2.1.1.4. Implementation of the proposed sampling. We defined a regular square grid with cells of 1° by 1° and generated two replicates. We applied the non-equal-area correction because the sampling grid was defined so as to coincide with PROBA-V images in geographic coordinates (Dierckx et al., 2014). The locations of the replicates were defined so as to correspond to pixel centroids. We propose to increase the sampling density in areas that were *a priori* problematic for cropland discrimination. Therefore a cropland probability map (Fritz et al., 2015) was utilised to build strata corresponding to areas with different misclassification probabilities. The probability map was constructed by computing the agreement of a series of global cropland maps. A high cropland probability means that the majority of these maps agreed on the presence of cropland whereas a null occurrence probability means that all products converged in the absence of cropland. A low cropland probability is to be interpreted as a high disagreement among maps. We generated four strata for which we defined specific sampling rates assuming that intermediate occurrence probabilities were most error prone (Table 1; Experts *vs.* Crowd). We refer to Fig. S1 for an overview of the spatial distribution of the sample.

As stratum 1 (0% cropland probability) covered most of the world's

Table 1
Definition of the strata and their associated sampling rates (at global scale) for the main Experts *vs*. Crowd and Experts *vs*. Experts samples. A sample of 4147 sampling units was collected to assess the consistency between the experts and the crowd. A subset of 398 sampling units were reinterpreted by a second group of experts to assess the within-expert variability.

Stratum	Occurrence probability	Sampling rate	Number of sampling units in sample	
			Experts *vs.* crowd	Experts *vs.* experts
1	0	Very low	100	18
2	>0–25%	Moderate	1160	94
3	25–75%	High	1962	192
4	>75%	Moderate	925	94
Total			4147	398

area and was assumed to be easier to map, a very small sample size was selected (100 sampling units). We selected points from replicate 1 for strata 2, 3 and 4, and extended the selection to replicate 2 for stratum 3. This means that points of replicate 1 were always selected unless they belonged to stratum 1. Points of replicate 2 were selected only if they fell in stratum 3. Following this procedure, a total of 4147 sampling units were obtained. Fig. 2 illustrates the resulting sampling for a subset of the conterminous United States and contrasts it against non-stratified systematic sampling in geographic coordinates. To assess the agreement between experts, about 10% of the main sample (n = 398) was selected to be evaluated by another expert following a stratified sampling approach (Table 1; Experts *vs*. Experts).

2.1.2. Response designs

Two response designs were defined to better accommodate the intrinsic characteristics of the expert and the volunteer groups during data collection. For the experts, each sampling unit (300 m × 300 m, corresponding to the size of a PROBA-V pixel) was divided into a block of polygons following state-of-the-art approaches devised in the GlobCover and ESA Land Cover Climate Change Initiative validation exercises, whereas blocks of 25 sub-pixels were preferred for the crowd (Fig. 3). We detail both response designs hereunder and describe the rationale for defining different response designs.

The experts that participated in the reference data collection have strong skills in satellite image and time series analysis as well as in land use land cover mapping. Most of them have also some fieldwork experience. They were allocated a set of sampling units to label according to their regional familiarity and agreed to carry out the task until completion. The expert response design consists of blocks of polygons that were generated using automated image segmentation (Fig. 3). Generic segmentation parameters were defined by trial and error so as to provide visually consistent polygons in the majority of the landscapes. Polygons were smoothed and simplified to reduce object complexity and facilitate interpretation. The experts' task was to label each polygon as cropland, non-cropland, or unknown, and to provide an overall confidence level (using “certain”, “reasonable” or “doubtful”). The validation interface displayed the corresponding block of polygons on very high resolution (VHR) images from two base maps (Google Maps, Virtual Earth). Pixel-level time series of the Normalised Difference Vegetation Index (NDVI) were provided to help distinguish crop from non-crop phenology. Available temporal profiles included weekly SPOT Vegetation (1 km spatial resolution; mean of 1999–2012) and PROBA-V time series (300 m spatial resolution; smoothed time series for 2014 and 2015). Sampling units with “unknown” polygons covering >25% of the sampling unit area were discarded as the class proportion estimates would be unreliable. Otherwise, they were neglected in the computation of the cropland proportion.

Volunteers were recruited by reaching out to the Geo-Wiki network and beyond. They differed from the experts in two main ways: their level of competency and the duration of their participation were unknown. This prompted us to define a second response design to better address these specific differences. We also proposed additional incentives to maintain a high level of participation throughout the campaign, *e.g.*, we used gamification and offered prizes or co-authorship to the best volunteers based on the quality and quantity of their contributions (Laso Bayas et al., 2017).

The crowd response design was simplified to limit the labelling workload because the polygon-based response design would have undermined the success of gamification. Volunteers were tasked with labelling blocks of 5 × 5 evenly-sized sub-pixels, which was considered as a good trade-off between the precision of the cropland proportion estimates and the duration of the interpretation and labelling process (Fig. 3). They were told to mark sub-pixels as cropland if they contained >50% cropland using a specific branch within the Geo-Wiki tool. Different background images were available to aid the visual interpretation such as Google Maps, Bing imagery, and cloud-free Sentinel-2

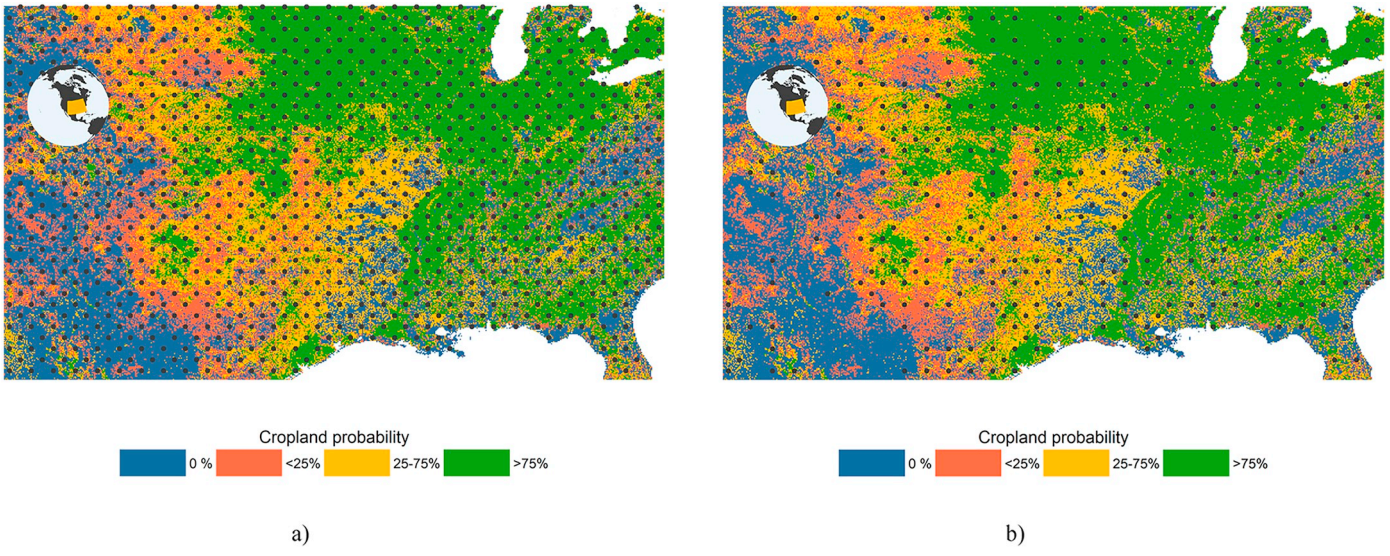

Fig. 2. Differences between (a) non-stratified systematic sampling and (b) the proposed stratified systematic sampling with a downgrading of the replicates by latitude and according to their stratum for most of the conterminous United States. The grid is based on 1° cells. The strata are derived from a cropland probability layer that defines the likelihood of cropland occurrence based on existing land cover maps.

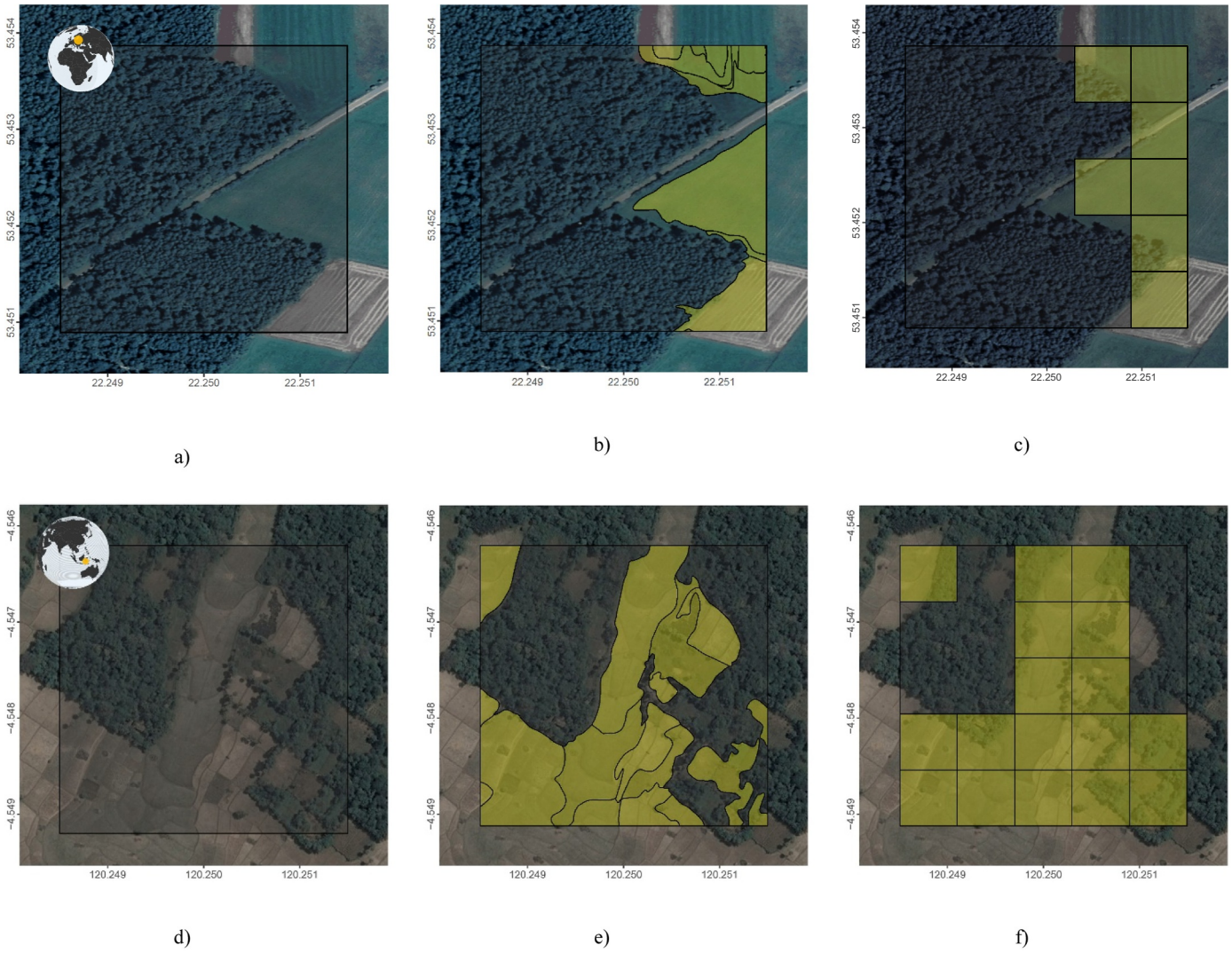

Fig. 3. Examples of sampling units located in a) Poland and d) Indonesia with the corresponding expert-based interpretations (b) and e)) and crowdsourced interpretations (c) and f)). The grey box represents the footprint of a 300-m pixel. Polygons and sub-pixels interpreted as cropland by experts and the crowd, respectively, are displayed in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

images in certain regions. Additionally, different NDVI time series could be displayed when clicking on a pixel. These were based on Landsat-7 and -8 (32-day composites) and MOD13Q1.005 (16-day composites). Volunteers were also encouraged to use a feature embedded within the Geo-Wiki tool to generate a KML file that is then opened automatically in the Google Earth desktop application to view available historical imagery. They were not asked to indicate confidence levels and could skip a sampling unit, indicating if there was no image available/just low-resolution imagery available/clouds present or if the sampling unit was too difficult to label. Finally, we encouraged volunteers to share, *via* social media, print screens of sampling units that were difficult to interpret so as to seek advice from their peers or from the Geo-Wiki team. We summarised the multiple contributions per sampling unit by taking the median of the reported cropland proportions.

Both groups of photo-interpreters were asked to follow the cropland definition from the Joint Experiment for Crop Assessment and Monitoring (JECAM) network: "*The annual cropland from a remote sensing perspective is a piece of land of a minimum 0.25 ha (minimum width of 30 m) that is sowed/planted and harvestable at least once within the 12 months after the sowing/planting date. The annual cropland produces an herbaceous cover and is sometimes combined with some tree or woody vegetation*" (JECAM, 2015; Waldner et al., 2016). In this definition, perennial crops and fallows are excluded from the cropland class. The year of interest was 2015.

2.2. Statistical analysis

Even though the prime focus of the paper is on binary maps, we started by analysing the continuous agreement between the expert and crowd observations of cropland proportions. First, the agreement between expert and crowd observations was measured in terms of Mean Absolute Error (MAE) as per Pontius et al. (2008). Secondly, Lin's Concordance Correlation Coefficient (CCC; Lin, 1989) can be calculated as an index of reliability. Ranging from −1 to 1, it evaluates the degree to which pairs of observations fall on the 1:1 line. Lin's CCC contains a measure of precision using Pearson's correlation coefficient and a bias correction for accuracy.

Next, we converted cropland proportions into two classes by setting the limit between cropland and non-cropland at 50%. We then constructed global and stratum-specific error matrices based on Table 1 and then computed accuracy measures such as the Overall Accuracy (OA), the Producers' Accuracy (PA), and the Users' Accuracy (UA) for the cropland class as this was the class of interest. For the global matrices, we properly weighted the sample data to account for the area of each stratum following Olofsson et al. (2014):

$$w_h = \frac{A_h}{n_h}$$

Table 2
Agreement metrics when comparing cropland proportions. N: Number of sampling units; MAE: Mean Absolute Error; Lin's CCC: Lin's Concordance Correlation Coefficient. Lin's CCC are not provided for stratum 1 due to the limited range of the cropland proportions observed in this stratum.

Stratum	Experts *vs.* crowd			Experts *vs.* experts		
	N	MAE	Lin's CCC	N	MAE	Lin's CCC
1	100	2.9	–	18	0.14	–
2	1160	8.33	0.60	94	3.38	0.82
3	1955	12.36	0.75	192	9.51	0.81
4	878	16.31	0.79	94	11.74	0.78
All	4093	11.83	0.79	398	8.17	0.83

where w_h is the weight of all sampling units belonging to stratum h, n_h is the number of sampling units in that stratum and A_h its area. When comparing with the crowd, experts were considered as reference. Note that for consistency between the Experts *vs.* Crowd and Experts *vs.* Experts analyses, the same group of experts was considered as reference.

To uncover patterns of agreement, we calculated the mean absolute error and the Overall Error (OE = 1 − OA) for bins of cropland proportions (estimated by the experts and by the crowd) as well as for bins of the crowd standard deviation. The crowd standard deviation indicates within-volunteer reliability and relates to the variability of all crowd observations collected for a given sampling unit. The crowd standard deviation was determined per sampling unit based on all available cropland proportion estimates provided by the volunteers.

2.3. Evaluation of the conflation strategy

We tested three scenarios for integration of crowd data in the validation data set (exclusion, partial conflation, and replacement) and their respective impact on accuracy estimates of two 30-m global cropland maps: Globeland30 (Chen et al., 2015) and Global Food Security-Support Analysis Data (GFSAD) (Gumma et al., 2017; Massey et al., 2017; Oliphant et al., 2017; Phalke et al., 2017; Teluguntla et al., 2017; Xiong et al., 2017; Zhong et al., 2017). First, a global GFSAD cropland map was generated from regional/continental maps by taking the maximum cropland extent where these maps overlapped. Then, the two global maps were resampled to 300 m to match the reference grid and per-pixel cropland proportions were computed. We then generated binary cropland/non-cropland maps following a majority rule.

In addition to exclusion and replacement, two partial conflation strategies were compared with conflation rates ranging from 5% to 95% by steps of 5%. The conflation rates indicate the percentage of expert observations that has been replaced by crowdsourced observations. In the first strategy, referred to as random conflation, we randomly swapped expert data by crowd data. Random conflation was repeated 50 times to account for chance. In the second strategy, referred to as informed conflation, we used a proxy to replace expert observations by crowd observations when they seem reliable. The identification of this proxy was based on the results of the statistical analysis. Three accuracy metrics were calculated to assess the impact of conflation on the accuracy estimates: the OA, the PA and the UA of the cropland class.

Finally, we measured the gain of conflation at global and continental scales by computing the number and proportion of conflated sampling units, the area-weighted proportion of conflated sampling units as well as the number of cropland sampling units that were conflated. The weights of the area-weighted proportions is derived by dividing the area of a stratum (either globally or by continent) by the number of sampling units that fall into it as detailed in Olofsson et al. (2014).

Table 3
Agreement metrics when comparing binary outcomes. N: Number of sampling units; OA: Overall Accuracy; UA: Users' Accuracy of the cropland class; PA: Producers' Accuracy of the cropland class.

Stratum	Experts *vs.* crowd				Experts *vs.* experts			
	N	OA	UA	PA	N	OA	UA	PA
1	100	1.00	1.00	–	18	1.00	1.00	–
2	1160	0.93	0.84	0.41	94	0.97	1.00	0.62
3	1955	0.90	0.82	0.67	192	0.89	0.94	0.74
4	878	0.86	0.92	0.80	94	0.88	0.92	0.85
All	4093	0.98	0.99	0.76	398	0.98	0.99	0.81

3. Results

3.1. Measures of agreement

Every sampling unit was validated by one expert and by at least four volunteers (maximum = 16, median = 5, mean = 5.3). Of the 4147 sampling units, 4093 sampling units were left for further analysis after removing the expert observations with an unknown proportion >25%. All sampling units from the Experts *vs.* Experts data set were kept.

The mean absolute error between the experts and the crowd reached 11.8% globally and increased with the cropland probability (2.9% to 16% for stratum 1 and 4, respectively; Table 2). Overall the concordance between the two groups reached 0.79 and was the highest in stratum 4. This coefficient is not provided for stratum 1 because of the limited range of cropland proportion observed in this stratum –most observations agreed on the absence of cropland. Interestingly, the highest disagreement between the experts and the crowd occurred in stratum 4, where cropland was most likely to occur and in larger proportions.

The mean absolute errors in the Experts *vs.* Experts data set were systematically lower than in the Experts *vs.* Crowd data set. For instance, it reached 8.2% globally and remained below 12% in all strata (Table 2). The concordance correlation coefficients were larger than 0.78 in all strata and had the same order of magnitude to those observed for the crowd except in stratum 2.

In the binary case, the overall agreement between the crowd and the experts was high (Table 3). Globally, the OA and the UA were 0.98 and 0.99, respectively while the PA did not exceed 0.76. Most errors occurred in strata 2 and 3 (0.41 and 0.67 respectively). Nonetheless, the analysis of the stratum-level accuracy metrics revealed stronger differences, *e.g.*, in stratum 4. The agreement measured by the OA spanned from 86% to 100% while the UA was systematically larger than the PA. Despite similar overall agreement, experts were systematically characterised by higher PAs.

3.2. Patterns of agreement

Fig. 4 illustrates the evolution of the mean absolute error and the overall error between the crowd and the experts as a function of the cropland proportion seen by the experts (a and d), by the crowd (b and e), and as a function of the crowd standard deviation (c and f). The mean absolute error increases with the cropland proportion as estimated by the experts from 7% in the absence of cropland to 27% where cropland is highly dominant. This translates in near error-free crowd contributions where the cropland proportion is <30%. Most errors were observed in the range of 50–60%, which is close to the legend cut-off proportion of 50%. Crowdsourced cropland proportions exhibited the most errors (>30%) around the cropland class cut-off value both in the continuous and discrete cases. Finally, the mean absolute error increased with the standard deviation of the crowd, *i.e.*, it is possible to infer the expected agreement based on the crowd standard deviation.

Error-free observations were more likely to occur when the crowd

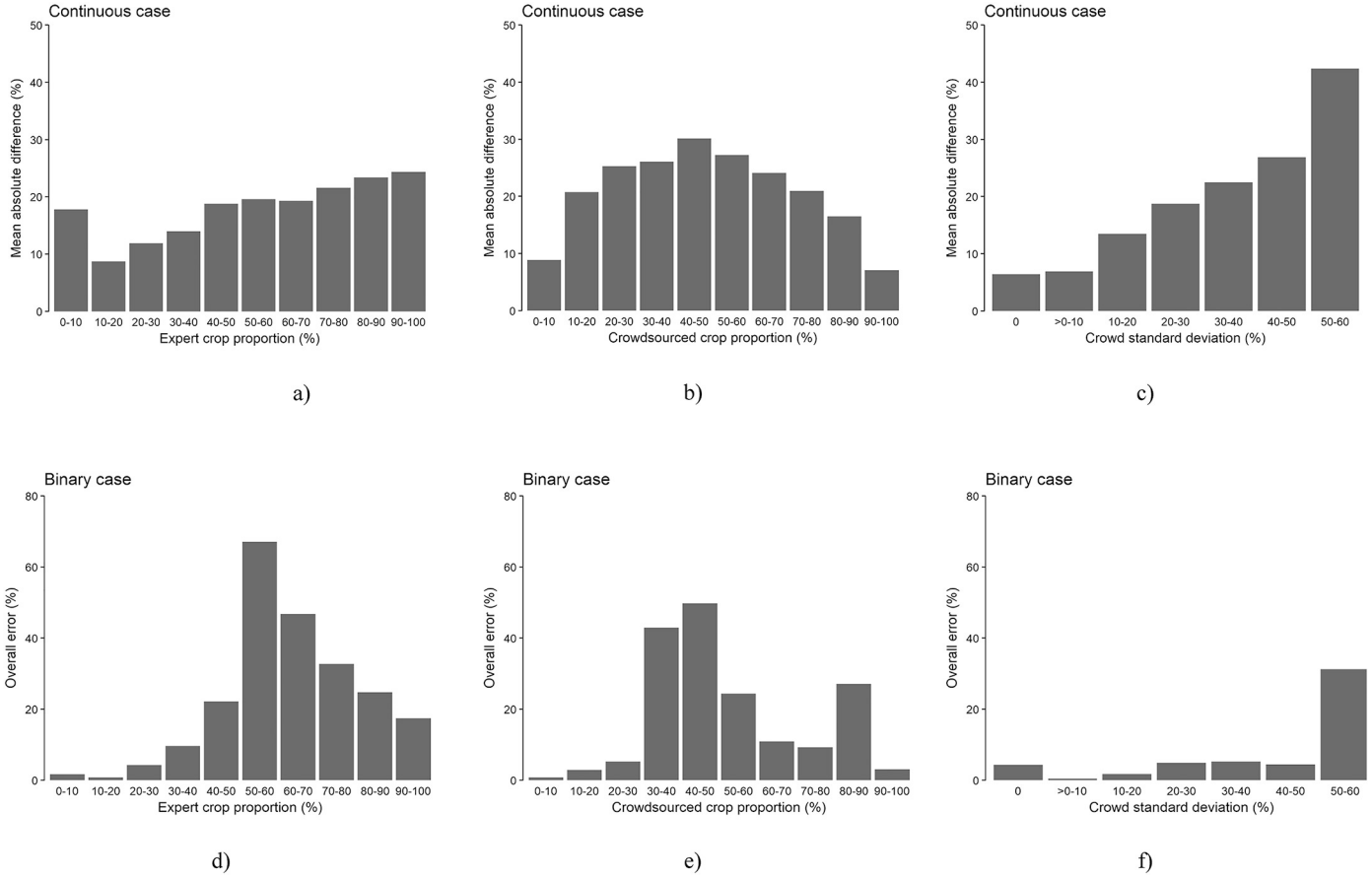

Fig. 4. Metrics of disagreement between the experts and the crowd. The first row presents the mean absolute error as a function of a) the expert crop proportion, b) the crowd crop proportion, and c) the standard deviation of the crowd. The second row presents the overall error as a function of d) the expert crop proportion, e) the crowd crop proportion, and f) the standard deviation of the crowd. Most discrete errors occur around the cut-off proportion between cropland and non-cropland. The crowd standard deviation is a reliable predictor of the disagreement with the experts.

standard deviation was <20%. About 10% of error was expected for crowd standard errors between 20% and 50%. The overall error increased sharply when the crowd standard deviation exceeded 50%. Therefore, the crowd standard deviation can be used as a simple proxy for the reliability of the crowd contributions.

3.3. Assessment of the conflation strategies

We validated the resampled GFSAD and Globeland30 (hereafter referred to as Globeland) maps using the expert data set (Table S1). The OAs were larger than 0.93 and the UAs of the cropland class were lower than 0.60. Accuracy strongly varies per stratum which highlights the relevance of the error-prone stratification approach. Based on the results presented in Section 3.2, we selected the crowd standard deviation to inform conflation. We conflated the expert and the crowd data sets for a range of rates, gradually swapping expert data by crowd data with increasing standard deviation. Note that 0% and 100% conflation correspond to the exclusion and the replacement scenarios, respectively. This informed conflation approach was benchmarked against random conflation. The hybrid data sets were then used to validate the two cropland maps.

Fig. 5 highlights that informed conflation is a valid approach to combine expert and crowd observations: it successfully kept the OA steady, up to a conflation rate of 70% in both cases. A similar trend was observed for the UA. The conflation strategy remained around the expert-only estimate until it abruptly decreased at around 70%, whereas the random conflation led to a steady decrease. In both maps however, the standard deviation criterion was less effective for the PA as the two strategies followed the same trajectory until 30% of conflation. From then on, informed conflation outperformed random conflation.

Fig. 6 shows the evolution of 1) the maximum crowd standard deviation and 2) the proportion of cropland sampling units for different conflation rate increases. Interestingly, only consensus crowd observations (standard deviation = 0%) were conflated until 50% conflation was reached. Introducing observations with a standard deviation of >30% led to sharp differences in accuracy compared to the exclusion case (conflation rate of 0%). About 30% of the sampling units had been conflated when the crowd standard deviation became different than zero.

Assuming a conflation rate of 70% (n = 2861), we computed the number and proportion of expert effort that could be spared by continent and by stratum (Table 4). The two continents where conflation was the most widespread were Asia (n = 919) and the Americas (n = 821). Accounting for the weight of the sampling units in the sampling scheme, 80% of the area of the four continents could be conflated. Almost all observations in stratum 1 could be conflated. Stratum 4 had the highest proportion of cropland sampling units in the conflation set (155 out of 445), underscoring the relevance of the stratification. Note that the vast majority of the observations were non-cropland observations (see also Fig. 6).

4. Discussion

4.1. On the sampling scheme

The sampling approach adopted in this study is an example of active

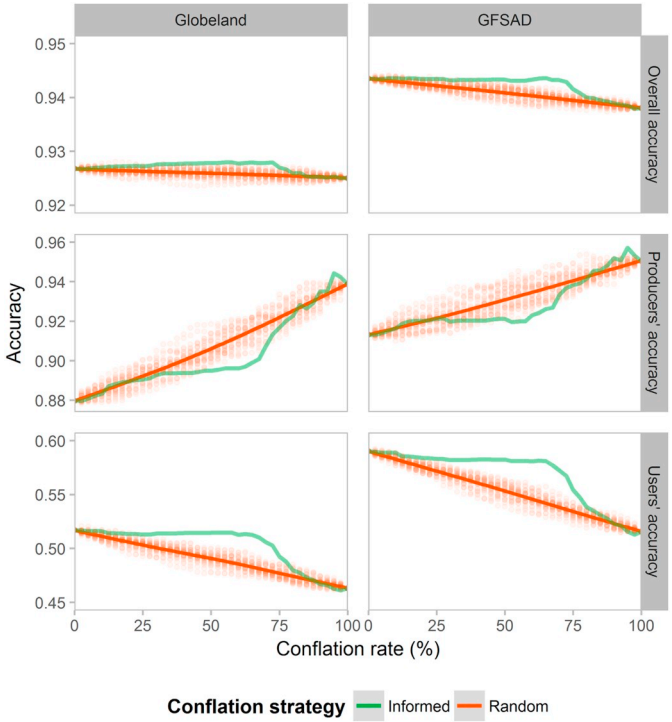

Fig. 5. Evolution of three accuracy metrics (overall accuracy, producers' and users' accuracy) as a function of the conflation rate. Zero percent and hundred percent of conflation correspond to the exclusion and the replacement scenarios, respectively whereas all the intermediate cases are partial conflation cases. In one case (green line), the crowd standard deviation was minimised and in the other (red), observations were randomly swapped. The red dots illustrate particular random realisations and the red line highlights the trend. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

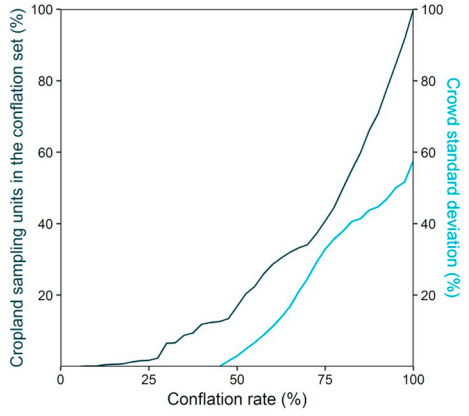

Fig. 6. Evolution of the crowd standard deviation and the proportion of crop observations that were conflated as a function of the conflation rate. When minimising the crowd standard deviation of the conflated sample, the conflation rate had to exceed 47% for observations that had not reached consensus to be selected.

crowdsourcing where volunteers are directed to collect data at specific locations specified by a probability sampling design that allows rigorous inference (see Stehman et al., 2018). To that aim, we introduced a systematic stratified sampling approach with a variable number of replicates. We based the stratification on a probability map of the class of interest to sample error-prone areas at a higher rate, *i.e.*, where the class occurrence probability was between 25% and 75%. Results show different error rates per stratum which supports the relevance of the sampling design. This finding converges with Lamarche et al. (2017) who reported that denser sampling rates in error-prone areas leads to more precise map accuracy estimates, which in turns improves map comparison.

Table 4
Number and proportion of the sampling units to be conflated assuming a conflation ratio of 70%. These quantities are provided per continent and per stratum.

	Number of conflated sampling units	Proportion of conflated sampling units (%)	Area-weighted proportion of conflated sampling units (%)	Number of cropland sampling units in the conflation
Continent				
Africa	648	70	88	11
Asia	919	70	86	119
Europe	390	65	80	56
Americas	821	69	84	77
Oceania	83	61	65	14
Stratum				
1	91	91	91	0
2	970	84	84	15
3	1355	69	69	107
4	445	51	51	155
Global	2861	70	71	277

As it is likely that practitioners need to augment the sample size of a specific region to increase the precision of the accuracy and/or area estimates, we reduced the rigidity of the systematic sampling by introducing a system based on a pattern of replicates. The proposed sampling design allows the sample size to be augmented by successively adding new replicates until the desired number of sampling units is reached (see Section 2.1.1). For instance, Laso Bayas et al. (2017) densified the size of the sample tenfold. However, increased flexibility is achieved at the expense of a slight reduction in the sampling efficiency compared to pure systematic sampling (Gallego et al., 2016). Note that findings on conflation are not bounded to a specific sampling design and could be adapted to other designs, *e.g.*, as proposed in Stehman et al. (2012).

Another important benefit of systematic sampling is the intrinsic traceability of the sample, which means it could facilitate its uptake as a shared reference data set. When random samples are used for the validation of remote sensing products without providing metadata, code, software versions, and sampling grid, readers or reviewers have no way of knowing whether difficult points have been eliminated without being documented. The benefits of traceability are not fully realised as the current sample used a correction for non-equal-area coordinate systems that require a random adjustment to the original pattern. In future efforts, traceability could be achieved with equal-area projection systems. For instance, the validation scheme of GlobCover relied on an equal-area projection per continent (Bontemps et al., 2011).

The proposed sampling scheme remains valid for validating other land cover classes but this would require adjustming of the stratification map accordingly. If the collection of the validation data needs to take place before the map is produced, it could be obtained following Fritz et al. (2015) similar to how the cropland probability strata were derived. Otherwise, probability information or classification uncertainty could be derived from the class memberships provided by the classifier (*e.g.*, Bogaert et al., 2016), or by using pixel-based accuracy estimates (*e.g.*, Khatami et al., 2017). The sampling design is also readily applicable for the validation of land cover changes such as cropland expansion (Morton et al., 2006), cropland abandonment (Löw et al., 2018) or urban sprawl (Taubenböck et al., 2009). Addressing land cover change would also require adjustments to the validation interface to include VHR images of the epochs of interest. In that regard, Picture Pile is an interesting example of the use of crowdsourcing for land cover change analysis (Danylo et al., 2018).

The response designs could also be adjusted to validate maps of higher spatial resolution, *e.g.*, 30 m. An important consideration, however, is the quality of the very high resolution images to be photo-interpreted. With smaller polygons and sub-pixels to label, it is critical that the photo-interpreter is provided with sufficiently detailed imagery.

4.2. On the accuracy of photo-interpretation and the need for quality checks

Both the experts and the crowd were subject to errors. Some sources of error affected both groups equally and others were specific to each group.

4.2.1. General sources of error

General factors affecting the performance of any photo-interpreter are primarily demographic, non-cognitive, and cognitive; external and technical factors impact performance to a lesser extent (Van Coillie et al., 2014). Additional factors affecting both groups were related to the experimental set up. They include: 1) out-of-season VHR images for which accurate discrimination between grassland and cropland is challenging, 2) noisy NDVI time series, 3) VHR images of poor quality, *e.g.*, low resolution or cloud cover, and 4) outdated VHR images. We refer to Lesiv et al. (2018) for a comprehensive analysis of the availability and topicality of VHR images in Google Earth and Bing Maps. Further inconsistencies could have occurred because of the diversity of data sources provided; there is, in fact, no guarantee that the different contributors based their interpretations on the same (combination of) data sources. Nonetheless, results demonstrate the high suitability of high-resolution satellite imagery for the validation of binary thematic maps, and the interest of using different sources to reach an efficient global result.

The different response designs (blocks of sub-pixels *vs.* blocks of polygons) and the limited precision of the cropland proportion estimates due to the number of sub-pixels (100/25 = 4%) might have introduced some inconsistencies in the comparison of the expert and the crowd observations. Indeed, the polygon-based response design is thought to be more precise than the pixel-based one because it is not affected by the Modifiable Area Unit Problem (MAUP), *i.e.*, polygons delineate actual image objects whereas pixels do not. A simulation experiment to evaluate the impact of using two different response designs revealed high agreement between cropland proportions derived with blocks of polygons or with a blocks of sub-pixels (adjusted R^2 of 0.987; see Fig. S2) and that the corresponding omission and commission errors were relatively balanced. This seems to indicate that the effect on the continuous assessment was limited and that class allocation differences likely cancelled each other out. A way to further reduce these discrepancies is to increase the number of sub-pixels, which would be especially relevant when validating continuous products. Attention should be paid to keep the number of sub-pixels to label sufficiently low in order not to reduce the participation of the crowd, which would in turn impact the reliability of standard deviation estimates. Nonetheless, blocks of sub-pixels represent an improvement to a direct presence/absence of cropland at the pixel level directly, such as implemented in previous crowdsourcing campaigns, *e.g.*, *See* et al. (2014). With 25 sub-pixels, the extra work for the crowd (compared to a direct interpretation of the sampling unit) did not significantly impact the participation.

4.2.2. Sources of error specific to each group

Photo-interpretation by a group of regional experts is often considered as the gold standard for large-scale validation. Indeed, local knowledge leads to higher accuracy, irrespective of the photo-interpreters' surveying experience (de Leeuw et al., 2011; Strand et al., 2002). We report a within-expert agreement (Lin's CCC = 0.83 and MAE = 8%) higher than the agreement between experts and the crowd (Lin's CCC = 0.79 and MAE = 12%), which tends to confirm that different levels of expertise and regional knowledge of the landscapes influence photo-interpretation. However, within-expert agreement was a far cry from perfect. These results are congruent with those obtained by Vancutsem et al. (2012), who reported different rates of between-expert agreement as a function of the cropland proportion in the sampling unit (83% and 45% agreement, respectively). Similarly, Powell et al. (2004) concluded that five interpreters were required to agree upon a specific class. This suggests that expert contributions should also be quality controlled to guarantee their accuracy and reliability. This could be achieved at no additional cost as conflation would help reduce the experts' workload. Particular attention should be paid 1) to verify that they have a thorough understanding of the phenomenon being observed and knowledge of the geographic region that they interpret (de Leeuw et al., 2011), 2) to ensure good working conditions (Van Coillie et al., 2014), and 3) to provide verification mechanisms during the labelling phase, as well as tools to manage fatigue and to avoid declines in vigilance (Van Coillie et al., 2014). Some errors might also be attributed to less than perfect segmentations. Analysis of the comments left by the experts during the labelling process reveals that segmentation issues were reported for $<3\%$ of the sampling units. $<0.5\%$ corresponds to cropland proportions between 40% and 60%, a range where labelling errors are more likely to impact class attribution. The generic segmentation algorithm might have resulted in over-segmentation in areas with very large field size and under-segmentation in complex landscapes. On the one hand, under-segmentation increases the workload (as image objects are represented by more than one polygon) but should marginally affect the labelling accuracy. On the other hand, over-segmentation is more likely to reduce the labelling accuracy. The magnitude of its impacts is a function of the image object sizes and is only relevant for cropland non-cropland cases (other land cover classes do not affect the estimates of cropland proportion). Examples of these situations include “roads between fields in field segments” or “trees [included] in field parcels”. The link between a reportedly poor segmentation and a subsequent poor labelling is thus not evident. In fact, experts remained confident in their labelling in 43% of the reported cases and were “doubtful” in only 12%. As a final note, results suggest that expert-based photo-interpretation suffers from an error of 8% (Table 2). This propagation of error during map validation could partially explain why the accuracy of global land cover maps seems to level off at around 70%.

Crowdsourcing was expected to reach a lower accuracy than that of the experts because it is open to all. For example, in a previous study by Fritz et al. (2013), the accuracy of crowdsourced observations ranged between 66% and 76%, and the agreement between volunteers reached 83%. In this study, however, the main contributors were engaged in remote sensing or geospatial sciences, or were students in related fields (Laso Bayas et al., 2017). Therefore some degree of expertise can be assumed. In addition, the performance and accuracy of volunteers can improve over time (See et al., 2013) as they are trained and feedback on their contributions is provided. This was encouraged by providing training material and encouraging the use of social media for difficult-to-interpret sampling units. Furthermore, the aggregation of multiple contributions per sampling unit (using the median) might have helped filter out extreme and unlikely contributions. Similarly, it might also have averaged out correct answers where most volunteers were mistaken. Advanced aggregation techniques should be explored to take the reliability of volunteered interpretations into account, *e.g.*, with latent class modelling (Foody et al., 2013, 2018) or by exploiting the knowledge of experts within the crowd (Mann and Helbing, 2017; Prelec et al., 2017).

It is important to reiterate here that ground truth data collection is also subject to error. For instance, it is estimated that the Land Use and Cover Area frame Survey (LUCAS) –a survey totalling a cost of €12.5 million for the collection of 337,855 points of which 200 K to 250 K will be *in situ* and the rest photo-interpreted– still suffers from 3% disagreement (Gallego and Delincé, 2010). Nonetheless, if *in situ*

observations were available at the location of the sampling, they could have been used to better assess the absolute quality of the photo-interpretation. In general, access to *in situ* data remains difficult in remote areas, so that ground truth observations are rare, or inaccessible. International programs such as the global strategy for improving agricultural and rural statistics (FAO, 2017) will hopefully provide more georeferenced *in situ* data in the future.

4.3. Achieving conflation

Results show that partial conflation of expert and crowd reference data can be achieved. However conflation is only practically viable when informed by a measure of the reliability of the crowd contributions such as the crowd standard deviation. Achieving informed conflation requires a reliable estimation of the selected reliability measure. For the crowd standard deviation, this can be achieved by defining a sufficiently high number of sub-pixels per sampling units and by ensuring that each sampling unit is labelled by a large number of volunteers. With our data set, a conflation rate of 70% could be realised without significant changes in the accuracy estimates. In practice, the conflation rate should be defined according to the application requirements.

Given the recent advances in image recognition and computer vision, *e.g.*, with deep learning methods (Xing et al., 2018), future studies could explore conflation of expert, crowd, and automated image labelling. Similar to the crowd standard deviation, class prediction probabilities –representing the prediction confidence of the model– could serve as a proxy to inform a conflation strategy.

4.4. Recommendations for hybrid data collection

Enhancing expert data sets with larger sets of crowdsourced contributions as proposed in this paper raises several challenges, mostly related to data quality and heterogeneity. Therefore, several interventions, *e.g.*, corrections and verifications (Fonte et al., 2015), must be implemented to increase the degree of reliability of the two data collection approaches while ensuring efficiency of the process and limiting intervention costs. Based on our results and on the literature discussed previously, a set of evidence-based guidelines can be formulated to move towards seamless integration of expert and crowd observations:

1) An **initialization set** consisting of a small number of sampling units should first be interpreted by experts to rank the volunteers and reward them not only based on the quantity but also on the quality of their contributions. *A priori* knowledge could inform the distribution of these sampling units in space so as to cover a wide range of complexity. In the present study, 2000 control sampling units were randomly selected and interpreted by a group of three trained interpreters (Laso Bayas et al., 2017). Sampling units where disagreement surpassed 12% (3 out of 25 sub-pixels) were discarded for quality purposes.
2) The **crowdsourcing campaign** is run using a user-friendly data collection tool designed to reduce task and interpretation complexity as well as to allow swift data collection. Volunteers should be trained to use the tool and are introduced to the legend. Gamification of the data collection as well as incentives should be considered to sustain their engagement (Fritz et al., 2017; Laso Bayas et al., 2016). In our crowdsourcing campaign, we prepared a tutorial video (https://bit.ly/2wCMUza), implemented a dynamic leader board, and rewarded volunteers with Amazon vouchers or co-authorship (Laso Bayas et al., 2017). Volunteers were also encouraged to share experiences and difficult locations with the wider community on social media.
3) Once the objective of the crowdsourcing campaign is met, one can strategically allocate experts to those sampling units where the crowd lacks consensus. Such **targeted verification of uncertain crowd observations by experts** should rely on a proxy variable for crowdsourcing reliability such as the crowd standard deviation. Experts should be provided training materials and clear guidelines on how to use confidence labels. A similar mechanism is implemented in the Virtual Interpretation of Earth Web-Interface Tool where conflicts in the majority cover class are reviewed by expert-level users whose input supersedes that from the users (Clark and Aide, 2011). Our results suggest that about 1200 sampling units (~30% of the sample) require expert verification. Additionally, sampling units with intermediate proportions (25–75%) could also be verified. It is likely that the additional workload is low because of the correlation between the crowd standard deviation and intermediate class proportions. For instance, only 11 additional sampling units would need to be verified in our data set.
4) The strategic allocation of experts allows to collect multiple expert observations per sampling unit contribution at no additional cost, thereby building up confidence in their interpretations. This **verification of expert observations by other experts** would help reach consensus (Powell et al., 2004) and ensures collection of reference data with high accuracy standards. This review process could occur for all sampling units or strategically. In the second case, confidence labels or large divergences between the crowd and the experts are criteria that ought to be considered in the selection process. In this study, a conflation rate of 70% could safely be realised. Therefore the remaining 30% of sampling units could be interpreted by two to three experts with an equivalent effort.

Conflation of crowdsourced and expert data under the umbrella of a collaborative global land cover information service would enable efficient collection and sharing of validation data, as well as further enhancing the value-added applications of land cover information (Chen et al., 2017).

5. Conclusions

In this study, we investigated whether the conflation of expert and crowd reference data collected *via* photo-interpretation to validate global binary maps was practically viable. To that aim, expert and crowd observations were collected at >4000 locations following a stratified systematic sampling approach. The stratified systematic sampling is based on a pattern of replicates that facilitates sample size adjustments and the stratification defines areas that are *a priori* problematic, so that the sampling rate could be increased in error-prone areas. The overall agreement between the experts and the crowd was high but varied by stratum and according to landscape complexity. Crowdsourcing appeared particularly cost-effective in areas that were easy to interpret and allowed difficult or problematic sampling units to be identified, *i.e.*, as evidenced by a lack of consensus between volunteers. Results suggest that crowd contributions can be integrated with validation data sets collected by experts but total conflation is not recommended. Partial conflation, however, maintains the accuracy standard of the expert data when informed by an indicator of the reliability of the crowdsourced labels. The crowd standard deviation, which indicates the level of consensus between volunteers for a specific sampling unit, was shown to be a good indicator of reliability of the crowd observations. We illustrated the proposed approach by validating two global cropland maps derived from Globeland 30 and GFSAD, and estimated that 70% of the expert data could be crowdsourced with little to no effect on the accuracy estimates. As a result, the spared expert effort can then be re-invested to strengthen the confidence of the expert contributions that were not conflated. We conclude that experts and crowd reference data collection should be integrated at the sampling and data analytics levels. While the approach presented here focused on binary assessment, the recommendations remain valid for multi-class validation.

Acknowledgements

This work was achieved within the framework of the SIGMA (Stimulating Innovation for Global Monitoring of Agriculture; http://www.geoglam-sigma.info/) project that received funding from the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement No 603719. SIGMA is a European contribution to the GEOGLAM (GLOBAL Agricultural Geo-Monitoring) initiative (https://www.earthobservations.org/geoglam.php). The authors would like to thank Roel Van Hoolst and all the volunteers who contributed to the classification of the reference sampling units. Partial support was also provided by the EU-funded ERC CrowdLand project under grant agreement No 617754. The authors thank the anonymous reviewers for their constructive comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2018.10.039.

References

Arino, O., et al., 2008. GlobCover: the most detailed picture of earth. In: ESA Bulletin-European Space Agency, pp. 136.
Bastin, L., Buchanan, G., Beresford, A., Pekel, J.F., Dubois, G., 2013. Open-source mapping and Services for web-based land-cover validation. Eco. Inform. 14, 9–16.
Bellhouse, D.R., 2014. Systematic sampling methods. In: Wiley StatsRef: Statistics Reference Online.
Bey, A., et al., 2016. Collect earth: land use and land cover assessment through augmented visual interpretation. Remote Sens. 8 (10), 807.
Bicheron, P., et al., 2008. GLOBCOVER Products Report Description and Products Description and Validation Report. 33(0). ESA Globcover Project Led by MEDIAS FrancePOSTEL, pp. 140–147.
Bogaert, P., Waldner, F., Defourny, P., 2016. An information-based criterion to measure pixel-level thematic uncertainty in land cover classifications. Stoch. Env. Res. Risk A. 1–16.
Bontemps, S., Defourny, P., Van Bogaert, E., Kalogirou, V., Perez, J.R., 2011. GLOBCOVER 2009 products description and validation report. ESA Bull. 1–53.
Bontemps, S., et al., 2013. Consistent global land cover maps for climate modelling communities: current achievements of the ESA' land cover CCI. In: ESA Living Planet Symposium, Retrieved. http://adsabs.harvard.edu/abs/2013ESASP.722E..62B.
Chen, J., et al., 2015. Global land cover mapping at 30 m resolution: a pok-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27. Retrieved. https://doi.org/10.1016/j.isprsjprs.2014.09.002.
Chen, J., Li, S., Wu, H., Chen, X., 2017. Towards a collaborative global land cover information service. Int. J. Digital Earth 10 (4), 356–370. Retrieved July 12, 2018. https://www.tandfonline.com/doi/full/10.1080/17538947.2016.1267268.
Clark, M.L., Aide, T.M., 2011. Virtual interpretation of earth web-Interface tool (VIEW-IT) for collecting land-use/land-cover reference data. Remote Sens. 3 (3), 601–620.
Comber, A., et al., 2013. Using control data to determine the reliability of volunteered geographic information about land cover. Int. J. Appl. Earth Obs. Geoinf. 23 (1), 37–48.
Danylo, O., et al., 2018. The picture pile tool for rapid image assessment: a demonstration using hurricane matthew. In: ISPRS Technical Commission IV Symposium, (Delft).
Defourny, P., Mayaux, P., Herold, M., Bontemps, S., 2012. Global land cover map validation experiences: toward the characterization of quantitative uncertainty. In: Remote Sensing of Land Use and Land Cover: Principles and Applications, pp. 207–223. Retrieved. http://dial.uclouvain.be/handle/boreal:114555.
Dierckx, W., et al., 2014. PROBA-V mission for global vegetation monitoring: standard products and image quality. Int. J. Remote Sens. 35 (7), 2589–2614.
Dunn, R., Harrison, A.R., 1993. Two-dimensional systematic sampling of land use. J. R. Stat. Soc.: Ser. C: Appl. Stat. 42, 585–601.
European Environment Agency, 2015. Copernicus Land Service – Pan-European Component: High Resolution Layers. pp. 2.
FAO, 2017. Global Strategy for Improving Agricultural & Rural Statistics. http://gsars.org/en/.
Fonte, C.C., Bastin, L., See, L., Foody, G., Lupia, F., 2015. Usability of VGI for validation of land cover maps. Int. J. Geogr. Inf. Sci. 29 (7), 1269–1291.
Foody, G.M., 2011. Impacts of imperfect reference data on the apparent accuracy of species presence-absence models and their predictions. Glob. Ecol. Biogeogr. 20 (3), 498–508.
Foody, G.M., 2013. Ground reference data error and the Mis-estimation of the area of land cover change as a function of its abundance. Remote Sens. Lett. 4 (8), 783–792.
Foody, G., et al., 2013. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. Trans. GIS 17 (6), 847–860.
Foody, G., et al., 2018. Increasing the accuracy of crowdsourced information on land cover via a voting procedure weighted by information inferred from the contributed data. ISPRS Int. J. Geo-Inf. 7 (3), 80. Retrieved July 6, 2018. http://www.mdpi.com/2220-9964/7/3/80.
Fritz, S., et al., 2009. Geo-Wiki.Org: the use of crowdsourcing to improve global land cover. Remote Sens. 1 (3), 345–354. Retrieved August 12, 2014. http://www.mdpi.com/2072-4292/1/3/345/.
Fritz, S., et al., 2011. Highlighting continued uncertainty in global land cover maps for the user community. Environ. Res. Lett. 6 (4), 044005 Retrieved September 6, 2014. http://stacks.iop.org/1748-9326/6/i=4/a=044005?key=crossref.e37ea06a1a56e58be6fbd1f907045b97.
Fritz, S., et al., 2013. Downgrading recent estimates of land available for biofuel production. Environ. Sci. Technol. 47 (3), 1688–1694.
Fritz, S., et al., 2015. Mapping global cropland and field size. Glob. Chang. Biol. 21 (5), 1980–1992.
Fritz, S., See, L., Brovelli, M.A., 2017. Motivating and sustaining participation in VGI. In: Mapping and the Citizen Sensor, pp. 93–117.
Gallego, J., Delincé, J., 2010. The european land use and cover area-frame statistical survey. In: Agricultural Survey Methods, pp. 149–168.
Gallego, J., Schucknecht, A., Waldner, F., 2016. Sampling to validate a global cropland map. In: Proceedings of Spatial Accuracy 2016.
GOFC-GOLD LC Office, 2015. The GOFC-GOLD reference data portal. Retrieved July 3, 2017. http://www.gofcgold.wur.nl/sites/gofcgold_refdataportal.php.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4), 211–221.
Gumma, M.K., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30-M for South Asia, Afghanistan and Iran. NASA EOSDIS Land Processes DAAC.
Howe, J., 2015. The wisdom of crowds. Wired Mag. 161 (4), 697.
Iwao, K., Nishida, K., Kinoshita, T., Yamagata, Y., 2006. Validating land cover maps with degree confluence project information. Geophys. Res. Lett. 33 (23).
JECAM, 2015. Guidelines for Field Data Collection. Retrieved. http://www.jecam.org/JECAM_Guidelines_for_Field_Data_Collection_v1_0.pdf.
Khatami, R., Mountrakis, G., Stehman, S.V., 2017. Mapping per-pixel predicted accuracy of classified remote sensing images. Remote Sens. Environ. 191, 156–167. Retrieved September 2, 2018. https://www.sciencedirect.com/science/article/pii/S0034425717300378.
Lamarche, C., et al., 2017. 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. 9 (1), 36. Retrieved July 4, 2017. http://www.mdpi.com/2072-4292/9/1/36.
Laso Bayas, J.C., et al., 2016. Crowdsourcing in-situ data on land cover and land use using gamification and Mobile technology. Remote Sens. 8 (11), 905. Retrieved July 13, 2018. http://www.mdpi.com/2072-4292/8/11/905.
Laso Bayas, J.C., et al., 2017. A global reference database of crowdsourced cropland data collected using the Geo-Wiki platform. Sci. Data 4.
de Leeuw, J., et al., 2011. An assessment of the accuracy of volunteered road map production in Western Kenya. Remote Sens. 3 (2), 247–256.
Lesiv, M., et al., 2018. Characterizing the Spatial and Temporal Availability of Very High Resolution Satellite Imagery in Google Earth and Microsoft Bing Maps as a Source of Reference Data. Land 7.4, 118.
Lin, Lawrence I-Kuei, 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.
Löw, F., et al., 2018. Mapping cropland abandonment in the Aral Sea basin with MODIS time series. Remote Sens. 10 (2).
Mann, R.P., Helbing, D., 2017. Optimal incentives for collective intelligence. PNAS 114 (20), 5077–5082. Retrieved. http://arxiv.org/abs/1611.03899%0A. https://doi.org/10.1073/pnas.1618722114.
Massey, R., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30m for North America: Cropland Extent Product (GFSAD30NACE). NASA EOSDIS Land Processes DAAChttps://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30NACE.001. Retrieved February 14, 2018.
Mayaux, P., et al., 2006. Validation of the global land cover 2000 map. IEEE Trans. Geosci. Remote Sens. 44 (7), 1728–1737.
Morton, D., et al., 2006. Cropland expansion changes deforestation dynamics in the Southern Brazilian amazon. In: Proceedings of the National Academy of Sciences of the United States of America. 103(39). pp. 14637–14641. Retrieved. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1600012&tool=pmcentrez&rendertype=abstract.
Oliphant, A.J., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30-m for Southeast & Northeast Asia: Cropland Extent Product (GFSAD30SEACE). NASA EOSDIS Land Processes DAAC.
Olofsson, P., et al., 2012. A global land-cover validation data set, part I: fundamental design principles. Int. J. Remote Sens. 33 (18), 5768–5788. Retrieved July 12, 2018. http://www.tandfonline.com/doi/abs/10.1080/01431161.2012.674230.
Olofsson, P., et al., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.
Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540 (7633), 418–422. Retrieved July 4, 2017. http://www.nature.com/doifinder/10.1038/nature20584.
Pesaresi, M., et al., 2013. A global human settlement layer from optical HR/VHR RS data: concept and first results. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6 (5), 2102–2131. Retrieved July 4, 2017. http://ieeexplore.ieee.org/document/6578177/.
Phalke, A., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30-m for Europe, Middle-East, Russia and Central Asia: Cropland Extent Product

(GFSAD30EUCEARUMECE). NASA EOSDIS Land Processes DAAChttps://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30EUCEARUMECE.001. Retrieved February 14, 2018.
Pontius, R., Thontteh, O., Chen, H., 2008. Components of information for multiple resolution comparison between maps that share a real variable. Environ. Ecol. Stat. 15 (2), 111–142.
Powell, R.L., et al., 2004. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. Remote Sens. Environ. 90 (2), 221–234.
Prelec, D., Seung, H.S., McCoy, J., 2017. A solution to the single-question crowd wisdom problem. Nature 541 (7638), 532–535.
Schepaschenko, D., et al., 2015. Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. Remote Sens. Environ. 162, 208–220.
See, L., et al., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS One 8 (7).
See, L., et al., 2014. Cropland capture: a gaming approach to improve global land cover. In: Connecting a Digital Europe Through Location and Place. Proceedings of the AGILE 2014 International Conference on Geographic Information Science, pp. 3–6.
See, L., et al., 2015. Harnessing the power of volunteers, the internet and Google earth to collect and validate global spatial information using geo-wiki. Technol. Forecast. Soc. Chang. 98, 324–335.
Stehman, S.V., 2009. Sampling designs for accuracy assessment of land cover. Int. J. Remote Sens. 30 (20), 5243–5272.
Stehman, S.V., Olofsson, P., Woodcock, C.E., Herold, Martin, Friedl, Mark A., 2012. A global land-cover validation data set, II: augmenting a stratified sampling design to estimate accuracy by region and land-cover class. Int. J. Remote Sens. 33 (22), 6975–6993. Retrieved July 12, 2018. http://www.tandfonline.com/doi/abs/10.1080/01431161.2012.695092.
Stehman, Stephen V., Fonte, Cidália C., Foody, Giles M., See, Linda, 2018. Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover. Remote Sens. Environ. 212, 47–59. Retrieved July 6, 2018. https://www.sciencedirect.com/science/article/pii/S0034425718301627.
Strahler, A.H., et al., 2006. Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps.
Strand, G.H., Dramstad, W., Engan, G., 2002. The effect of field experience on the accuracy of identifying land cover types in aerial photographs. Int. J. Appl. Earth Obs. Geoinf. 4 (2), 137–146.
Taubenböck, H., Wegmann, M., Roth, A., Mehl, H., Dech, S., 2009. Urbanization in India – spatiotemporal analysis using remote sensing data. Comput. Environ. Urban. Syst. 33 (3), 179–188. Retrieved July 12, 2018. https://www.sciencedirect.com/science/article/pii/S0198971508000604#fig3.
Teluguntla, P., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30-m for Australia, New Zealand, China, and Mongolia: Cropland Extent Product (GFSAD30AUNZCNMOCE). NASA EOSDIS Land Processes DAAChttps://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30AUNZCNMOCE.001. Retrieved February 14, 2018.
Van Coillie, F.M.B., et al., 2014. Variability of operator performance in remote-sensing image interpretation: the importance of human and external factors. Int. J. Remote Sens. 35 (2), 754–778.
Vancutsem, C., Marinho, E., Kayitakire, F., See, L., Fritz, S., 2012. Harmonizing and combining existing land cover/land use datasets for cropland area monitoring at the African continental scale. Remote Sens. 5 (1), 19–41. Retrieved December 27, 2012. http://www.mdpi.com/2072-4292/5/1/19/.
Waldner, F., Fritz, S., Di Gregorio, A., Defourny, P., 2015. Mapping priorities to focus cropland mapping activities: fitness assessment of existing global, regional and National Cropland Maps. Remote Sens. 7 (6), 7959–7986.
Waldner, F., et al., 2016. Towards a set of agrosystem-specific cropland mapping methods to address the global cropland diversity. Int. J. Remote Sens. 37 (14), 3196–3231.
Wolter, K.M., 1984. An investigation of some estimators of variance for systematic sampling. J. Am. Stat. Assoc. 79 (388), 781–790.
Woodcock, C.E., Gopal, S., 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. Int. J. Geogr. Inf. Sci. 14 (2), 153–172.
Xing, H., Meng, Y., Wang, Z., Fan, K., Hou, D., 2018. Exploring geo-tagged photos for land cover validation with deep learning. ISPRS J. Photogramm. Remote Sens. 141, 237–251. Retrieved July 13, 2018. https://www.sciencedirect.com/science/article/pii/S0924271618301333.
Xiong, J., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30-m Africa: Cropland Extent Product (GFSAD30AFCE). NASA EOSDIS Land Processes DAAChttps://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30AFCE.001. Retrieved February 14, 2018.
Yu, L., et al., 2013. FROM-GC: 30 m global cropland extent derived through multisource data integration. Int. J. Digital Earth 6 (6), 521–533. Retrieved September 13, 2018. http://www.tandfonline.com/doi/abs/10.1080/17538947.2013.822574.
Zhong, Y., et al., 2017. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-Support Analysis Data (GFSAD) @ 30-m for South America: Cropland Extent Product (GFSAD30SACE). NASA EOSDIS Land Processes DAAChttps://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30SACE.001. Retrieved February 14, 2018.