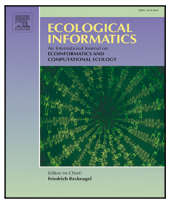
Ecological Informatics 91 (2025) 103300

Contents lists available at ScienceDirect

Ecological Informatics

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



Factors influencing the consistency in crowdsourced interpretations of aerial photographs to measure tree canopy cover

Jill M. Derwin [a], Valerie A. Thomas [a,*], Randolph H. Wynne [a], Karen G. Schleeweis [e], John W. Coulston [d], S. Seth Peery [b], Kurt Luther [c], Greg C. Liknes [f], Stacie Bender [g], Susmita Sen [a]

[a] Virginia Tech, Forest Resources and Environmental Conservation, Blacksburg, VA 24061, USA
[b] Virginia Tech, Enterprise GIS, Blacksburg, VA 24060, USA
[c] Virginia Tech, Department of Computer Science, Alexandria, VA 22305, USA
[d] United States Forest Service, Southern Research Station, Blacksburg, VA 24061, USA
[e] United States Forest Service, Rocky Mountain Research Station, Ogden, UT 84401, USA
[f] United States Forest Service, Northern Research Station, St Paul, MN 55108, USA
[g] United States Forest Service, Geospatial Technology and Applications Center, Salt Lake City, UT 84119, USA

ARTICLE INFO

Keywords:
Human intelligence task
Citizen science
Map accuracy
Worker reliability
Ensemble models
Machine learning
Crowd sourcing

ABSTRACT

Machine learning models are typically data-hungry algorithms that require large data inputs for training. When they produce wall-to-wall remote sensing products, model validation also requires large sets of temporally harmonized field observations. Crowdsourcing may offer a potential solution for the collection of photointerpretations for the training and validation of spatial models of tree canopy cover (TCC), as it harnesses the power of a large anonymous crowd in the completion of repetitive discrete analyses or human intelligence tasks (HITs). This study explores the factors that determine the consistency of TCC interpretations collected by an anonymous crowd to those collected by a control group. The crowd interpretations were obtained through an anonymous platform with a task-reward framework, while those collected by the control group were collected by known interpreters in a more traditional setting. Both groups carried out this task using an interface developed for Amazon's Mechanical Turk platform. We collected multiple interpretations at sample plot locations from both crowd and control interpreters, and sampled these data in a Monte Carlo framework to estimate a classification model predicting the consistency of each crowd interpretation with control interpretations. Using this model, we identified the most important variables in estimating the relationship between a location's characteristics and interpretation behaviors which affect consistency in interpretations between crowd workers our control group. Overall, we show low agreement between crowdsourced and control interpretations, as well as interpretations from individual control group members. This warrants caution in considering the crowdsourced photointerpretation of TCC as a data source for model training and validation without adequate interpreter training as well as significant quality control measures and consistency standards. We show that the number of plots interpreted was the strongest indicator of the reliability of an individual's interpretations, further evidenced by apparent fatigue effects in crowd interpretations. The second most important variable related to the use of the false color display during interpretation followed by a variable related to the use of the natural color display during interpretation, reflecting the differences in interpretation methodologies used by crowd workers and control group interpreters and the impact display has on the interpretation of tree canopy cover. Finally, we discuss recommendations for further study and future implementations of crowdsourced photointerpretation. These include the enhanced use of existing mechanisms within Mechanical Turk such as worker qualifications to identify and reward more attentive workers, as well as enhanced attention to quality control measures throughout the data collection process and measures to increase intrinsic motivation. For our study we also recommend a minimum time on task or other measures to reduce the punishment of access to HITs for workers who took their time providing detailed interpretations.

* Corresponding author.
E-mail addresses: jmd59@vt.edu (J.M. Derwin), thomasv@vt.edu (V.A. Thomas), wynne@vt.edu (R.H. Wynne), karen.schleeweis@usda.gov (K.G. Schleeweis), john.coulston@usda.gov (J.W. Coulston), sspeery@vt.edu (S.S. Peery), kluther@vt.edu (K. Luther), greg.liknes@usda.gov (G.C. Liknes), stacie.bender@usda.gov (S. Bender), susmita@vt.edu (S. Sen).

https://doi.org/10.1016/j.ecoinf.2025.103300
Received 18 February 2025; Received in revised form 23 June 2025; Accepted 23 June 2025
Available online 10 July 2025

We also recommend using optimized default interface settings instead of providing a variety of options to the interpreter.

1. Introduction

Modern environmental research relies on the ability of scientists to model the relationships between various ecosystem components in order to monitor trends, understand the current state of our resources, and to more accurately forecast potential threats and facilitate improved decision-making. As algorithms become more complex and study areas become more expansive, the models used to estimate these relationships require the collection of more training and validation data, a process that is usually time-consuming and costly (Congalton and Green, 2009; Tipton et al., 2012; McRoberts and Tomppo, 2007). For remote sensing applications, harmonizing the collection of in situ observations with satellite imagery could pose additional logistical challenges and costs (Congalton and Green, 2009), particularly over large spatial scales. Crowdsourcing presents a potential source of high quality training data, particularly for applications like tree canopy cover (TCC) modeling, where predictor variables can be interpreted from high quality aerial imagery in place of field observations.

Crowdsourcing is a method of data collection which has risen to prominence as an avenue to reduce time and cost of data collection by "outsourcing tasks…to an undefinably large, heterogeneous mass of potential actors" (Anon, 2012). Crowdsourcing processes, called Human Intelligence Tasks (HITs) typically utilize an anonymous voluntary crowd to meet a specific goal for an identified solicitor. If performed by experts, these tasks would be considerably more expensive (Haklay and Weber, 2008). HITs are typically difficult to automate, therefore requiring some activity be performed by a human actor. To reduce bias most HITs are designed to involve only simple, objective analyses (Galloway et al., 2006). While photointerpretation typically relies on experts to interpret imagery, crowdsourcing could be operationalized to collect interpretations even more quickly to better meet modern data needs, or to cost-effectively obtain multiple interpretations for data aggregation methods.

TCC is defined as the "proportion of the forest floor covered by the vertical projection of the tree crowns" (Jennings et al., 1999). It is an important measurement in forest management, as it can provide insight into forest health, disease, degradation, growth, and productivity. One method for the collection of TCC values is photointerpretation, which uses high resolution aerial imagery. Even when employed by forestry experts, this approach reduces the cost of data collection by minimizing travel time and resources. Esseen et al. (2006) show that photointerpretation of forest edge lengths was found to be eight times faster and five times more cost effective than field measurements (Esseen et al., 2006). Additionally, where constraints to physical access to sites could bias sample design, photointerpretation can often still be performed (Congalton, 1991). To obtain photointerpreted measures of TCC, interpreters use a dot grid over an image depicting a defined spatial area. The interpreter then decides if each dot falls on a tree canopy or non-tree canopy feature in the image. The percentage of dots that fall on tree canopies yields a TCC value. Expert-interpreted TCC values have been collected from imagery across the United States to train and test nationwide models for the National Land Cover Database (NLCD) since 2011 (Coulston et al., 2012; Derwin et al., 2020). Photointerpreted TCC measurements are suitable for model training and validation, and can be collected more efficiently than the common field methods. However, collecting these interpretations at all of the 364,240 photo plot locations across the United States that made up the sampling frame for this study would still require considerable cost and time. To address this, we assess the potential of crowdsourcing to obtain TCC interpretations.

This study builds on a body of research which has employed crowdsourcing to collect geographic information by way of classification, digitization, and conflation of satellite or aerial imagery (de Albuquerque et al., 2016), and to improve and validate existing remote sensing models. Yu et al. (2015) demonstrates the value of crowdsourced photointerpretations for the identification of cloud-covered pixels in Landsat imagery. This study showed that crowd interpretations improve cloud and cloud-shadow detection. They show that crowd interpreters were more successful at identifying and delineating clouds compared to automated models, particularly when using a majority voting rule across interpreters (Yu et al., 2015). Additionally, the platform Geo-wiki.org uses imagery from Google Earth to collect data across the world. Geo-wiki.org has been used to generate training data for classification models and to improve upon existing classifications for global landcover (Fritz et al., 2009) models in regions where there is disagreement. Schepaschenko et al. (2015) created a forest probability layer and a map of percent forest cover from Geo-wiki.org's interpretations in combination with several commonly used models of forest cover. Their forest classification model has an overall accuracy of 93% and has an R^2 of 0.87 which rivals or exceeds the accuracy of all other models tested (Schepaschenko et al., 2015).

There remains skepticism that crowd workers can meet the high data quality standards of science and produce data that is comparable to experts' (Wiggins and Crowston, 2011), however several studies address this. See 2013, which also utilized data from Geowiki.org to identify areas impacted by humans and landcover, compared interpretations from more than 60 crowd workers to a set of three expert interpreters at control points where initial analysis suggested non-experts should be able to interpret these features, and collected interpreter ratings of confidence in interpretation (See et al., 2013). Their analysis shows that volunteers underestimated the degree of human impact in most land cover types, but they were more consistent than experts in the estimates of human impact they did provide, and they were also slightly more consistent in their interpretation of landcover type. Experts were more adept at classifying landcover than non-experts, showing a higher users and producers accuracies for most classes across relevant control point groups. They found that, for all interpreters, decisions were more consistent in locations where they were highly confident.

An interpreter could also have various personal qualities that could improve photointerpretation performance. Van Coillie et al. (2014) cites a long history of psychological studies related to remote sensing image interpretation. The earliest examples relate to 'vigilance', or the opposite of fatigue, observed in WWII operators watching radar screens (Van Coillie et al., 2014). In these studies, fatigue effects result in a drop in performance over time (Parasuraman, 1986) as referenced by Van Coillie et al. (2014). Van Coillie et al. (2014) also reviewed research that demonstrated the importance of task-specific training to create higher levels of agreement between interpreters (Cooper et al., 2007), and that training for systematic methods of interpretation as opposed to random search caused higher performance (Wang et al., 1997). These authors measured the effects of these influences using photointerpretation by enlisting volunteers to digitize various ground features from remote sensing imagery. The results demonstrate that human factors such as intrinsic motivators and personal characteristics, including experience and the amount of time the interpreter was willing to spend were responsible for 26% of inter-individual differences in performance. They also observed a clear fatigue effect (Van Coillie et al., 2014). Fatigue affects were also observed in Yu, 2015, where they showed that interpretation accuracy deteriorated with time on task (Yu et al., 2015). This study also demonstrated that quality of interpretations was insensitive to wage, an extrinsic motivator, which

is a finding also reflected in other studies (Mason and Watts, 2010; Buhrmester et al., 2011).

Successful photointerpretation requires some experience, such as knowledge of subject, geographic region, and remote sensing system (Campbell and Wynne, 2011). Due to the anonymous nature of crowdsourcing, crowd workers traditionally have unknown or variable levels of experience with all of these. To increase the likelihood of high quality interpretations, voting mechanisms are common in the analysis of crowdsourced observations.

1.1. Objectives

We evaluated the effectiveness of crowdsourcing for the collection of validation-quality TCC interpretations. The primary objective of this study was to determine how "reliable" photointerpreted TCC values can be when collected by crowd workers on Amazon's Mechanical Turk. We deployed multiple requests at each sample plot location to better understand how workers perform individually and in aggregate. Second, we explored possible factors that influence that reliability. For this study, the "reliability" of crowd interpretations was determined by their consistency with control interpretations.

To accomplish this, we measured the differences between interpretations collected by a non-crowd control group of interpreters and those collected by crowd workers. Our control group was made up of three individuals with prior interpretation experience in forestry, no profit motive or time limitations, and general familiarity with the landscapes in our sample. To understand the factors which impact the differences in interpretations between these groups, we identified the plot-specific variables and worker-interpretation-behavior variables that influence the level of reliability in crowd interpretations as they compare to the control group's interpretations. We have used this information to propose recommendations for the design of future studies in order to improve consistency and agreement in photointerpretation of TCC by crowd workers.

2. Data and methods

2.1. Study area and sample site selection

This study was conducted at sites across the conterminous United States. Public coordinates from the Forest Inventory and Analysis (FIA) program were selected as a sampling frame since the continuous hexagonal sampling grid ensures that our plots would not be clustered, and is unlikely to be aligned with regularly spaced landscape features (Brand et al., 2000). However areas that were predicted to have 0% TCC based on the 2016 USFS Analytical Tree Canopy Cover product (United States Department of Agriculture, Forest Service, 2019; Ruefnacht et al., 2022) were masked from sampling as well as pixels categorized as background, non-tree, water, ice, snow, and agriculture. 0% TCC pixels represented a large portion of the landscape and therefore would have represented a large portion of our sample, however the interface we developed would not have yielded any information about interpretation behavior for plots that were interpreted to have 0% TCC, and, though time required to interpret a 0% TCC plot is negligible, the payment would have been the same, resulting in a relatively costly set of data yielding little information. Therefore we opted to remove those areas.

As stated in our objectives, one application of this study was to assess the value of crowdsourced interpretations for the validation of a continuous remote sensing product. As such, the target variable for this analysis was model error. We used a stratified random sampling protocol with strata delineated using the 1% classes from the standard error layer provided with the USFS Analytical Tree Canopy Cover product (United States Department of Agriculture, Forest Service, 2019; Ruefnacht et al., 2022) (See Fig. 1). By stratifying on the this layer, we would be minimizing the variance in the expected error within each stratum since the standard error of predictions across random forest trees is likely related to model error. The standard error layer was created by calculating the standard error of the many predictions produced by the random forest model at each pixel location in the TCC model. The resulting map represents the prediction uncertainty of each pixel. In this case, the standard error layer ranges from 0 to 45% TCC in 2011 and from 0 to 46% TCC in 2016, yielding 46 strata for 2011 and 47 strata for 2016. Using this sample design, we hoped to select samples across the full range of possible model error values, based on the assumption that the standard error layer should follow the expected error distribution of the model. The number of plots sampled from each stratum was determined using optimal allocation which distributes sample units using a combination of within-stratum variance and stratum area (Avery and Burkhart, 1994) in order to increase sampling intensity over areas with a higher prediction uncertainty. Stratified random sampling with optimal allocation is a standard probability sampling design with known inclusion probabilities, where the number of sample plots (n_h) in stratum h is defined by:

$$n_h = n * \frac{a_h * SE_h}{\sum (a_h * SE_h)} \tag{1}$$

where n is the total sample size, a_h is the area of stratum h, a is the population area, and SE_h is the standard error within stratum h (Avery and Burkhart, 1994; Freese, 1962; Cochran, 1977). Fig. 1 outlines this mapping process and the subsequent sampling method described above. Strata definitions and standard error values were obtained from the standard error layer of the 2016 USFS Analytical Tree Canopy Cover product (United States Department of Agriculture, Forest Service, 2019).

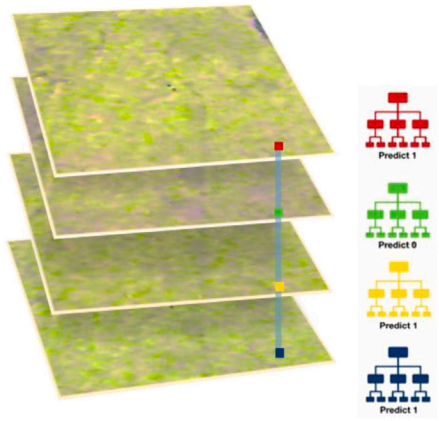
A target sample size of 2000 plots was confirmed to be appropriate to achieve the desired precision (E) and confidence interval for this study based on Eq. (2).

$$n = \frac{(\sum (a_h * SE_h))^2}{\frac{N^2 * E^2}{t^2} + \sum (a_h * SE_h^2)} \tag{2}$$

where n is the total sample size, a_h is the stratum area h, SE_h is the standard error value of stratum h, N is the population size based on the number of unmasked pixels in the population (comprising 2,989,614,429 30-meter pixels across the conterminous United States), t is *student's t* based on the 95% confidence interval and large population size (1.960), and E is the desired precision of the sample mean to the population mean (which we determined to be 1% TCC) (Avery and Burkhart, 1994). This calculation results in a value of 1384, which is significantly lower than the target sample size.

Our sampling method resulted in the allocation of fractional plots to each stratum, which we rounded to the nearest whole number to calculate n_h (See Fig. 1). After rounding these values, the total sample size across all strata for crowd interpretations is 1998. For the control sample, a subset of the crowd plots were selected using the same sample allocation method but a 25% target sample size to accommodate the limited time our control interpreters had to contribute to this study. A target sample of 500 plots was used for control sampling and after fractional plots were rounded during optimal allocation sampling the final number of control sample plots is 502.

Because we distributed our sample using stratified random sampling with optimal allocation, inclusion probabilities were defined by proportional strata areas as well as the standard error value. We hoped to concentrate our sampling in areas with higher prediction uncertainty since error values in those areas were expected to be more variable, requiring more sample plots. In order to avoid over-emphasizing areas with a high uncertainty or under-emphasizing areas with low uncertainty in our analysis, sample-wide statistics were calculated by determining the estimate within each 1% interval of standard error and then combining those estimates using a weighted average based on area.



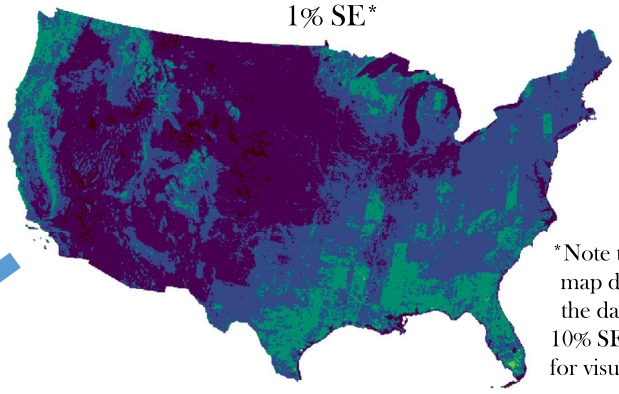
Obtain pixel-level predictions from the random forest model

Calculate the standard error (SE) for each pixel

$$SE = \frac{\sigma}{\sqrt{n}}$$

σ = the standard deviation across predictions
n = sample size

Standard error layer with each stratum (h) equal to an interval of 1% SE*



*Note that this map displays the data with 10% SE classes for visual effect

Calculate areas of each stratum (a_h), and then calculate the inclusion probabilities

SE_h	Area (ha) (a_h)	(a_h*SE_h)/∑(a_h*SE_h)	n_h=(a_h*SE_h)/∑(a_h*SE_h)*n (n=2000)	Inclusion Probability	n_h Rounded for Crowd
0-1	140048.6	2.7E-05	0.05	3.5E-08	0
2	91905.84	3.6E-05	0.07	7.0E-08	0
3	253017.9	1.5E-04	0.30	1.1E-07	0
4	590557.5	4.6E-04	0.92	1.4E-07	1
5	1249044	1.2E-03	2.45	1.8E-07	2
6	2181138	2.6E-03	5.12	2.1E-07	5
7	3654157	5.0E-03	10.02	2.5E-07	10
8	5454724	8.5E-03	17.09	2.8E-07	17
9	7300739	1.3E-02	25.73	3.2E-07	26
…	…	…	…	…	…
46	0.09	8.11E-10	1.62E-06	1.6E-6	0

Using the calculated sample size of 2000 as our target, after rounding we had 1998 plots total, distributed throughout the map

Combine classes with fewer than two samples

SE_h	n_h=(a_h*SE_h)/∑(a_h*SE_h)*n (n=2000)	n_h Rounded for Crowd
0-4	1.35	1
5	2.45	2
6	5.12	5
7	10.02	10
8	17.09	17
9	25.73	26
…	…	…
38	1.08	1
39-46	0.93	1

Fig. 1. Sample allocation workflow.

2.2. Mechanical Turk HIT interface development

Amazon's Mechanical Turk (https://www.mturk.com/) is a network interface for crowdsourcing tasks. This interface enables potential analysts to be trained for and undertake HITs for a small monetary payment. The preliminary design of the Amazon Mechanical Turk interface used for this study has two sectors. A local interface which hosts spatial data such as plot locations, dot grids, and NAIP imagery (from existing ArcGIS online web services (Esri and USDA Farm Service Agency, 2014b,a)) is located on Virginia Tech's GIS servers (as in Yu et al. 2015) (Yu et al., 2015), and a Mechanical Turk Shell, which has the capacity for recruitment, worker identification, and payment, interpretation tasks, database output, and interface interaction monitoring. The design harmonizes these two sectors into a single interface using ESRI's ArcGIS API for Java script.

Color-infrared NAIP imagery was obtained using ArcGIS online with dates close to 2011 for time 1 and 2016 for time 2, corresponding to timing for the 2011 and 2016 USFS Analytical Tree Canopy Cover products. 1 m spatial resolution NAIP imagery is collected by the USDA Farm Service Agency from aerial surveys with partial coverage each year. Since 2009, coverage of NAIP imagery is scheduled every three years for the contiguous lower 48 US states. Images are collected on clear days to ensure that images have less than 10 percent cloud cover per quarter quad tile. The images used for this study have three visible bands and one near infrared band, enabling a natural-color or false-color composite (USDA Farm Service Agency, 2018)

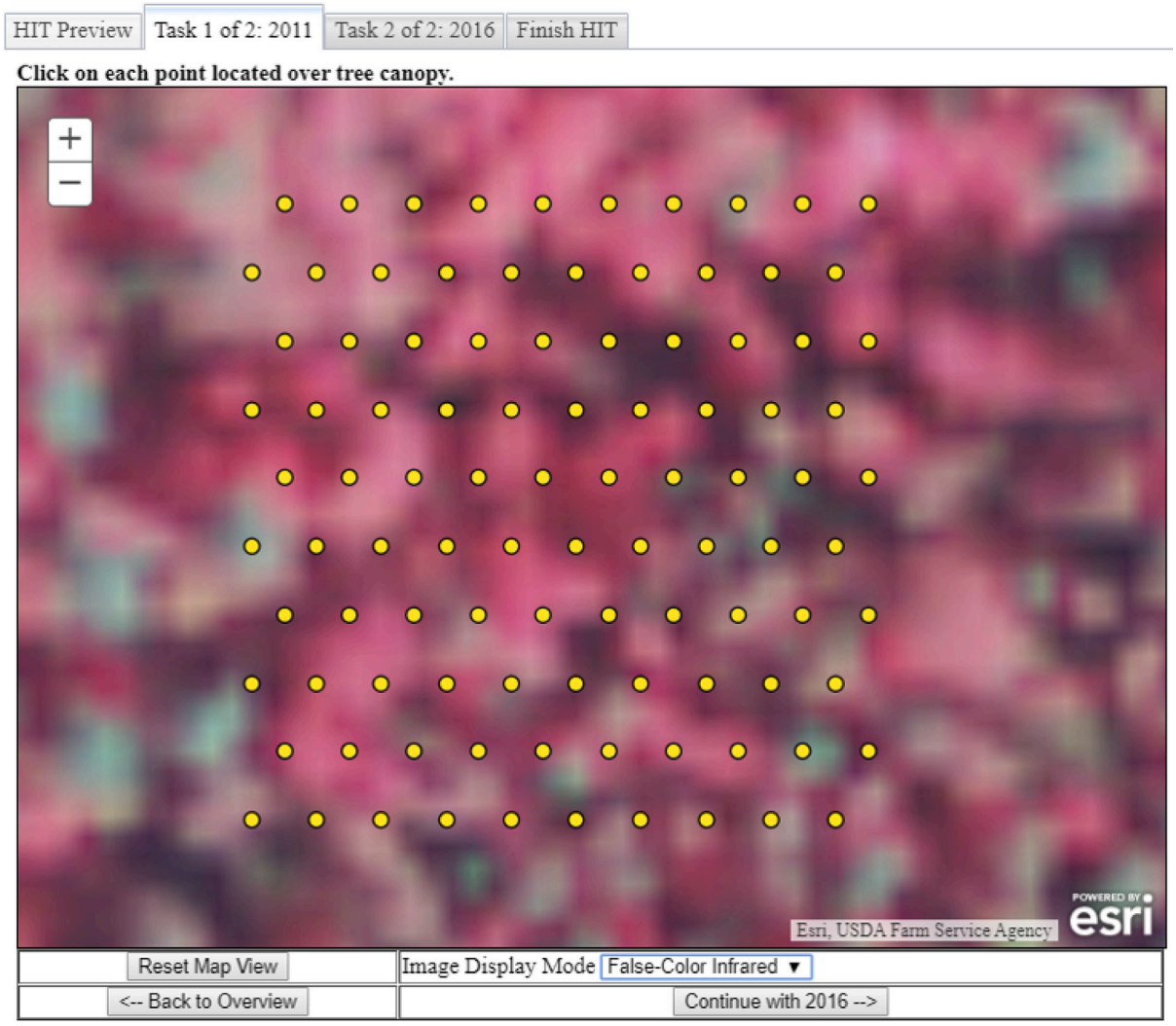


Fig. 2. Mechanical Turk interface showing 2011 image in false color display.

For this research, we served 1998 HITs, one for each sample plot selected. Each HIT involves the interpretation of a 100-dot grid at each plot location for two time periods. The dot grid consists of evenly-spaced dots aligned with a 30-m Landsat pixel which intersects the interpretation plot location. This dot grid covers a smaller area and uses fewer dots compared to the training data described in Coulston et al. (2012) and the data used for the USFS Analytical Tree Canopy Cover product, as described in Ruefnacht et al. (2022) but our objective was to obtain a user-friendly validation. We chose a 30-m grid area to provide a validation at the pixel-level, and we chose to use 100 dots so that our validation would be provided within 1% TCC, which is the precision of the 2016 USFS Analytical Tree Canopy Cover product being tested (United States Department of Agriculture, Forest Service, 2019; Coulston et al., 2012; Ruefnacht et al., 2022).

Upon the completion of each HIT, a crowd worker received a payment of $0.22 and Amazon received a 20% payment of $0.05. We took reasonable measures to ensure that interpretations were completed by different interpreters. There were 9 available assignments for each HIT, meaning that 9 different interpreter accounts accessed and completed the same plot interpretation. One interpreter's account cannot access and interpret the same HIT twice, however if an interpreter had multiple accounts they may have been able to work the same HIT twice.

During interpretation, the interpreter could change the display between natural color and false color, zoom in and out, and reset the zoom to be centered around the dot grid. To interpret TCC on the provided plot, interpreters would click on the dots in the dot grid. The first time the interpreter clicked a dot, the color would turn blue and the designation would change to indicate that the dot fell on tree canopy. If they clicked a dot a second time, this would revert the dot's appearance and change the designation to indicate that the dot did not fall on tree canopy. Once the interpreter completed the 2011 interpretation they would repeat the process, this time using a 2016 NAIP image. They could also return to the 2011 interpretation to review and modify their work. Once both 2011 and 2016 interpretations were completed, the interpreter would certify and submit their HIT (Fig. 2).

The back-end of the interface consisted of a database which collected information on the interpretations for each dot from each interpreter. In addition to the interpretation, when an interpreter clicked on a dot, the interface collected the display status (false color/natural color), the image scale, and a time-stamp. The data were organized by plot ID and dot ID and additionally included fields with the x and y coordinates of each dot. After each phase of this study, these data were compiled into a JSON file and imported into R where they were read to a table.

A training document was developed by our team in order to ensure that both control and crowd interpreters had the background needed to understand and complete this task. This document was available for review at any time from a link included in the interpretation interface. To qualify for this HIT, crowd workers were expected to review the training document and then pass a qualification exam, which consisted

Table 1
Schedule and characteristics of each HIT release including phase, description, number of plots, dates, and timing.

	Description	Number of plots	Initiated	Completed	Approximate timing
Phase 1	Quality control	1	24-Oct	29-Oct	6 days
Phase 1b	Quality control re-launch	50	19-Nov	23-Nov	5 days
Phase 2	Control group & crowd	451	18-Dec	21-Dec	4 days
Phase 3	Crowd only	1,498	2-Mar	5-Mar	4 days

of 11 questions each requiring the interpretation of a single dot over imagery. Crowd workers who passed the exam with an 80% were qualified for the HIT. Qualification also included the review of an informed consent document outlining the data that were being collected and any possible risks, and a captcha task to reduce interference in our HIT by bots (Stokel-Walker, 2018; Moss and Litman, 2018; Ahler et al., 2021).

2.3. HIT release and quality control

Crowd HITs were released in three phases. The first and second phases consisted of plots that corresponded to control plots, while the third phase represented the remaining 1496 plots. The first phase was limited to only 50 plots for the purpose of quality control and to ensure that the interface was working properly. These 50 plots were randomly selected from the set of 502 that the control group was also interpreting (see Table 1).

Technical issues with the initial phase 1 request resulted in individual workers interpreting the same plot location many times. As an exploratory quality control measure, these interpretations were plotted, compiled, and mapped to show how many of their own dot interpretations agreed. This information was qualitatively assessed to be satisfactory before the HITs were re-released. Once phase 1 was completed, plot interpretations were mapped out over NAIP and qualitatively spot checked to ensure that workers were interpreting the imagery correctly and that the interpretations were properly aligned with NAIP in case of image projection issues. Interpretations that resulted in 0% TCC were also identified and visually compared to ensure that workers were not submitting a disproportionate number of empty HITs to increase their wage. After phase 2, histograms were produced to further confirm there was not an over-abundance of 0% TCC interpretations compared to the expected distribution. Direct quantitative comparisons were not performed for quality control as one of the objectives of this study was to test the ability of crowd workers to collect consistent interpretations with the control group, and adjusting the interface based on the results of these comparisons could bias that outcome.

2.4. Data analysis

This study generated up to nine crowd interpretations at 1998 plot locations and up to three control interpretations at 477 plot locations across the United States. While the control group had 502 plots available to them to interpret, because of various technical issues we were only able to collect data at 477 of those locations. In this study, no clear benchmark for "truth" exists among these photointerpretations of TCC, and so we rely on analysis of agreement between crowd and control interpreter groups. A reliability score was calculated for each crowd worker using the proportion of plots where that crowd worker's interpretations agreed with each of the available control interpretations within 10% TCC. This proportion was calculated using an area-weighted average for each 1% interval of the standard error layer based on the proportion of an individual crowd worker's interpretations that are within 10% of control interpretations(p_{ai}) (Eq. (3)).

$$p_{ai} = \frac{n_{ai}}{n_i} \tag{3}$$

where n_{ai} is the number of interpretations from worker a in agreement in a 1% interval of the standard error layer i and n_i is the sample size within interval i.

The proportion of plots with agreement within 10% TCC was also calculated between individual control interpreters to better understand their level of agreement. Control interpretations were also compared with one another and to crowd interpretations using an estimate of Pearson's correlation coefficient. Estimated Pearson's correlation coefficients (r_{XY}) (Eq. (5)) between groups textit$X$ and$Y$ were calculated by:

$$r_{XY} = \frac{\sum (w_i * (X_i - \bar{X}) * (Y_i - \bar{Y}))}{\sqrt{\sum (w_i * (X_i - \bar{X})^2) * \sum (w_i * (Y_i - \bar{Y})^2)}} \tag{4}$$

where X_i and Y_i are the values of from groups X and Y within interval i of the standard error layer, $\bar{X}$ and $\bar{Y}$ the estimated means of X and Y across all intervals (Pozzi et al., 2012; Tim, 2021), and w_i is the weight representing the proportional area of interval i (a_i) to the full study area (a), defined by:

$$w_i = \frac{a_i}{a}; \sum w_i = 1 \tag{5}$$

Crowd plot interpretations were compared to control group interpretations using estimated means and estimated Pearson's correlation coefficients (r).

2.5. Reliability model

Crowd workers' interpretations were compared to control group interpretations to obtain ratings of crowd worker reliability. Because there were multiple crowd and control group interpretations, a data frame was created with records for all possible comparisons of crowd worker and control group interpretations at each plot. The percentage of these comparisons that were within 10% TCC was obtained for each crowd interpreter using the estimated mean agreement as described above (Eq. (3)).

Benchmarks for crowd worker reliability were set by the upper and lower bounds of agreement between the three interpreters in the control group. The minimum proportion of all 2011 and 2016 observations that fell within 10% TCC between any two control group interpreters was 0.20. This value was used to define the upper bound of "unreliable" crowd workers, by categorizing any crowd worker whose interpretations agreed within 10% TCC with less than 0.20 of the control interpretations. The maximum proportion of all 2011 and 2016 observations that fell within 10% TCC between any two control group interpreters was 0.50. Likewise, this value was used to define the lower bound of reliable crowd workers, whose interpretations agreed within 10% TCC with more than 0.50 of the control interpretations. Crowd workers who only interpreted HITs in phase 3, which were not interpreted by the control group, and those individuals whose agreement frequency fell between 0.20 and 0.50 were considered to be "Other/Unknown' and were left out of the training sample.

While data were collected for both 2011 and 2016, initial testing on both years of data showed strong correlations between results, so while the reliability analysis presented in this study relates only to 2016 data, similar trends were observed for 2011. The benchmarks described above, however are defined using both 2011 and 2016 data.

This process resulted in a training sample with 95 reliable interpretations at 93 plots from 10 workers and 5168 unreliable interpretations at 1952 plots from 23 workers. While these benchmarks were obtained using data from both 2011 and 2016, the reliability model was only estimated and analyzed for the 2016 data.

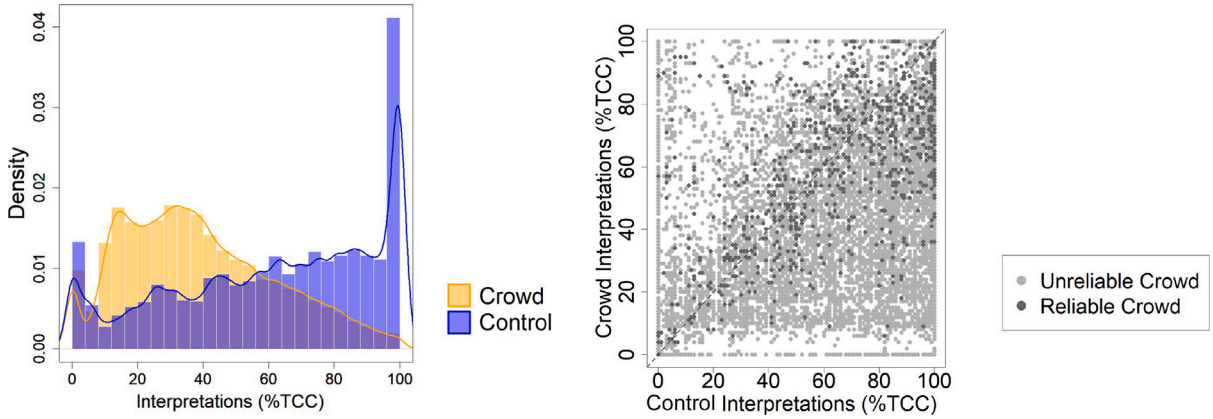


Fig. 3. (Left) Distribution of control and crowd interpretation values for 2016. Data from 2011 show similar trends and distributions. (Right) All combinations of control vs. crowd interpretations from 2016 split into unreliable and reliable classifications based on model predictions.

A subset of variables relating to plot characteristics and interpretation behavior were collected for each interpretation. A training sample was then selected using a Monte Carlo framework and used to estimate a series of 5000 random forest classification models predicting reliability class using R's randomForest package, version 4.6-14 (R. Core Team, 2018; Breiman et al., 2018). There was a disparity in the number of reliable interpretations in the training sample compared with the more prevalent unreliable class. Because classification algorithms are typically designed to minimize the output model's overall error, a large disparity in training class size could result in poor quality classifications for the smaller class since the small number of training sample plots would have a relatively small impact on the models error. To mitigate this issue in our reliability model, bootstrapped sampling was used to randomly choose a balanced sample from the training data for each of the 5000 models following Chen et al. (2004). The reliable interpretations were randomly selected with replacement to identify an in-bag sample and an out-of-bag testing sample. The resulting in-bag dataset was then sampled with replacement again to amplify the number of interpretations in the sample until the reliable class accounted for 50% of interpretations used to train our model. The unreliable data were also sampled using this method to select an unreliable sample also accounting for 50% of interpretations used to train our model. The resulting training dataset had two classes with 2637 interpretations in each. Unreliable interpretations that were not selected for the training data set were held out for the out-of-bag testing dataset. The sample interpretations from reliable and "unreliable" classes were combined to create a training data set and a testing dataset, and a random forest classification model was estimated to predict interpreter reliability. This process was repeated 5000 times, creating training and testing data for each model.

The final reliability classification separated all interpretations into reliable and unreliable classes using these scores. If more than 50% of models classified an interpretation as reliable, the final classification was reliable. If less than or equal to 50% of the models classified an interpretations reliable, the final classification was unreliable.

Variable importance metrics and in-bag and out-of-bag predictions were obtained for each of the 5000 reliability models based on the rank-score method described in Schroeder et al. (2017). Out-of-bag predictions were used to estimate overall model accuracy for each model iteration and predictions on all data (in-bag and out-of-bag) were used to calculate a percent reliability metric for each interpretation. This metric reflects the percent of models which predicted that interpretation to be reliable. To test variable importance in our reliability model, the mean increase in error resulting from each predictor's removal from the model was collected and ranked for each iteration. The top 10 most important variables were given a score between 10 (for the most important variable) and 1 (for the 10th most important variable) and those scores were added by variable across iterations to get an overall variable importance score (Schroeder et al., 2017).

Different subsets of crowd interpretations (all crowd, 2011, 2016, reliable, and unreliable) were compared to control group interpretations to test if they came from statistically different distributions. These statistics were approximated using a combination of all available interpretations at each plot location. To ensure representation of all interpretations we randomly sampled a single interpretation from (up to 9) available crowd interpretations and a single interpretations from the (up to 3) available control group interpretations at each of the 477 plots where both crowd and control interpreters worked and repeated this process 5000 times. For each of the 5000 resulting data sets, subsets of data were obtained by querying by the interpretation year and, for 2016 data, by interpretation reliability class. Control group interpretations were compared to crowd interpretations for each subgroup using estimated Pearson's correlation coefficient (r).

3. Results

3.1. All interpretations

Crowd interpreters returned 17,690 assignments during this study at 1998 plot locations, each representing a TCC interpretation for 2011 and 2016. Interpreters from the control group returned a total of 1078 assignments at 502 locations. The 170 crowd workers who participated in this study completed between 1 and 1927 assignments individually, with the five most prolific interpreters producing between 1208 and 1927 interpretations and the ten most prolific producing between 830 and 1927. The majority of workers submitted at least 6 assignments, but many workers submitted fewer, and 49 workers submitted only 1 assignment.

Mean TCC values from crowd interpretations are substantially lower than those from control interpreters (Table 2). When phase 3 is accounted for, the estimated mean of the crowd interpretations is even lower for both years, however there were no control interpretations at phase 3 plots for comparison. The disparity in crowd and control interpretations is apparent in the histograms from 2016 which show that crowd interpreters interpreted lower TCC values more often than the control group. By contrast, the histograms also show that control group interpreters observed very high TCC values more often than crowd workers (Fig. 3). Data from 2011 show similar trends and distributions.

Disagreement was measured using the standard deviation of interpretations for each plot by group. Disagreement between crowd workers is higher overall than it is for the control group in 2011 and 2016 for phase 1 and 2 data only. This trend is maintained when all three phases of crowd data are accounted for (Table 2). Histograms also

Table 2
Summary values for interpretations as collected. Disagreement refers to the standard deviation of all interpretations of each plot by worker type.

Worker type	Year	Phases	Number of workers	Number of plots	Number of interpretations	Estimated mean % TCC	Estimated mean disagreement (% TCC)
Control Group	2011	1 & 2	3	502	1,078	63.6	16.8
Control Group	2016	1 & 2	3	502	1,078	63.6	16.4
Crowd	2011	1 & 2	108	502	4,432	41.7	23.3
Crowd	2016	1 & 2	108	502	4,432	43.6	21.9
Crowd	2011	1, 2, & 3	170	1,998	17,690	38.8	21.2
Crowd	2016	1, 2, & 3	170	1,998	17,690	39.6	20.5

Table 3
Inter-interpreter comparison for the control group showing number of plots interpreted between the interpreters in question, proportion of interpretations within 10% TCC, and estimated Pearson's correlation coefficient (r).

Comparison	Number of interpretations	Est. Mean Proportion within 10% (2011 and 2016)	Est. r (2011)	Est. r (2016)
Control Worker 1 v Control Worker 2	388	0.53	0.64	0.74
Control Worker 2 v Control Worker 3	484	0.32	0.77	0.82
Control Worker 3 v Control Worker 1	708	0.20	0.73	0.67
All Control	1,580	0.31		

show that disagreement values tend to be higher for crowd interpreters than for the control group in 2016 (similar results were observed for 2011), but control interpreters' disagreement levels have more plots with high disagreement (Fig. 3).

Diagnostic plots for individual crowd workers and maps of plot interpretations show several patterns in crowd interpreter behavior. For example, a scatter plot between an individual crowd worker's interpretations (later deemed unreliable) and control interpretations show a concentration of crowd interpretation values between 5 and 25% TCC. Comparing interpretation values to the number of plots interpreted show a sharp decline in TCC values as the worker interpreted more plots, with recovery at the beginning of each HIT release phase which were separated by up to 72 days. Additionally, interpretation times decline as the worker interpreted more plots, also recovering at the beginning of each interpretation phase (Fig. 7).

Maps of selected crowd interpretations show the following cases:

- Crowd workers clicked a series of disconnected dots throughout the plot resulting in an overestimate of the TCC value compared to the control interpretation (Top-left, Fig. 4)
- A plot with heavy texture and shadow where a crowd worker underestimated TCC compared to control interpretations (Top-right, Fig. 4)
- A plot showing features which a crowd worker interpreted as trees, and which the control group member interpreted as non-tree features resulting in a higher crowd interpretation value compared to the control (Bottom-left, Fig. 4)
- A plot with heavy shadow and nearby trees where a crowd worker overestimated TCC compared to the control (Bottom-right, Fig. 4)

Finally, scatter plots pairing all control interpretations against all crowd interpretations show that crowd workers routinely underestimate TCC compared to the control, and that there is little correlation between the groups' interpretations (Fig. 3).

3.2. Control group interpreters and agreement

Control interpretations were compared by interpreter to determine if their observations were similar. For all control interpreters, 31% of interpretation were within 10% TCC of the other control interpreters, with the lowest agreement frequency between control interpreters 1 and 3 and the highest frequency of agreement between control interpreters 1 and 2. These agreement proportions account for interpretations from both 2011 and 2016 (Table 3).

Table 4
Estimated Pearson's correlation coefficients (r) between all crowd and control group interpretations for 2011 and 2016 and all dates, and between reliable crowd and control group interpretations and unreliable crowd and control group interpretations for 2016. Records with NA values were removed from calculations. All comparisons are calculated using bootstrapped interpretations (one at each plot) so *n Interpretations* is also equivalent to the number of plots represented.

Year	Number of interpretations	Estimated r
2011	477	0.19
2016	477	0.23
All	954	0.21
Reliable (2016)	41.2	0.53
Unreliable (2016)	434.9	0.21

Finally, estimated Pearson's correlation coefficients (r) show correlations between control interpreters (comparing data from individual years) that range between 0.64 and 0.82. Scatterplots of TCC interpretations between control group interpreters show some correlation. Plots show little difference in agreement between interpretations of 2011 imagery and 2016 imagery (Fig. 5).

3.3. Reliability

3.3.1. Reliability model

The Monte Carlo modeling framework produced 5000 models for interpretation reliability, each with their own class predictions (reliable or unreliable) for all plots, out-of-bag error estimates, and variable importance rankings. The out-of-bag accuracy of all classification models are normally distributed with a mean of 99.5%, a minimum model accuracy of 99.0%, and a maximum model accuracy of 99.8%. The aggregate confusion matrix for these models shows that, on average, 73% of interpretations that were observed to be reliable were also predicted to be reliable, and >99% of plots that were observed to be unreliable were also predicted to be unreliable. The final class of reliable interpretations has a higher level of agreement with control workers' interpretations than those interpretations classified as unreliable. Where the complete set of crowd data generally shows crowd workers under-estimating TCC values compared to control interpreters,

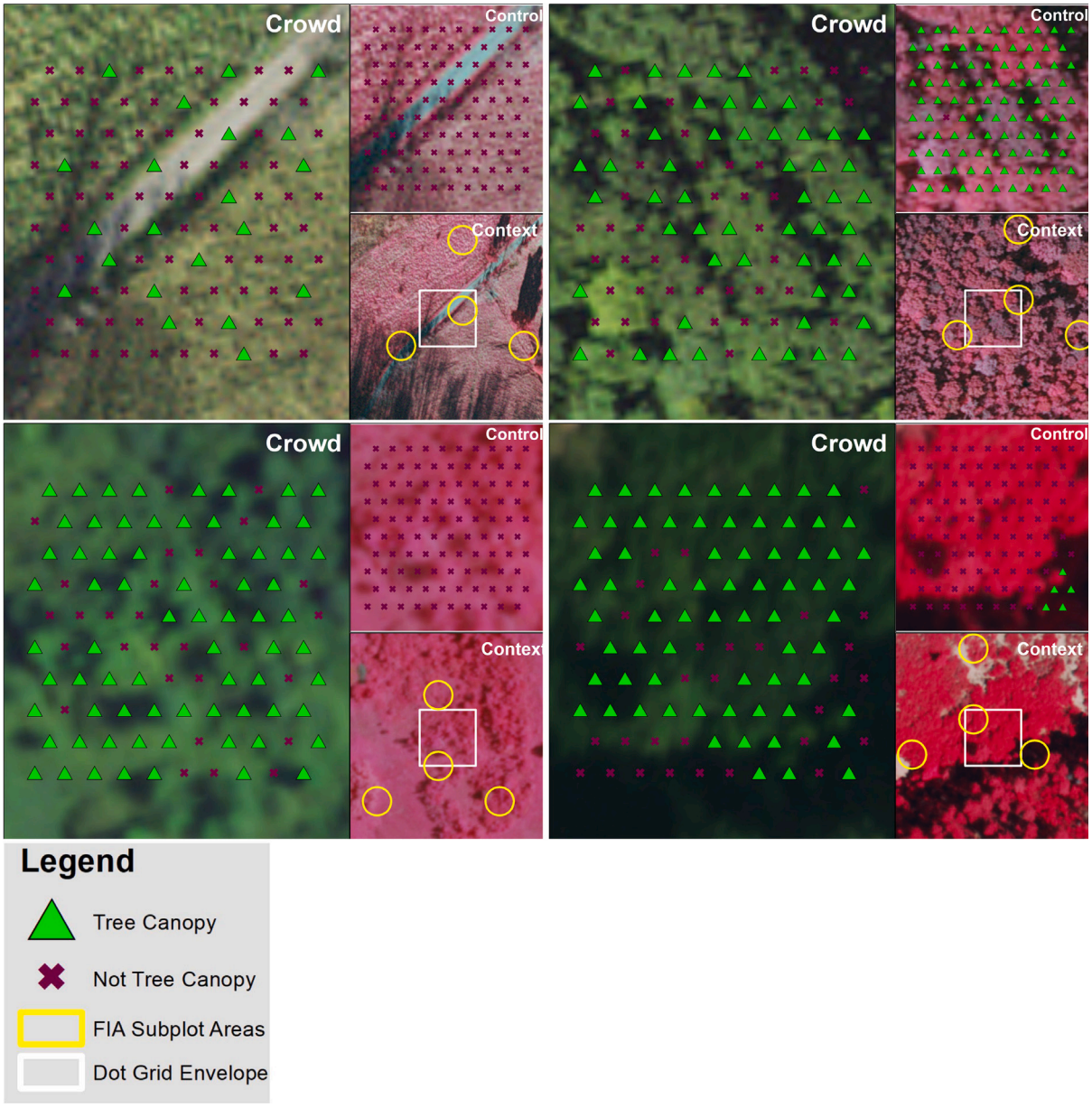


Fig. 4. Each image shows a panel with a crowd interpretation (and the underlying NAIP imagery in the display used during the majority of that interpretation) at an approximate scale of 1:700, a control interpretation (and the underlying NAIP imagery in the display used during the majority of that interpretation) at an approximate scale of 1:1000, and a context image, zoomed to an approximate scale of 1:3250. Public FIA plot footprints are also shown in the context panel to demonstrate where the plot coordinates fell relative to the area covered by the interpretation dot grid. Only publicly available plot locations, which are approximations of actual FIA plot locations, were used in this image and in this study. Base layers from ESRI and the USDA Farm Service Agency.

a larger proportion of the interpretations deemed reliable are close to the 1:1 line or above the 1:1 line (Fig. 3).

Selected interpretations were evaluated for correlation using an estimate of the Pearson's correlation coefficient (r). This metric were calculated using all available plots for the 2011, 2016, and full set of data, to identify different patterns across the dates interpreted. There were 477 plots that both crowd and control group workers interpreted for 2011 and 2016 (adding to 954 single-date interpretations). Crowd and control interpretations were randomly selected from the pool of available interpretations for each plot location using a resample-aggregation approach. This meant that one of the up to three available control interpretations was selected and one of the up to nine available crowd interpretations were selected for each resample draw with 5000 resamples total. The r values were calculated for each resample and then an average r was used as for the estimate. This process was repeated with reliable and unreliable interpretations, however the number of plots that were identified as reliable and unreliable varied based on the interpretation selected by the resample draw. Therefore, the number of interpretations reported is an average of the number of reliable and unreliable interpretations used to calculate r for each resample. Reliable interpretations accounted for a smaller proportion of the interpretations in all of the 5000 samples, accounting for about 8.6% of each sample plot on average.

When comparing all crowd and control interpreters, 2016 data had a slightly higher correlation compared to 2011, but data from individual years and both years combined showed a low correlation between crowd and control interpreters. For unreliable 2016 interpretations, many of the statistics follow the same patterns as the full set

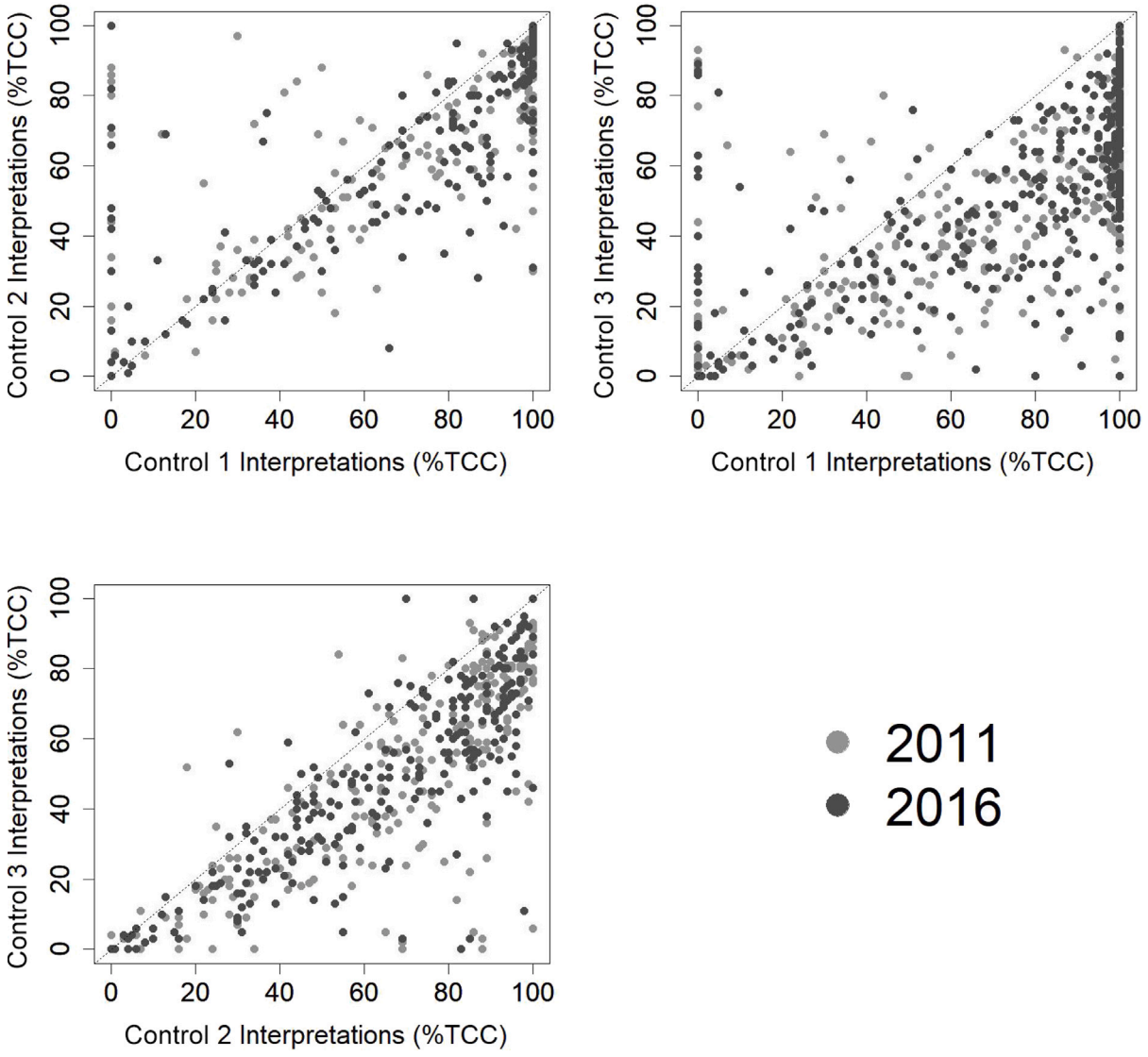


Fig. 5. Scatter plot between individual control group members for 2011 and 2016.

of 2016 interpretations, with even lower correlations between crowd and control group interpretations. However, reliable 2016 data show a stronger (though unremarkable) correlation between crowd and control group interpretations, indicating that reliable crowd interpretations and control group interpretations are not significantly different (see Table 4).

3.4. Important variables

Variable importance scores, demonstrate that *the number of plots interpreted* is the most significant variable in determining the reliability of a plot interpretation (Schroeder et al., 2017). The second most important variable overall was *the % of dots interpreted with a false color display*, which was ranked most important for 90 models, second most important in 4599 models, and lower rankings for 311 models, with an overall score of 44,733. The *% of dots interpreted with a natural color display* and *the interpreted TCC value* are closely ranked, with overall scores of 36,531 and 35,915, respectively. Categorically, two out of four of the interpretation-scale variables (*Number of plots interpreted* and *Number of HIT's available*) have importance scores ranked in the top five, as well as two out of four of the dot-level interpretation variables (*% of dots interpreted using false color display* and *% of dots interpreted using natural color display*). One of the TCC variables was also ranked in the top five (*interpreted TCC*). Neither of the two ecological plot variables have importance scores ranked in the top five, however *level-2 ecoregion* was the lowest ranked variable in the top half (Fig. 6).

3.4.1. Number of plots interpreted

Number of plots interpreted is ranked the most important variable overall in determining crowd interpretation reliability in photointerpretation of TCC compared to the control group. It was ranked most important variable for 4910 out of 5000 models, and second for the remaining 90 models. The number of plots interpreted is higher among all crowd interpretations and unreliable crowd interpretations, and even lower among reliable crowd interpretations than among control group interpreters who were only expected to complete 502 of the 1998 total plots theoretically available to crowd workers. This indicates that reliable interpretations were more likely to occur early on in a worker's interpretation sequence (Table 5).

When the number of plots interpreted (n) is plotted against the mean absolute difference between the n'th crowd interpretation and all available control interpretations, the trend line has a positive slope of 0.02. Based on this linear model, the predicted error of the first interpretation all workers interpreted is 27.4%, while the predicted error of the 400th plot interpreted is 35.4%, and the predicted error of the 894th plot (the maximum value of number of plots interpreted

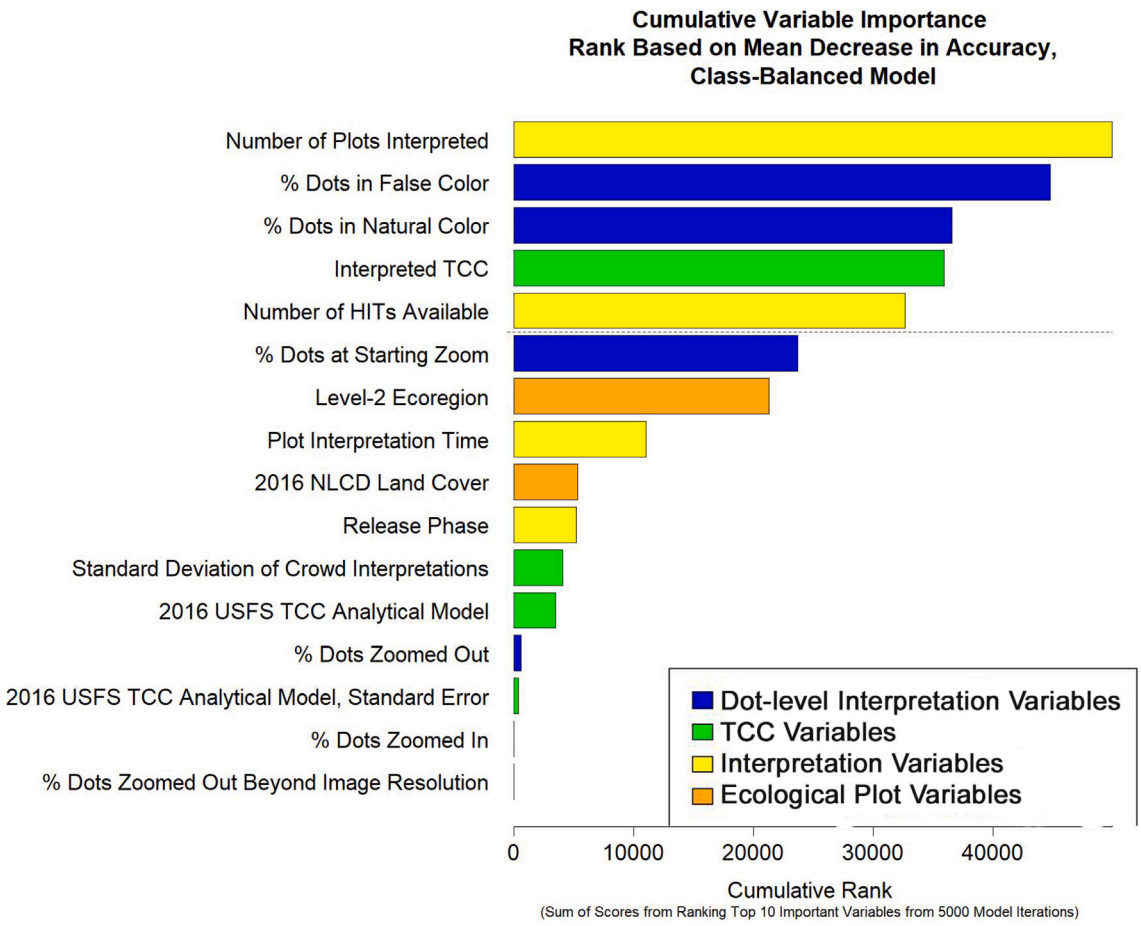


Fig. 6. Variable importance scores for reliability model, calculated using a rank-score method which assigns points to the top ten most important variables in each of the 5000 models and adds them together (2016 only). Dashed line indicates the top five most important variables across all models. These scores were derived from the mean decrease in accuracy metric supplied by the randomForest function in R.

Table 5
Estimated mean value for the top five most important variables by interpreter class for 2016 only.

variables	All crowd interpretations	Unreliable crowd interpretations	Reliable crowd interpretations	Control group interpretations
Number of Plots Interpreted	1095.3	1124.8	30.8	378.9
% Dots in False Color	5.8	4.7	46.5	62.0
% Dots in Natural Color	34.8	35.3	17.0	1.4
Interpreted TCC	39.6	39.0	64.3	62.9
Number of HITs Available	707.7	715.6	398.9	186.9

by a crowd worker across plots that were also interpreted by the control group) interpreted is 45.3% (Fig. 7).

3.4.2. % of Dots in False Color display and % of Dots in Natural Color display

The variables "% of Dots in False Color" and "% of Dots in Natural Color" refer to the percent of all 100 dots in each single-date plot interpretation that were clicked in each display. For control interpretations, the majority of plots were interpreted using the false color display, while crowd workers usually interpreted images in a natural color display, which was the default. Like control interpretations, reliable crowd interpretations utilized the false color display more than unreliable crowd interpretations, but reliable crowd workers were more likely to use a combination of displays, changing display in the course of 31.1% of their interpretations compared to all other groups who changed display settings in the middle of an interpretation less than 10% of the time (Table 5).

3.4.3. Interpreted TCC

Interpreted TCC represents the resulting value of the interpreted plots. The plot interpretation value for all crowd workers was significantly lower than the control group's with an average of 39.6% TCC compared to 62.9% TCC for control interpreters at the same plots. Reliable crowd interpretations have a higher TCC value on average than control interpretations, with a mean of 64.3% TCC, and interpreted TCC values from the control group have a wider distribution than reliable interpretations. Unreliable crowd interpretations have the lowest value, with an average of 39.0% TCC. (Table 5).

3.4.4. Number of HITs available

The number of HITs available represents the actual number of available HITs for each worker at the time of interpretation. Unreliable crowd interpretations were performed when there was a larger number of HITs available than reliable interpretations. For unreliable

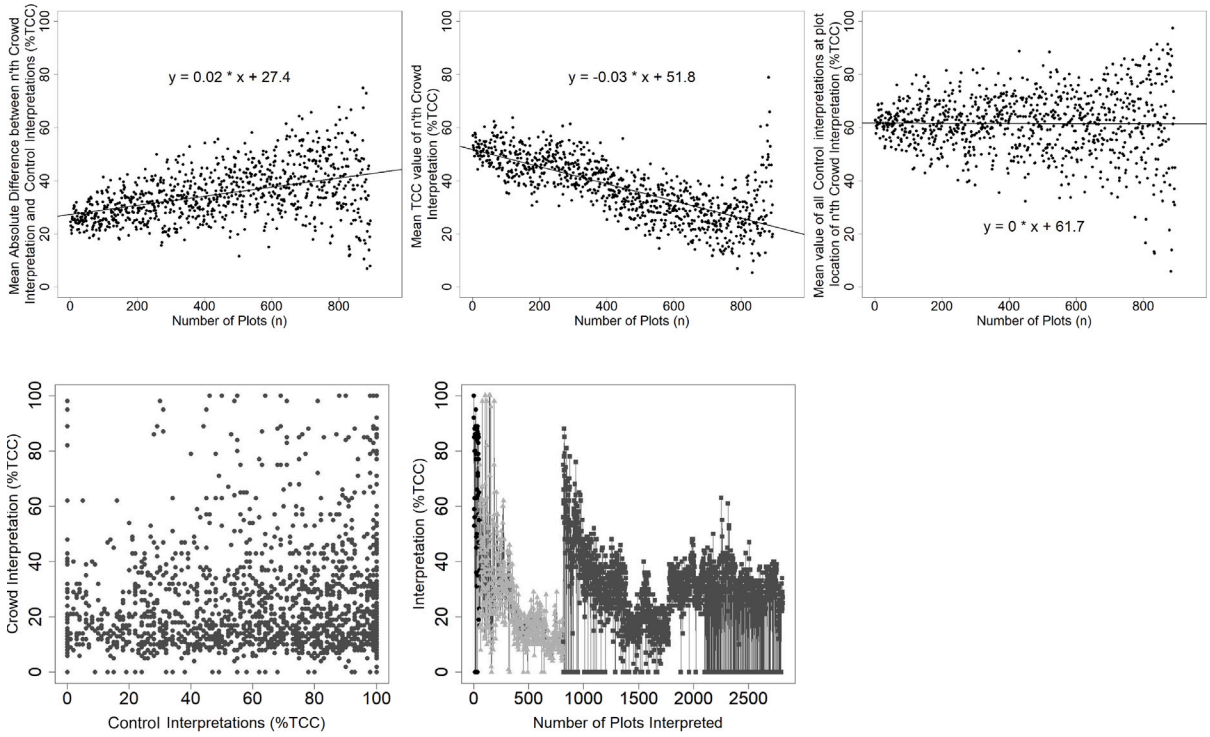


Fig. 7. Scatter plots demonstrating an increase in error and decrease in TCC value with additional plots as interpreted by crowd workers. (top-left) Absolute difference of the n'th crowd interpretation, averaged across all workers who interpreted n plots vs. the number of plots interpreted (n). (top-center) Mean TCC value of the n'th crowd interpretation, averaged across all workers who interpreted n plots vs. the number of plots interpreted (n). Please note that HITs were assigned randomly and not in sequence, therefore the n'th plot for one worker is likely to fall in a different location than the n'th plot for another worker. Therefore, these data represent the average value of crowd interpretations across multiple locations, but at the same time in their sequence of interpretations. (top-right) Mean TCC value of the control group's interpretations at the plots locations corresponding to the n'th crowd interpretation, averaged across all available control group interpretations at plots that were interpreted as any crowd worker's n'th interpretation vs. the number of plots interpreted (n). Dots displayed in all three plots represent interpretations of 2016 imagery at plots that were interpreted by the control group and also by crowd workers. (bottom-left) Plots from an individual crowd worker's interpretations showing the control group's interpretations vs. the worker's interpretations, and (bottom-center) the number of the worker's plots interpreted (n) vs. the interpretation's value.

and all crowd interpretations, box plots show a wider, more balanced distribution of values than they do for reliable crowd interpretations, which has a skewed distribution. The majority of reliable crowd interpretations were completed while there were fewer than the mean number of HITs available for that group. Control interpretations show a narrow, balanced distribution, around a fairly low mean number of HITs available, reflecting the smaller number of control group interpreters who accepted fewer assignments overall with no competition for assignments (Table 5).

4. Discussion

4.1. Consistency in photointerpretation was difficult to achieve, for both the control group and crowd workers

The inter-comparison of control interpreter data shows that each worker's interpretations follow a different distribution and the control group members with the highest level of agreement interpreted only 53% of plots to be within 10% TCC of one another. While the maximum estimated correlation is 0.82, the minimum correlation is only 0.64, suggesting that the control interpreters had difficulty with many plot interpretations and lacked consistency in some instances (Table 3). Some reasons for this inconsistency can be seen in the scatter plots comparing their interpretations. These plots show plots where one member of the control group consistently interpreted very low or 0% TCC and another interpreted higher % TCC, vice versa pointing to a lack of consensus in what constitutes trees and the interpretation sub-canopy tree cover or ground within canopy gaps.(Fig. 5).

Several methodological decisions may improve consistency among interpretations in future studies. First, our control interpreters were not trained using a consistent protocol as seen in Jackson et al. (2012), where between 63%–90% of TCC interpretations from expert interpreters were within 10% agreement with other experts (Jackson et al., 2012) and in Riemann et al. (2016) where re-measured TCC photointerpretations showed a high level of agreement between interpreters (Riemann et al., 2016). Each of our control interpreters had a background in remote sensing and were comfortable interpreting landscape features from high resolution remote sensing imagery across the United States. They were also given the same training protocol as the crowd interpreters, which identified various features that were known to be difficult to interpret, such as shrubs and shadows, and provided guidance on their interpretation. In order to allow for intercomparison between the interpretation behaviors of the crowd and control groups, we did not implement a more in-depth training protocol that resembled the comprehensive training used in many expert-driven photointerpretation studies. A consistent and rigorous training protocol for control and crowd interpreters could improve results going forward. Our control group had not visited the locations they were interpreting, and they interpreted plots across a broader region than the studies mentioned above. Our workers also did not interpret plots that were modeled to have 0% TCC based on the 2016 USFS Analytical Tree Canopy Cover product, which may have improved their level of agreement. Finally, TCC is scale-dependent (Coulston et al., 2012, 2010). Though the definitions of tree and the photointerpretation methods that produced this dataset are similar to those for the original training data for the Forest Service Tree Canopy Cover model (United States Department of Agriculture, Forest Service, 2019; Ruefnacht et al., 2022), the interpretations are performed using a 100-dot grid that covered a 30 × 30 meter plot, as compared to the model's training data which used a 109-dot grid over a 144 ft radius plot (6052 sq. m) (Ruefnacht

et al., 2022), while the resolution of the base imagery (NAIP) remained the same. This change could greatly impact all interpreters' ability to resolve features across the landscape, and introduce measurement error and inconsistency in their interpretations.

This study found that crowd interpretations were generally unreliable with respect to control group interpretations. Visual inspection of various interpretations revealed that crowd workers had some difficulty distinguishing trees from vegetated background, and identifying canopy gaps from shadows in dense canopy areas. Additionally, crowd workers appeared to have difficulty resolving separated tree crowns when they were situated among heavily vegetated shrublands. They also appeared to miss contextual cues which could be determined by zooming out from a plot to see other trees for comparison. This resulted in erroneously high TCC values interpreted in croplands and shrublands. By contrast, crowd interpreters were successful when identifying trees with separated canopies on bare, un-vegetated soil.

It also appears that some workers clicked randomly and erratically, not following any pattern related to the vegetation on the landscape. For some individuals, scatter plots of their interpretations compared to control interpretations showed many of their interpretations fell between 10 and 30% TCC with no association with control interpretations. Some scatter plots for individual workers show that 0% TCC interpretations were prevalent even when control group members interpreted high values of TCC, possibly an indication that they occasionally submitted empty interpretations to increase their payout. Quality of crowd work has declined on Mechanical Turk in the last few years (Stokel-Walker, 2018; Moss and Litman, 2018). Starting in 2018, research states that "troll" workers who provide "insincere" responses have become more prevalent on Mechanical Turk. Ahler, Roush, and Sood (2021) cited an increase in "untrustworthy" responses from 6%–13% in 2015 to 25% in 2018. These combined interpretations characterized by misrepresented location and "insincere" responses (Ahler et al., 2021). Another study by Chmielewski and Kucker (2019) performed a 4-wave survey data collection over the summer of 2018 and used response-validity indicators to ensure the quality of responses. They showed that in waves 1 and 2, only 10.4% and 13.8% of respondents were flagged. In wave 3, a notable decrease in crowd performance was identified with 62.0% of respondents flagged. This was reduced to 38.2% in wave 4 (Chmielewski and Kucker, 2019).

We also looked at the standard deviation in crowd and control interpretations at each plot to understand the level of disagreement. Estimated mean standard deviation values is 16.8% for 2011 imagery and 16.4% for 2016 imagery for control interpreters. The standard deviation of crowd workers is higher, as expected, ranging between 20.5–23.3% for different subsets of data and years. Between groups of workers there is very little agreement. The estimated correlation between crowd and control interpretations is 0.21 overall. By comparison, reliable crowd interpretations have an estimated correlation of 0.53 with control interpretations (Table 2).

4.2. Crowd workers' used the interface differently leading to different interpretations

Display characteristics proved to be a significant determinant of crowd interpretation reliability with respect to control interpretations. The second most important variable in our reliability model was the percentage of dots that were clicked using the false color composite and the third most important variable in our reliability model was the percentage of dots that were clicked using the natural color composite.

Since reliable interpretations and control interpretations were typically performed using the false color display, while crowd interpretations were perfomed using natural color, it is possible that the visible differences between the natural color imagery and the false color imagery led to the different interpretations of TCC at the same plots. In general, crowd workers underestimated areas where control interpreters indicated there should be high TCC. There are examples of this sort of behavior in the literature, where inexperienced workers view shadows on other canopies as gaps to the ground (Avery 1969, page 30) (Avery, 1969). The false color display, which uses the near infrared band to distinguish strong vegetation signatures, even in some cases when they are obstructed by shadow could have resulted in higher TCC values at plots with a lot of gaps and shadows. We see examples of this in this study as well (See Fig. 4).

While the false color display was provided to both crowd and control workers, and information regarding the usefulness of false color imagery for TCC interpretation was described in training, crowd workers rarely used the false color composite imagery. The natural color display was the default setting when a HIT was initiated, as we expected most people would be better acquainted with this view. In order to change the display, the map would need to be reloaded which would add time to the interpretation, so perhaps those workers who were motivated to work quickly chose to use the default display setting. Additionally, Campbell and Wynne (2011) suggest that the use of non-visible bands can be a source of confusion to non-expert interpreters (Campbell and Wynne, 2011). Thus it is possible that workers who provided the use of false color imagery is actually an indicator of an interpreter's experience with photointerpretation. However, since false color and natural color imagery highlight different landscape characteristics, more consistent interpretations may have been obtained if all interpreters used the same display.

4.3. The reliability of crowd interpretations was impeded by extrinsic financial motivators

Fig. 3 demonstrates that crowd interpretations were systematically lower than control interpretations. Many of the HITs requested received one or more very low value interpretations as compared to the other assignments submitted by crowd workers and compared to the interpretations submitted by control group members, as evidence by Fig. 7. Reliable crowd interpretations tended to have higher TCC values, with an average value that exceeded the control group's average 5.

The structure of our HIT request and the incentive structure of crowdsourcing in general led to a dataset that was biased toward lower crowd TCC values. Our HIT only required workers to click on dots they identified as landing on tree canopies, and workers were paid for each assignment submitted without regard for the time invested. Crowdsourcing is commonly associated with task-defined payments rather than time-defined payments, so it follows that spending less time on task by submitting systematically lower interpretations would yield a higher wage.

The visible patterns in Fig. 7 may indicate different methods for reducing average time-on-task such as submitting a number of fully-interpreted assignments followed by a few empty uninterpreted assignments or clicking only a portion of the dots that landed on trees in each assignment. Patterns characterized by progressively lower TCC interpretations or intermittent 0% TCC observations at a higher frequency than would be expected based on the distribution of the data were observed in several individual crowd workers' interpretations, though they could also be explained by technical issues preventing the interpreter from visualizing the assignment or inexperience interpreting this type of imagery. These patterns may also indicate broader differences in the interpretation approach taken between crowd workers and the control group or possibly the effects of fatigue on crowd workers.

Assignments on Mechanical Turk are a no-sum game. They are usually distributed on a first-come, first-serve basis and the number of assignments available for a given HIT are finite. Thus working too slowly could reduce the number of HITs a worker has access to and, in turn, their potential earnings. The competition for available assignments incentivizes workers to work for many hours continuing through fatigue to ensure they had access to as many HITs as possible.

The literature related to crowdsourcing and photointerpretation suggests that fatigue is a strong indicator of worker performance (Van

Coillie et al. (2014), See et al. (2013), Szalma et al. (2006), Yu et al. (2015) and Parasuraman (1986) as referenced by Van Coillie et al. (2014)). In our study, the most important variable for determining plot reliability was the number of plots interpreted, which was our indicator of worker fatigue. Many workers show a sharp decline in TCC value and interpretation time with interpretation sequence (Figs. 7 and 4).

Fig. 7 demonstrates that, as a worker interprets more plots and stays on task longer, the more the value of their interpretations tends to diverge from control interpretations. The spread of the scatter plot widens as the number of plots interpreted increases likely due to the fact that substantially fewer workers interpreted a larger number of plots.

The fifth-most important variable in our reliability model was the number of HITs available. This variable was included in our assessment to get an understanding of how the potential earnings that workers had access to at a given point in time were affecting their attention to their interpretations. For workers on Mechanical Turk, the number of HITs available could be multiplied by the wage to get an estimate of potential earnings if a worker completed their assignments quickly enough to access all of the HITs in the request. Our study showed that reliable interpretations were often completed toward the end of a request when there were relatively few HITs available, and less competition to complete those HITs. High potential earnings for a HIT may create a motivation to move through HITs faster in order to get to the potential pay out before another worker.

Some of the time-money motivations mentioned here are further discussed in the literature related to Mechanical Turk. Rogstadius et al. (2011) demonstrates that intrinsic motivators can improve the quality of a crowdsourcing task and that extrinsic motivators (like financial incentives) could actually create a "crowding out" effect where workers have less interest in tasks they are paid to perform (Rogstadius et al., 2011). Various methods have also been explored to improve quality and reduce the influence of workers who prioritize speed over quality (Eickhoff and de Vries, 2013; Eickhoff et al., 2012).

4.4. Recommended improvements

There are possible ways to improve this study for future work. The biggest improvement would be systematic quality control throughout the study. We performed thorough qualitative analysis on the first phase of interpretations, but the observed fatigue effects and time-money motivation indicate that this study would have benefited from additional assessments at later stages of data collection. Requiring workers to review and certify their work using a visual inspection could also be helpful. Yu et al. (2014), suggest a combination of "social" and "learning strategies" to increase intrinsic motivation to incentivize higher quality work and improve participant engagement (Yu et al., 2014). This means establishing a clear training goal or skill-development outcome for crowd participants to work toward and also enabling a collaborative or social aspect to participation, such as a chat window or a way of comparing interpretations with other workers. Their findings show that learning strategies improved data quality while social strategies improved participant engagement.

Requiring workers to click every dot in a plot would encourage a more detailed evaluation. This recommendation is an example of a "defensive design" element, ensuring that "cheating" at the task is not easier than doing it well (Allahbakhsh et al., 2013). This could also ensure that workers spent more time on task, reducing the incentive to speed through tasks prioritizing quantity over quality of HITs completed. The increased attention required per assignment could also reduce the likelihood of interpretations that do not follow the prompt, such as those where dots are randomly clicked without following the pattern in the image or empty interpretations submitted for images that are densely forested.

Payment should reflect time on task and a study should be conducted to assess the required time using a group of non-expert interpreters. We underestimated the time it would take an untrained worker to complete an interpretation in our study significantly, predicting that most interpretations could be completed in under 2 min. This time was estimated using a few complex examples from our control group, with the assumption that interpretation time would be significantly reduced in areas with sparse vegetation. Feedback from workers as well as the interpretation times reported in this study communicated that this payment was too low for a task of this complexity. While other studies (Yu et al., 2015; Mason and Watts, 2010; Buhrmester et al., 2011) suggest that wage does not have a strong influence on interpretation quality, efforts should be made to estimate a fair wage based on time required to complete reliable crowd interpretations if similar studies are to be attempted in the future. Mason and Watts (2010) study the effect of payment on crowd performance. They observed an "anchoring" effect where workers who were paid more perceived the value of their work to be greater and were therefore not more motivated than workers who were paid less for the same task. They showed that a quota system resulted in better work for less pay, which would be the best option if it could be operationalized (Mason and Watts, 2010). Additionally, based on workers' comments, with a higher wage we might expect to see better retention of workers who were willing and motivated to spend time and perform high quality interpretations with a higher wage. In future endeavors it would be wise to perform a separate study using untrained workers to estimate the average time required for interpretation and in turn, payment.

Limiting the number of HITs a worker could complete before a new qualification or by releasing HITs in smaller groups could also overcome the motivation to speed through tasks since the number of HITs available would be smaller. This could also enable quality control on subsets of the data without long delays, reducing repercussions to the requester from negative online reviews about payment time. A smaller number of HITs requested per release would also result in more manageable file sizes for compiling and downloading data for analysis.

This study showed that most work was completed by crowd workers using the interface defaults. We recommend that a decision be made on best practices for interpretation, including both interpretation scale and display. These settings should be used in the interface as the default settings. This would create greater consistency between the conditions affecting interpretations across group, and could reduce the difference in interpretation values caused by inconsistent display settings .

Additionally, adding elements like further contextual information with LiDAR (Gatziolis, 2012; Ma et al., 2017; Smith et al., 2009) or intrinsic motivators like certifications (Allahbakhsh et al., 2013; Rotman et al., 2012; Goodchild, 2007) may be helpful in improving the reliability of interpretations.

Not all of the interpretations in this study were unreliable. It is clear that many workers on Mechanical Turk participate with the intention of providing good work and getting paid for it. Fig. 3 and Table 5 show clear separation between reliable and unreliable interpretations. Almost all of the workers provided some reliable interpretations, however, there is still a need for mechanisms to increase interpretation quality if crowdsourcing platforms are ever to be used operationally.

5. Conclusion

This study was aimed at identifying the factors that could predict whether an interpretation could be classified as reliable or unreliable and how those variables were related to reliability in order to make recommendations for future studies, both with regards to interface design and quality control.

We conclude that, while crowd workers are capable of measuring TCC using photointerpretation, agreement between interpreters was difficult to achieve, even among our control group. Factors that affect the agreement between crowd and control interpretations include financial motivations, fatigue, and differences in the utilization of the user interface. In particular, our data show trends that indicate a priority for timely interpretations over high quality interpretations,

which is supported by the incentive structure of Mechanical Turk. The reliability of crowd interpretations also appears to be affected by the potential earnings a worker perceives, with reliable interpretations occurring more often when the number of HITs available (and therefore the potential payout) was lower. Fatigue effects were apparent as consistency between crowd and control interpretations declined with the number of plots interpreted for crowd workers. Finally, crowd workers prioritized the default settings of our interface. Incidentally, these settings included a natural color composite which may be more intuitive for an untrained crowd worker to interpret, but would likely have less available information compared to the false color composite, which was less commonly used. Differences in interpretations arise from both the display settings used and each worker's level of experience interpreting color-infrared imagery and resolving differences between background and foreground in highly vegetated landscapes. These factors may have contributed to differences in quality and consistency among crowd workers compared to the control group.

The lack of consistency observed in this study demonstrates a need for higher quality control standards and more thorough training, both for photointerpreters with a known level of experience, and for those engaging in these activities through a crowdsourcing environment.

We have several recommendations in the event that this work were to be operationalized, utilizing crowd interpretations to obtain measures of TCC from photointerpretation. First, increased quality control measures could help identify and reduce interpretations that fail to follow the prompt, such as those where dots are randomly clicked without following the pattern in the image or empty interpretations submitted for images that have greater than 0% TCC. We suggest that workers be required to evaluate and click every dot in the grid rather than only clicking dots over tree canopies, reducing the likelihood of incomplete interpretations. Feedback from our crowdworkers also highlighted the need for a thorough analysis of required interpretation time for untrained workers to ensure payment is fair and reasonable. Limiting the number of HITs per worker and reducing the number of HITs in each request could mitigate some of the fatigue effects we observed and reduce the perception of competition and the potential earnings from a HIT. To make interpretations more consistent, the user interface should be set up so that the ideal interpretation environment is the default setting since most crowd workers did not change these settings. Additionally, more contextual information about each plot could be provided (in the form of LiDAR data or land cover classifications) to improve a user's understanding of the landscape they are interpreting. Finally, incentives to reward high quality interpretation or quality contingent pricing (Wang et al., 2013) could yield better results.

While our models show promise in being able to distinguish between reliable and unreliable interpretations, the insights drawn from this study are aimed at improving our interface to yield better quality observations in future studies. Using models such as this one to filter interpretations or select preferred data would risk eliminating sincere interpretations or including erroneous data. The result would therefore be inappropriate for use in model validation or model training.

What is more, control interpreters did not agree perfectly with one another during this exercise, highlighting the subjective nature and innate challenges involved in the photointerpretation of TCC and the risk of using an imprecise benchmark to select crowd interpretations. This lack of precision highlights the need for caution in using photointerpreted TCC values for model training and validation if there are absent or unknown standards for training and consistency. These results additionally show that special consideration is needed, especially when multiple crowd interpreters are used to collect these observations at different sites, without overlap or ratings of interpretation difficulty.

CRediT authorship contribution statement

Jill M. Derwin: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Valerie A. Thomas:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Randolph H. Wynne:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Karen G. Schleeweis:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **John W. Coulston:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **S. Seth Peery:** Visualization, Software, Resources, Data curation, Conceptualization. **Kurt Luther:** Supervision, Project administration, Methodology, Conceptualization. **Greg C. Liknes:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Stacie Bender:** Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization. **Susmita Sen:** Writing – review & editing, Validation, Software, Resources, Methodology, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dr. Randolph H. Wynne and Dr. Valerie Thomas reports financial support was provided by USDA. Dr. Randolph H. Wynne and Dr. Valerie Thomas reports financial support was provided by US Department of Agriculture Forest Service. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the USDA National Institute of Food and Agriculture McIntire Stennis project entitled "Precision Forestry for Southern Pine Carbon Monitoring" and the United States Forest Service (USFS) under the grant "TCC 2021 Research and Development, 18-JV-11242305-035". Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the National Institute of Food and Agriculture (NIFA) or the United States Department of Agriculture (USDA). We would like to thank Muna Khatiwada for her contributions to the data collection for this study and we would like to acknowledge the contributions of the crowd workers who participated in our HIT as well as those who provided meaningful feedback.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103300.

Data availability

Some of the data for this study have been published in the supplemental appendix. Data collected from Amazon's Mechanical Turk including worker identifiers, interpretations, and behavioral indicators about the interpretation environment have been excluded in accordance with our interpreter consent agreement and Institutional Review Board guidelines. The data for this study may be made available to other researchers for research purposes by request via a Data Sharing Agreement with Virginia Tech.

References

Ahler, D.J., Roush, C.E., Sood, G., 2021. The micro-task market for lemons: Data quality on Amazon's Mechanical Turk (working paper). Political Sci. Res. Methods 1–39, URL http://www.gsood.com/research/papers/turk.pdf.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E., Dustdar, S., 2013. Quality control in crowdsourcing systems: Issues and directions. IEEE Internet Comput. 17 (2), 76–81. http://dx.doi.org/10.1109/MIC.2013.20.

Anon, 2012. Crowdsourcing. Bus. Inf. Syst. Eng. 3, http://dx.doi.org/10.1007/s12599-012-0215-7.

Avery, T.E., 1969. Forester's guide to aerial photo interpretation. In: Agriculture Handbook No. 308. U.S. Department of Agriculture Forest Service, Washington, D.C.

Avery, T.E., Burkhart, H.E., 1994. Forest Measurements, fourth ed. McGraw-Hill.

Brand, G.J., Nelson, M.D., Wendt, D.G., Nimerfro, K.K., 2000. The hexagonal/panel system for selecting FIA plots under an annual inventory. In: McRoberts, R.E., Reams, G.A., Deusen, P.C.V. (Eds.), Gen. Tech. Rep. NC-213. Department of Agriculture, Forest Service, North Central Research Station, pp. 8–13.

Breiman, L., Cutler, A., Liaw, A., Wiener, M., 2018. Breiman and Cutler's random forests for classification and regression. URL https://www.stat.berkeley.edu/~breiman/RandomForests/.

Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. 6 (1), 3–5. http://dx.doi.org/10.1177/1745691610393980, PMID: 26162106.

Campbell, J.B., Wynne, R.H., 2011. Image interpretation. In: Introduction to Remote Sensing, fifth ed. Guilford Press, New York, NY, pp. 130–153.

Chen, C., Liaw, A., Breiman, L., 2004. Using Random Forest to Learn Imbalanced Data. Tech. Rep., Department of Statistics, UC Berkeley.

Chmielewski, M., Kucker, S.C., 2019. An MTurk crisis? Shifts in data quality and the impact on study results. Soc. Psychol. Pers. Sci. 11 (4), 464–473. http://dx.doi.org/10.1177/1948550619875149.

Cochran, W., 1977. Sampling Techniques, third ed. John Wiley & Sons, Inc, ISBN: 9780471162407.

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote. Sens. Environ. 37, 35—46.

Congalton, R.G., Green, K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, second ed. CRC Press, Boca Raton, FL, ISBN: 978-1-4200-5512-2.

Cooper, J.S., Mukherji, S.K., Toledano, A.Y., Beldon, C., Schmalfuss, I.M., Amdur, R., Sailer, S., Loevner, L.A., Kousouboris, P., Ang, K.K., Cormack, J., Sicks, J., 2007. An evaluation of the variability of tumor-shape definition derived by experienced observers from CT images of supraglottic carcinomas (acrin protocol 6658). Int. J. Radiat. Oncol. Biol. Phys. (ISSN: 0360-3016) 67 (4), 972–975. http://dx.doi.org/10.1016/j.ijrobp.2006.10.029.

Coulston, J.W., Moisen, G.G., Wilson, B.T., Finco, M.V., Cohen, W.B., Brewer, C.K., 2012. Modeling percent tree canopy cover: A pilot study. Photogramm. Eng. Remote. Sens. (ISSN: 2220-9964) 78 (7), 715–727, URL https://www.fs.usda.gov/treesearch/pubs/40860.

Coulston, J.W., Oswalt, S.N., Carraway, A.B., Smith, W.B., 2010. Assessing forestland area based on canopy cover in a semi-arid region: a case study. For.: An Int. J. For. Res. (ISSN: 0015-752X) 83 (2), 143–151. http://dx.doi.org/10.1093/forestry/cpp039.

de Albuquerque, J.P., Herfort, B., Eckle, M., 2016. The tasks of the crowd: A typology of tasks in geographic information crowdsourcing and a case study in humanitarian mapping. Remote. Sens. 8 (859), 1–22. http://dx.doi.org/10.3390/rs8100859.

Derwin, J.M., Thomas, V.A., Wynne, R.H., Coulston, J.W., Liknes, G.C., Bender, S., Blinn, C.E., Brooks, E.B., Ruefenacht, B., Benton, R., Finco, M.V., Megown, K., 2020. Estimating tree canopy cover using harmonic regression coefficients derived from multitemporal landsat data. Int. J. Appl. Earth Obs. Geoinf. (ISSN: 0303-2434) 86, 101985. http://dx.doi.org/10.1016/j.jag.2019.101985.

Eickhoff, C., de Vries, A.P., 2013. Increasing cheat robustness of crowdsourcing tasks. Inf Retr. 16, 121–137.

Eickhoff, C., Harris, C.G., de Vries, A.P., 2012. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In: SIGIR' 12. Portland, Oregon, pp. 871–879.

Esri and USDA Farm Service Agency, 2014a. USA NAIP imagery: Color infrared. URL https://naip.arcgis.com/arcgis/rest/services/NAIP/ImageServer. Revised in 2020.

Esri and USDA Farm Service Agency, 2014b. USA NAIP imagery: Natural color. URL https://naip.arcgis.com/arcgis/rest/services/NAIP/ImageServer. Revised in 2020.

Esseen, P.-A., Jansson, K.U., Nilsson, M., 2006. Forest edge quantification by line intersect sampling in aerial photographs. For. Ecol. Manag. (ISSN: 0378-1127) 230 (1), 32–42. http://dx.doi.org/10.1016/j.foreco.2006.04.012.

Freese, F., 1962. Elementary forest sampling. In: Agriculture Handbook, vol. 232, U.S. Department of Agriculture, Forest Service, Washington, DC, URL http://www.fs.fed.us/fmsc/ftp/measure/cruising/other/docs/AgHbk232.pdf.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., Obersteiner, M., 2009. Geo-wiki.org: The use of crowdsourcing to improve global land cover. Remote. Sens. (ISSN: 2072-4292) 1 (3), 345–354. http://dx.doi.org/10.3390/rs1030345, URL http://dx.doi.org/10.3390/rs1030345.

Galloway, A.W.E., Tudor, M.T., Haegen, W.M.V., 2006. The reliability of citizen science: A case study of oregon white oak stand surveys. 34 (5), 1425–1429.

Gatziolis, D., 2012. Comparison of LiDAR- and photointerpretation-based estimates of canopy cover. In: Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists. e-Gen. Tech. Rep. SRS-157, U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, pp. 231—235, URL https://www.fs.usda.gov/treesearch/pubs/41012.

Goodchild, M., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. http://dx.doi.org/10.1007/s10708-007-9111-y.

Haklay, M., Weber, P., 2008. OpenStreetMap: User-generated street maps. IEEE Pervasive Comput. 7 (4), 12–18.

Jackson, T.A., Moisen, G.G., Patterson, P.L., Tipton, J., 2012. Repeatability in photo-interpretation of tree canopy cover and its effect on predictive mapping. In: Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists. e-Gen. Tech. Rep. SRS-157, U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, pp. 189—192, URL https://www.fs.usda.gov/treesearch/pubs/41006.

Jennings, S., Brown, N., Sheil, D., 1999. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. For.: Int. J. For. Res. (ISSN: 0015-752X) 72 (1), 59–74. http://dx.doi.org/10.1093/forestry/72.1.59.

Ma, Q., Su, Y., Guo, Q., 2017. Comparison of canopy cover estimations from airborne LiDAR, aerial imagery, and satellite imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 10 (9), 4225–4236.

Mason, A.W., Watts, J.D., 2010. Financial incentives and the performance of crowds. ACM SIGKDD Explor. Newsl. 11 (2), 100–108.

McRoberts, R., Tomppo, E., 2007. Remote sensing support for national forest inventories. Remote. Sens. Environ. 110, 412–419. http://dx.doi.org/10.1016/j.rse.2006.09.034.

Moss, A.J., Litman, L., 2018. After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it [blog post]. CloudResearch URL https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.

Parasuraman, R., 1986. Vigilance, monitoring, and search. In: Boff, J.R., Kaufmann, L., Thomas, J.P. (Eds.), Handbook of Human Perception and Performance, Volume 2: Cognitive Processes and Performance. Wiley, New York, pp. 1–49.

Pozzi, F., Di Matteo, T., Aste, T., 2012. Expoential smoothing weighted correlations. Eur. Phys. J. B 85 (175), 1–21. http://dx.doi.org/10.1140/epjb/c2012-20697-x.

R. Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Riemann, R., Liknes, G., O'Neil-Dunne, J., Toney, C., Lister, T., 2016. Comparative assessment of methods for estimating tree canopy cover across a rural-to-urban gradient in the mid-atlantic region of the USA. Env. Monit. Assess. 188 (297), 1–17. http://dx.doi.org/10.1007/s10661-016-5281-8.

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., Vukovic, M., 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. 5, (1), pp. 321–328, URL https://ojs.aaai.org/index.php/ICWSM/article/view/14105,

Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., Jacobs, D., 2012. Dynamic changes in motivation in collaborative citizen-science projects. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. In: CSCW '12, Association for Computing Machinery, New York, NY, USA, pp. 217–226. http://dx.doi.org/10.1145/2145204.2145238.

Ruefnacht, B., Heyer, J., Johnson, V., Goetz, W., Bender, S., Schleeweis, K., Megown, K., 2022. Forest Service Tree Canopy Cover Mapping: 2016 Product Suite and Methods. Tech. Rep. GTAC-10264-RPT1, U.S. Department of Agriculture, Forest Service, Geospatial Technology and Applications Center, 47 p.

Schepaschenko, D., See, L., Lesiv, M., McCallum, I., Fritz, S., Salk, C.F., Moltchanova, E., Perger, C., Shchepashchenko, M., Shvidenko, A., Kovalevskyi, S., Gilitukha, D., Albrecht, F., Kraxner, F., Bun, A., Maksyutov, S., Sokolov, A., Dürauer, M., Obersteiner, M., Karminov, V., Ontikov, P., 2015. Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics. Remote. Sens. Environ. (ISSN: 0034-4257) 162, 208—220. http://dx.doi.org/10.1016/j.rse.2015.02.011.

Schroeder, T.A., Schleeweis, K.G., Moisen, G.G., Toney, C., Cohen, W.B., Freeman, E.A., Yang, Z., Huang, C., 2017. Testing a Landsat-based approach for mapping disturbance causality in U.S. forests. Remote. Sens. Environ. 195, 230–243. http://dx.doi.org/10.1016/j.rse.2017.03.033.

See, L., Comber, A., Salk, C.F., Fritz, S., van der Velde, M., Perger, C., Schill, C., McCallum, I., Kraxner, F., Obersteiner, M., 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. PLoS One 8 (7), 1–11. http://dx.doi.org/10.1371/journal.pone.0069958.

Smith, A.M., Falkowski, M.J., Hudak, A.T., Evans, J.S., Robinson, A.P., Steele, C.M., 2009. A cross-comparison of field, spectral, and lidar estimates of forest canopy cover. Can. J. Remote. Sens. 35 (5), 447–459. http://dx.doi.org/10.5589/m09-038.

Stokel-Walker, C., 2018. Bots on Amazon's Mechanical Turk are ruining psychology studies. New Sci. 1—4, (Accessed 8 November 2018).

Szalma, J.L., Hancock, P.A., Warm, J.S., Dember, W.N., Parsons, K.S., 2006. Training for vigilance: Using predictive power to evaluate feedback effectiveness. Hum. Factors: J. Hum. Factors Ergon. Soc. 48, 682–692.

Tim, 2021. Such thing as a weighted correlation?. arXiv:https://stats.stackexchange.com/q/222107. Cross Validated, (Accessed 24 March 2021).

Tipton, J., Moisen, G., Patterson, P., Jackson, T.A., Coulston, J., 2012. Sampling intensity and normalizations: Exploring cost-driving factors in nationwide mapping of tree canopy cover. In: Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists. e-Gen. Tech. Rep. SRS-157, U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, pp. 201—208, URL https://www.srs.fs.usda.gov/pubs/41008.

United States Department of Agriculture, Forest Service, 2019. 2016 FS analytical TCC. URL https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/.

USDA Farm Service Agency, 2018. NAIP coverage 2002–2018. URL https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/status-maps/pdfs/naipcov{_}2018.pdf.

Van Coillie, F., Gardin, S., Anseel, F., Duyck, W., Verbeke, L.P.C., Wulf, R.R.D., 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. http://dx.doi.org/10.1080/01431161.2013.873152.

Wang, J., Ipeirotis, P.G., Provost, F., 2013. A framework for quality assurance in crowdsourcing. (CBA-13-06).

Wang, M.-J.J., Lin, S.-C., Drury, C.G., 1997. Training for strategy in visual search. Int. J. Ind. Ergon. 20, 101–108.

Wiggins, A., Crowston, K., 2011. From conservation to crowdsourcing: A typology of citizen science. In: 2011 44th Hawaii International Conference on System Sciences. pp. 1–10.

Yu, L., André, P., Kittur, A., Kraut, R., 2014. A comparison of social, learning, and financial strategies on crowd engagement and output quality. pp. 967–977. http://dx.doi.org/10.1145/2531602.2531729.

Yu, L., Ball, S.B., Blinn, C.E., Moeltner, K., Peery, S., Thomas, V.A., Wynne, R.H., 2015. Cloud-sourcing: Using an online labor force to detect clouds and cloud shadows in landsat images. Remote. Sens. (ISSN: 2072-4292) 7 (3), 2334–2351. http://dx.doi.org/10.3390/rs70302334.