Remote Sensing of Environment 208 (2018) 145–153

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A comparison of resampling methods for remote sensing classification and accuracy assessment

Mitchell B. Lyons[a,*], David A. Keith[a,b], Stuart R. Phinn[c], Tanya J. Mason[a], Jane Elith[d,e]

[a] Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, UNSW Australia, Sydney 2052, Australia
[b] New South Wales Office of Environment and Heritage, Sydney 1232, Australia
[c] Remote Sensing Research Centre, School of Earth and Environmental Sciences, University of Queensland, Brisbane 4072, Australia
[d] School of BioSciences, University of Melbourne, Melbourne 3010, Australia
[e] ARC Centre of Excellence for Environmental Decisions, University of Melbourne, Melbourne 3010, Australia

ARTICLE INFO

Keywords:
Validation
Bootstrapping
Cross validation
Bias
Variance
Land cover mapping
Vegetation mapping
Class area proportion
Population parameter

ABSTRACT

Maps that categorise the landscape into discrete units are a cornerstone of many scientific, management and conservation activities. The accuracy of these maps is often the primary piece of information used to make decisions about the mapping process or judge the quality of the final map. Variance is critical information when considering map accuracy, yet commonly reported accuracy metrics often do not provide that information. Various resampling frameworks have been proposed and shown to reconcile this issue, but have had limited uptake. In this paper, we compare the traditional approach of a single split of data into a training set (for classification) and test set (for accuracy assessment), to a resampling framework where the classification and accuracy assessment are repeated many times. Using a relatively simple vegetation mapping example and two common classifiers (maximum likelihood and random forest), we compare variance in mapped area estimates and accuracy assessment metrics (overall accuracy, kappa, user, producer, entropy, purity, quantity/allocation disagreement). Input field data points were repeatedly split into training and test sets via bootstrapping, Monte Carlo cross-validation (67:33 and 80:20 split ratios) and *k*-fold (5-fold) cross-validation. Additionally, within the cross-validation, we tested four designs: simple random, block hold-out, stratification by class, and stratification by both class and space. A classification was performed for every split of every methodological combination (100's iterations each), creating sampling distributions for the mapped area of each class and the accuracy metrics. We found that regardless of resampling design, a *single* split of data into training and test sets results in a large variance in estimates of accuracy and mapped area. In the worst case, overall accuracy varied between ~40–80% in one resampling design, due only to random variation in partitioning into training and test sets. On the other hand, we found that all resampling procedures provided accurate estimates of error, and that they can also provide confidence intervals that are informative about the performance and uncertainty of the classifier. Importantly, we show that these confidence intervals commonly encompassed the magnitudes of increase or decrease in accuracy that are often cited in literature as justification for methodological or sampling design choices. We also show how a resampling approach enables generation of spatially continuous maps of classification uncertainty. Based on our results, we make recommendations about which resampling design to use and how it could be implemented. We also provide a fully worked mapping example, which includes traditional inference of uncertainty from the error matrix and provides examples for presenting the final map and its accuracy.

1. Introduction

Categorical maps (e.g. land cover, land use, vegetation community type, soil type etc.) are still one of the fundamental underlying information sources for decision making for many scientific, conservation, and management activities. There is a range of strategies for making these maps and assessing their accuracy. Remote sensing approaches are common, falling into the general category of "image classification". Often, these approaches involve using some kind of modelling approach to map, from image data, a set of known classes using known cases of those classes for training. This contrasts with unsupervised approaches, which do not use operator-controlled

* Corresponding author.
E-mail address: mitchell.lyons@gmail.com (M.B. Lyons).

https://doi.org/10.1016/j.rse.2018.02.026
Received 1 September 2017; Received in revised form 1 February 2018; Accepted 16 February 2018
Available online 20 February 2018
0034-4257/

training information. Informative, transparent and statistically robust presentation of the accuracy and reliability of such maps is critical to enable their use in scientific, legal and economic decisions (Foody 2004, 2015; Olofsson et al. 2014).

Greater accuracy is of course desirable, but just as importantly it is critical that informative estimates of accuracy and uncertainty are provided with a map. This is particularly true when accuracy values directly inform a decision-making process. Accuracy is typically reported in terms of a predictive accuracy metric describing the agreement between mapped values and the known values for those cases (i.e. 'overall accuracy'). Most other common metrics are variations of the concept of overall accuracy. For example, metrics for individual classes based on commission and omission error are common. Kappa metrics, which correct for chance agreement, are also widely reported, though they have recently been criticised for their assumptions about the accuracy of a "random" classification (e.g. Pontius & Millones 2011). Use of overall accuracy itself has also been questioned for its relevance when used across different mapping scenarios (e.g. Foody 2004; Stehman et al. 2008). The other statistics commonly estimated from maps are mapped area or estimates of population parameters. Likewise, it is important to understand the accuracy and uncertainty in these values, and there has been much research on this topic also (e.g. McRoberts et al. 2011; McRoberts 2014).

The sampling design for acquiring input data, and the type of model used for mapping, varies widely and influences the accuracy of resultant maps (Stehman et al. 2008; Zhen et al. 2013). The sampling design can vary, both in terms of how the underlying data are collected and how they are partitioned to train and test the mapping procedure (e.g. Foody 2002; Zhen et al. 2013; Olofsson et al. 2014). Modelling approaches range from simple methods, such as maximum likelihood and nearest neighbourhood classifiers to more complex methods such as random forests, support vector machines and boosted regression trees (e.g. Brenning 2009).

For data partitioning, the most common strategy is to choose some ratio to split the data into training and test sets; the training set informs the model, and a single test set is held out to calculate accuracy metrics post hoc. The split ratio varies, but the training sample commonly comprises 50–80% of the full dataset. The training set may be selected based on one of several alternative strategies, including: simple random sample, or a random sample stratified by class, by class and spatial location, or split spatially by blocks or circles (Olofsson et al. 2014). Split ratio and sampling design can also affect both the map and the estimate of its accuracy (Zhen et al. 2013). Regardless, the use of a *single* split of data into training and test sets may provide misleading information about estimates and their uncertainty. This is because any one split could be an unrepresentative sample of the data, so the user has no idea how close the class area or accuracy estimates are to the *truth*.

The purpose of accuracy assessment is to estimate the error and uncertainty of the output classification, to either choose the most appropriate mapping procedure or to inform interpretation of the output. This information is used in combination with estimates of population parameters (and their uncertainty). Much of the focus on development of accuracy metrics and best practice in model evaluation has been to provide more meaningful estimates of map accuracy and population parameters (Olofsson et al. 2014). This research has concluded that associated measures of uncertainty are critical for use and interpretation (McRoberts 2014; Olofsson et al. 2014). Additionally, better scientific, conservation and management outcomes result when knowledge of uncertainty is incorporated into decisions (Burgman et al. 2005; Guisan et al. 2013; Foody 2015). Indeed, for many applications, accuracy estimates explicitly inform decision-making, and yet it is still not common place to include estimates of both accuracy and uncertainty along with maps. This is despite the growing range of methods for doing so.

Resampling procedures such as bootstrapping and cross-validation can be used to estimate map accuracy and associated uncertainty (i.e. variance or confidence intervals) in a relatively unbiased manner (Efron and Tibshirani 1997). These methods are commonly implemented for predicting geographic distributions, for example for species or ecological communities (Roberts et al. 2017). They can also be effective in remote sensing frameworks and have been employed in various ways, often with a focus on estimation of mapped areas or population parameters (e.g. Weber and Langille 2007; Brenning 2009; McRoberts et al. 2011; Champagne et al. 2014; Gallaun et al. 2015; Hsiao and Cheng 2016). However, resampling approaches remain uncommon for assessing mapping accuracy and its uncertainty (e.g. standard error/confidence intervals). The premise is quite simple; instead of using a single split of the data to produce the accuracy metric, the splitting is repeated multiple times using a chosen resampling framework. Both the map and accuracy results are produced for every iteration, giving a sample distribution of map and accuracy results. This sampling distribution can then be summarised to provide an empirical estimate of accuracy along with its uncertainty.

Independent samples and subsequent construction of the error matrix have been consistently viewed as the desirable way to estimate predictive performance. Indeed, this construct has an important role, however, truly independent samples are rarely available. Resampling can provide some of the advantages of an independent sample and it provides accurate estimates along with meaningful information about variance. Alternatives to repeated classifications have been shown to be useful (e.g. McKenzie et al. 1996; Hess and Bay 1997; Gallaun et al. 2015), and have been advocated for some time now (e.g. Foody 2004). These methods have often been motivated by reducing computational cost, but modern classification methods and more powerful computers mean that resampling frameworks are now tractable for most users. However, there are no comprehensive comparisons of different resampling strategies for categorical mapping and accuracy assessment, nor are there easy to use workflows in common image processing and GIS software packages for developing and applying resampling frameworks.

In this paper, we compared resampling approaches to the traditional single-split, train and hold-out test set approach. Our primary objective was to compare the way mapping accuracy is assessed, but we also compared the estimated areas of each mapping class. We tested whether the accuracy and area estimated depended on the design of the test and training sets, that is, whether bootstrapping or cross-validation (Monte Carlo or *k*-fold) was used, if and how stratification was used (simple random, thematic and/or spatial stratification), and the ratio at which samples were partitioned into training and test sets. We also tested several aspects of the classification framework, including the classification model (maximum likelihood and random forest) and the accuracy assessment metric used (overall accuracy, kappa, user/producer accuracy, entropy, purity and quantity/allocation disagreement). Using a worked example, we compared a resampling approach to more traditional approaches based on a single hold-out test set for validation. The data used in this study included a dense set of field observations of vegetation communities from south east Australia, and ADS40 high resolution (40 cm) multispectral imagery. We show that all resampling approaches gave consistent measures of accuracy, as well as providing useful estimates of uncertainty in both class area and accuracy. We discuss the limitations of commonly used approaches in this context, and identify practical options for users seeking to improve the robustness of maps and their accuracy assessments by implementing a resampling based mapping approach.

2. Methods

2.1. Field and image data

The study area (~50 km^2) is within the O'Hares Creek catchment in Dharawal National Park and Nature Reserve, nearby Sydney, Australia.

This study site belongs to Australia's Long Term Ecological Research Network (Mason et al. 2017). As part of this network, 25,500 × 500 m blocks were established using constrained random placement within the 10 × 5 km study area. The blocks were constrained to be non-overlapping and not on any obviously disturbed areas. The density of points needed for stable proportion estimates was determined (sensu Nascimento and Laurance 2002) to be about 10 points per hectare. Points were placed randomly within each block and were constrained such that all points were > 5 m apart (Fig. S1). There were ~230–260 points per block, and most points were at least 10–15 m apart. Three different observers allocated each of the points to one of four vegetation classes: Eucalyptus woodland (n = 3732), Wet heath (n = 1404), Banksia thicket (n = 783) or Tea tree thicket (n = 231). Allocations were assigned via manual interpretation of the ADS40 imagery. The three observers were trained in image interpretation, which included site visits to gain familiarity with vegetation types. Observers unanimously agreed for ~75% of points and ~98% of points had a two-observer majority. For vegetation class labels in this study, we used the majority agreement, and randomly broke the remaining ties. The image data used for the modelling were captured in 2008 with a Leica Geosystems ADS40 sensor (40 cm pixels), as 12-bit data (scaled to 16-bit) with blue (428–492 nm), green (533–587 nm), red (608–662 nm), and near-infrared (833–887 nm) bands. The data supplier applied industry standard orthorectification and colour matching corrections to produce the image mosaic. For each point we extracted the mean of a 3 × 3 pixel neighbourhood for each image band.

2.2. Analysis workflow

The data analysis routine included four main steps (Fig. 1): 1) draw a sub-sample from the whole field data set, 2) select a test and training set from the sub-sample, 3) fit a supervised classification model to image data corresponding to a training set, and 4) calculate accuracy and area metrics. At step two, many thousands of paired test and training sets were selected via a range of resampling strategies, and to each of these, steps 3 and 4 were applied. The whole process was repeated for 50 different sub-samples. We use the nomenclature of 'training and test' sets, which equates to 'calibration and validation' or 'map and reference' sets in other literature.

All analysis was performed in the R programming language. Maintained code and classification data are available at the corresponding authors github page: https://github.com/mitchest/rs-accuracy-variance, and a static release as per this paper is available on Zenodo (https://doi.org/10.5281/zenodo.1146099). Unfortunately, the full original image data set is unable to be redistributed.

2.2.1. Sub-sampling of field data

The very dense set of "ground" observations enabled sub-sampling in step one. Since the aim of this paper was to compare the behaviour of accuracy metrics, we decided to fix the way the sub-sample was drawn from the full data set, to mimic a well-designed field survey. We randomly drew 10% of the data for each sub-sample. The sub-sample was only constrained to sample across each survey block with a probability that mimicked the full field data set. Indicator variograms for each class indicated a small amount of spatial correlation between sample points, particularly Eucalypt Woodland where a reasonable proportion of points were separated by less than the variogram range (Fig. S2). Spatial autocorrelation is discussed in detail later.

2.2.2. Resampling methodology

The sub-sample was then used to generate the training and test set pairs, using a range of resampling designs. The first resampling design was the bootstrap (Walker's alias method, as implemented in base R). In this case, the training set was generated by resampling the whole sub-sample with replacement (i.e. bootstrapped sensu Efron and Tibshirani 1997). Bootstrapping results in some sites not being selected, and this set of sites was taken as the test set (i.e. the out-of-bag sample). The next resampling design was k-fold cross-validation, and we nominally chose five folds as it has previously been used in related literature (Brenning 2009; Roberts et al. 2017). k-fold cross validation involves partitioning the data once into k folds, so for example a 5-fold cross-validation divides the data into five partitions. Then, in an iterative manner, each fold is used once as the test set whilst the remaining $k - 1$ folds form the training sets. The final resampling design was Monte

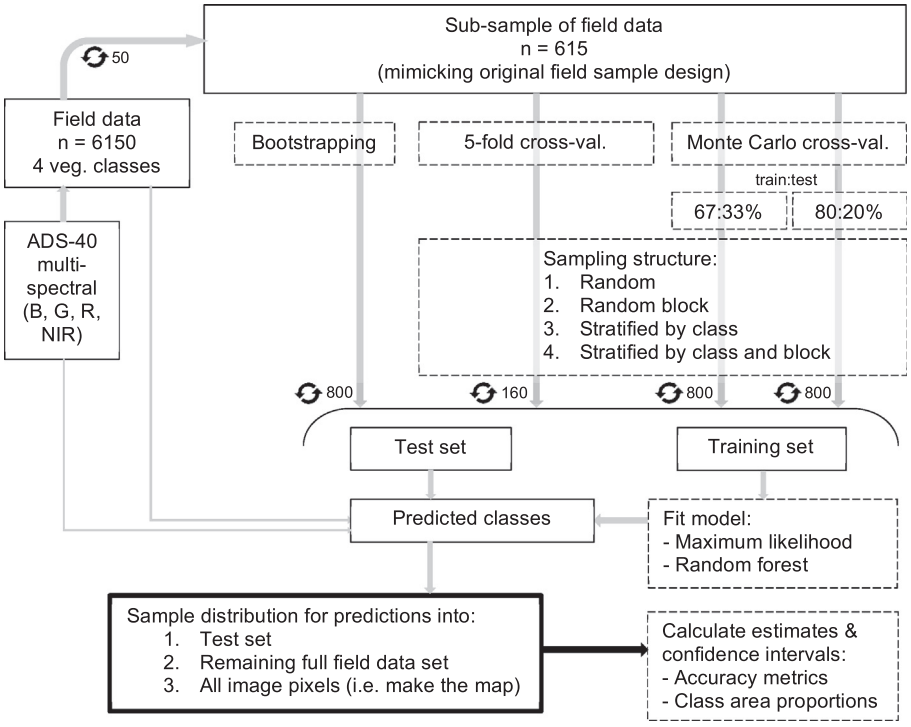

Fig. 1. Flow chart of the resampling framework. A sampling distribution is generated for every methodological combination, for which an estimate and confidence interval are calculated.

Carlo cross-validation, also known as repeated random cross-validation. Monte Carlo cross-validation involves simply repeatedly splitting the data at some ratio into a training and test data set. We chose two ratios (67:33 and 80:20) that are commonly seen in the literature for splitting data for mapping and independent accuracy assessment. Additionally, 67:33 and 80:20 are loosely analogous to the bootstrap and a 5-fold cross-validation, respectively (Hastie et al. 2009). The important thing to note at this point is that any individual split of the data into a training and test set (from any of the resampling designs) is analogous to a typical mapping scenario where the data are split once, and one portion is used to train the map and the other to assess its accuracy.

For each cross-validation approach we also imposed different structures for how the training set was selected. Using the most common strategies (Zhen et al. 2013), we selected training sample units that were: simple random; block hold-out (training and test data were from different survey blocks); stratified by vegetation class (preserving class proportion); or stratified by both vegetation class and by survey block. Note that in the last option, stratification was additionally constrained so that the same number of sample units were taken from each survey block (entirely different design to the block hold-out).

2.2.3. Classification methods

To each of the training sets, we fitted two common remote sensing classification models to the ADS40 image data – discriminant analysis (commonly seen as maximum likelihood in the remote sensing literature) and random forest. For the maximum likelihood classification we used a moments estimator with a flat prior. For the random forest classifier we used 250 trees and allowed two of the image bands to be used at each node. Other less important parameterisation can be found in the code provided (see beginning Section 2.2).

2.2.4. Accuracy estimates

Accuracy assessment and class metrics were calculated for all the methodological combinations of resampling. For the accuracy metrics, we used: overall accuracy (sample counts); Kappa (standard); user and producer accuracy; entropy (Shannon); purity; and quantity and allocation disagreement. These are all standard metrics that can be calculated from the error matrix; descriptions are provided in Table S1, Supplementary Material. Accuracy assessments usually focus on performance in the test sample, presented as an error matrix. We additionally calculated the metrics for predictions into remaining sites in the full field data set (i.e. before sub-sampling in 2.2.1. occurs), to provide a closer test of “truth”. The result of a resampling approach is a sampling distribution for each of the accuracy metrics for every combination of the resampling methodology. The median of the empirical sampling distribution was taken as the accuracy estimate, and a percentile method was used to create the relevant confidence intervals from the sampling distribution (i.e. 5th and 95th percentile for a 90% confidence interval; 2.5th and 97.5th percentile for a 95% confidence interval).

2.2.5. Area estimates

There is considerable interest for calculating area estimates (with uncertainty) in the literature (e.g. McRoberts et al. 2011; Olofsson et al. 2014). Thus we opted to also include estimates of class area and uncertainty into our comparison of resampling approaches. To do this we predicted the classification out into the entire image data set (i.e. ‘made the map’) for all resampling designs, and calculated the mapped area of each of the vegetation class. Class areas were estimated using the population error matrix with mapped areas corrected for classification error (sensu Olofsson et al. 2014). As for accuracy, the median was taken as the area estimate for a given class, and the relevant percentile as the confidence interval.

2.2.6. Resampling parameterisation

For resampling, we used 800 bootstrap iterations, 800 iterations of Monte Carlo cross-validation for each of the two split fractions, and 160 repeats of the 5-fold cross-validation (to generate 800 test-training pairs). Stabilisation of the resample estimate is more important than the absolute number of resamples (Efron and Tibshirani 1997), so we used a greater number of iterations than what we expected to be sufficient (Hastie et al. 2009), and examined the stabilisation post hoc. The number of iterations can be considered sufficient when the resampling estimates stabilize. We repeated the analysis for 50 different sub-samples (i.e. as in Section 2.2.1). We calculated the accuracy metrics for all 50 sub-samples, but only calculated map area estimates for one of the sub-samples (the computational overhead for predicting the full map is very large).

2.3. Example application

To demonstrate a more tractable example of applying a resampling framework to a mapping scenario, a small, fully-worked example is provided. We took a smaller subset of the field and image data (Fig. S1) and using a maximum likelihood classifier, implemented a Monte Carlo cross-validation (67:33 split) with simple random sampling and 800 iterations. Accuracy and class area estimates and confidence intervals were calculated as per Section 2.3 above. We then calculated additional estimates and confidence intervals for each mapping iteration based on two existing approaches. Firstly, we calculated accuracy and area estimates with confidence intervals based on standard error from the full error matrix (sensu Olofsson et al. 2014). Secondly, we calculated the same values based on a simple bootstrapping approach (resampling the error matrix sensu Hess and Bay 1997) with 1000 iterations. This provided a reference to what might be performed in a typical scenario using a single-split train and hold-out test approach. We also provide this mapping example as a separate fully-reproducible example, including image data (via R code; see beginning Section 2.2).

3. Results

The full field data set had 6150 points, so each 10% sub-sample iteration contained 615 points for analysis. Combining the 50 sub-samples and the various methodological combinations, around 550,000 classifications were trained. As there were no major differences in results or patterns between the two classification models, we focus on results from the maximum likelihood classifier, for simplification and familiarity (random forest results are in Supplementary Material).

3.1. Effect of resampling and stratification design on accuracy metrics

The median values for the accuracy metrics calculated on the test sets were remarkably similar across the different resampling designs (Fig. 2, Figs. S3–4). Variance differed somewhat among resampling and stratification designs, but was mostly related to the size of the test set. As such, the bootstrap and 67:33 split with Monte Carlo cross-validation had the smallest variance, though most of the difference occurred in the extremities. For example, the 90% interval for overall accuracy was comparable (spanning ~15%) for most methodological combinations. Inducing stratification by class reduced variance in entropy, purity and quantity/allocation disagreement (which are more sensitive to inter-class errors). All metrics were more variable under the block hold-out design, and the median values indicated slightly lower performance.

Within each vegetation class, median values for user and producer accuracies on the test set were consistent across resampling and stratification design (Figs. S5–6). Variance was somewhat consistent within class, but was again greater in the block hold-out design and smaller with class stratification. The range in variance differed greatly among classes, which was related to the test set sample size and the resulting smaller sample sizes. For all designs, user accuracy and producer accuracy for Tea Tree Thicket had 90% confidence intervals that

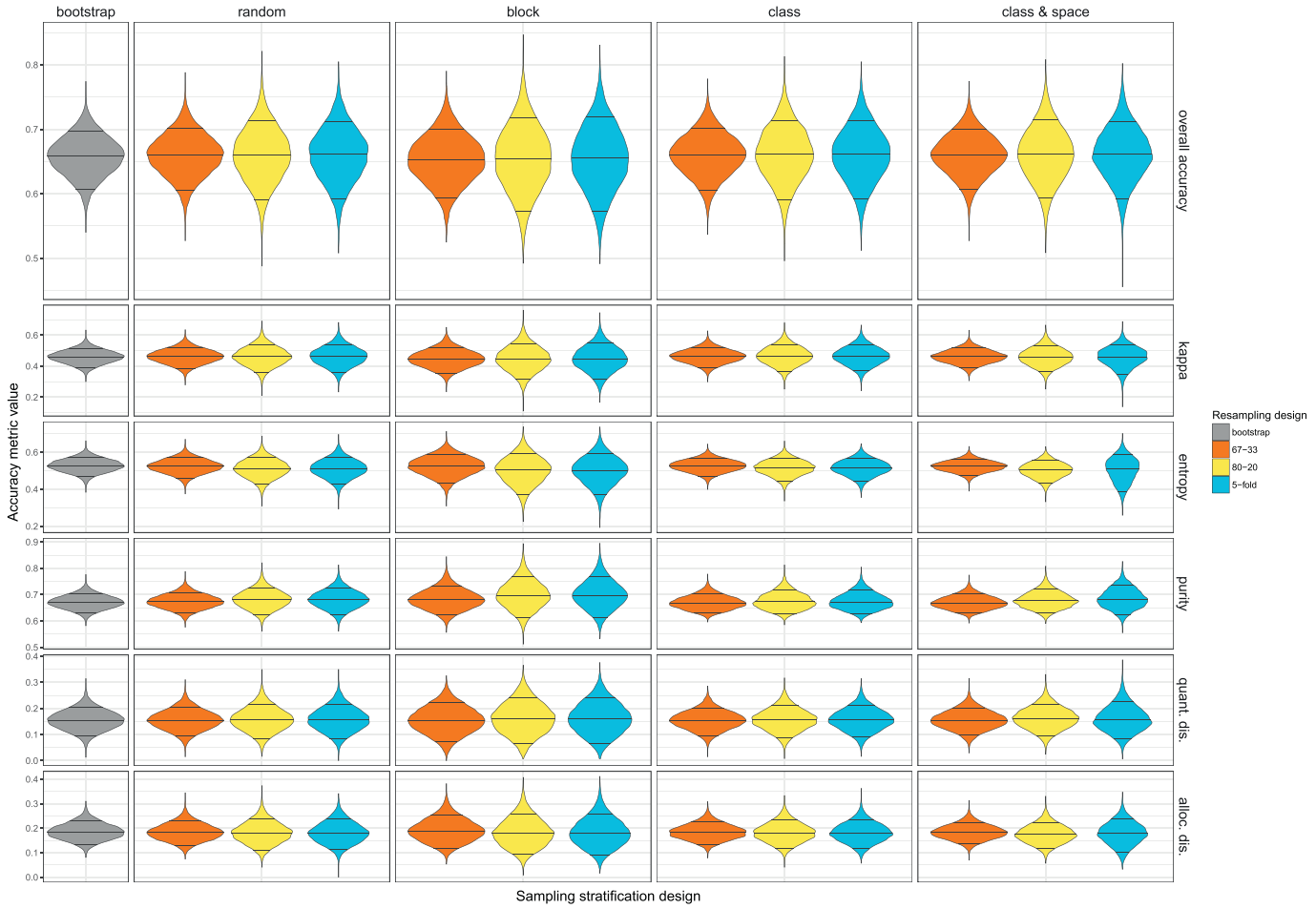

Fig. 2. Comparison of accuracy assessment metrics obtained via a resampling framework for a maximum likelihood classification of four vegetation classes from 40 cm imagery (B,G,R,NIR). Horizontal lines on violins indicate the median and the 90% confidence interval. Resampling procedures included bootstrapping, and Monte Carlo (train:test ratios of 67:33 and 80:20) and 5-fold cross validation. Cross-validation resampling was structured in four ways: simple random sampling; stratified spatially; stratified by vegetation class; or stratified by both vegetation class and spatially.

spanned > 50% in accuracy values, often > 75%.

For all accuracy metrics across all resampling and stratification designs, the median values calculated on the test set were very close to the median values calculated on the full field data set (Figs. 2, S2–3). The variance for the full field data set was understandably smaller due to the size of test set.

Kappa metrics have been critiqued in recent times (sensu Pontius and Millones 2011). Here we do not intend to add to the discussion, we simply note that Kappa did not alter the interpretation of overall accuracy in any way with respect to performance of either the classification models or resampling designs (as evident in Figs. 2, S2–3).

3.2. Effect of resampling and stratification design on class area estimates

The median values for the estimated proportions of each vegetation class were consistent across resampling designs (Fig. 3). Variance was larger in the block hold-out design and for the bootstrap (90% interval ~4–7% area), but under the other resampling designs there were no great differences relating to the size of the test set (90% interval mostly ~2–3% area).

3.3. Sampling efficiency

We also inspected results for each sub-sample separately (i.e. each of the 50 repeats of taking 615 cases from the full field data set, Fig. 1), to see whether any particular selection of a sub-sample was having undue influence on the results (Fig. S7). Variance differed slightly among sub-samples, but the median values were very stable. Taking overall accuracy as an example, the median values were within 1–2% of each other among sub-samples. We note here that we use the median area and accuracy values as an approximation of the "truth". Our results suggest that the medians (or indeed the mean) are a good approximation in this context, but concede that the actual truth is generally unattainable for remote sensing of vegetation communities.

Realising that we used relatively computationally demanding iteration numbers for resampling (i.e. we produced 800 samples for each), we evaluated efficiency by comparing the metric values to number of resampling iterations. Again taking overall accuracy as an example, median values stabilised after about 200 iterations, and variance was well captured after about 50 iterations (Fig. S8). At severely reduced numbers of iterations (i.e. < 25), *k*-fold cross-validation seemed to be most effective at estimating the variance, but the median values were usually inaccurate.

3.4. Example application

Fig. 4 shows the results of the mapping example. Table 1 gives accuracy and area estimate results that compare the resampling approach to a more traditional accuracy assessment approach applied to two selected iterations of the sampling. For accuracy, the confidence intervals based on standard error (from the population error matrix sensu Olofsson et al. 2014) and the bootstrapped confidence intervals (from

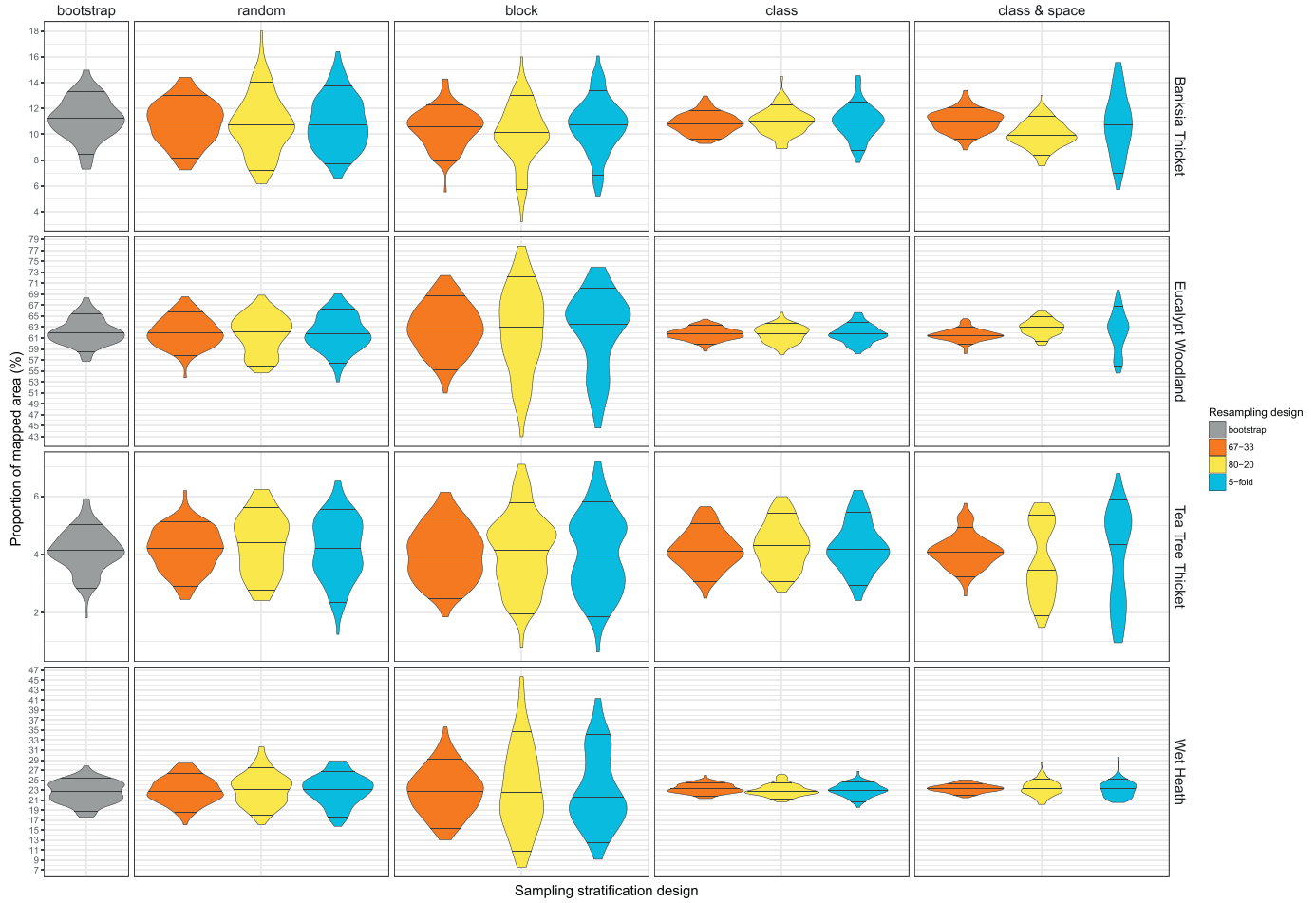

Fig. 3. Comparison of estimated class area for four vegetation classes obtained via a resampling framework for a maximum likelihood classification of 40 cm imagery (B,G,R,NIR). Horizontal lines on violins indicate the median and the 90% confidence interval. Resampling procedures included bootstrapping, and Monte Carlo (train:test ratios of 67:33 and 80:20) and 5-fold cross validation. Cross-validation resampling was structured in four ways: simple random sampling; stratified spatially; stratified by vegetation class; or stratified by both vegetation class and spatially.

the sample count error matrix sensu Hess and Bay 1997) were mostly comparable, and they usually contained the median estimate from the full resampling routine. However, the estimates themselves (that is the actual 'accuracy value') for both approaches were quite variable among mapping iterations, and were often at the upper or lower limits of the confidence interval obtained from the full resampling routine. As noted in the literature (e.g. Olofsson et al. 2014), the confidence intervals based on standard error overestimated the error in mapped areas. Conversely, the bootstrapped confidence intervals were mostly underestimated. Both the bootstrap and standard error confidence interval approaches were misleading when all (or almost all) the test samples for a class were classified correctly (resulting in standard errors of zero).

4. Discussion

The take-home message from this paper is that using a single hold-out test set, as done commonly in remote sensing applications, gives an appreciable chance of a misleading estimate of accuracy, and makes it difficult to accurately quantify uncertainty. This is true across different resampling designs for accuracy assessment, different accuracy metrics and different classification methods applied to the training data. The core of the problem is that with a single split, the modeller has no idea whether they happen to have chosen a "good" or "bad" training and test sample, that is, how representative it is of their data. In the worst case, if our data were split into a training set for mapping (maximum likelihood classifier) and a test set for an independent assessment of overall accuracy, then the estimated accuracy could have been almost any value between ~40–80% (Fig. 2). If we used a resampling design with a larger test set, that range reduced to ~50–75%. However, using a resampling approach we inferred that the (median) accuracy was 62% for both cases, but that the confidence intervals were different (90% intervals of 53–71% and 57–68% respectively). In fact, whatever resampling design we used we always inferred that our map was around 62% accurate (random forest classifier ~65%), rather it was the variance that changed among methodological options. Calculating confidence intervals from one individual iteration (either approach we tested) could sometimes account for this problem (Table 1), but the estimates were still unrepresentative and uninformative in many cases. Furthermore, our resampling approach was also useful for estimating class areas and their corresponding uncertainty. As for accuracy, class area proportions varied greatly among individual mapping iterations, but the area proportions estimated from the sampling distributions were consistent across all resampling designs (Fig. 3).

4.1. Variance is a critical aspect of classification assessment

In line with existing literature, we assert that quantifying variance is as important as any other methodological consideration for either generating (Hess and Bay 1997, Brenning 2009, Champagne et al. 2014, Gallaun et al. 2015, Hsiao and Cheng 2017) or comparing (McKenzie et al. 1996; Foody 2004; Stehman et al. 2008) map accuracy. The

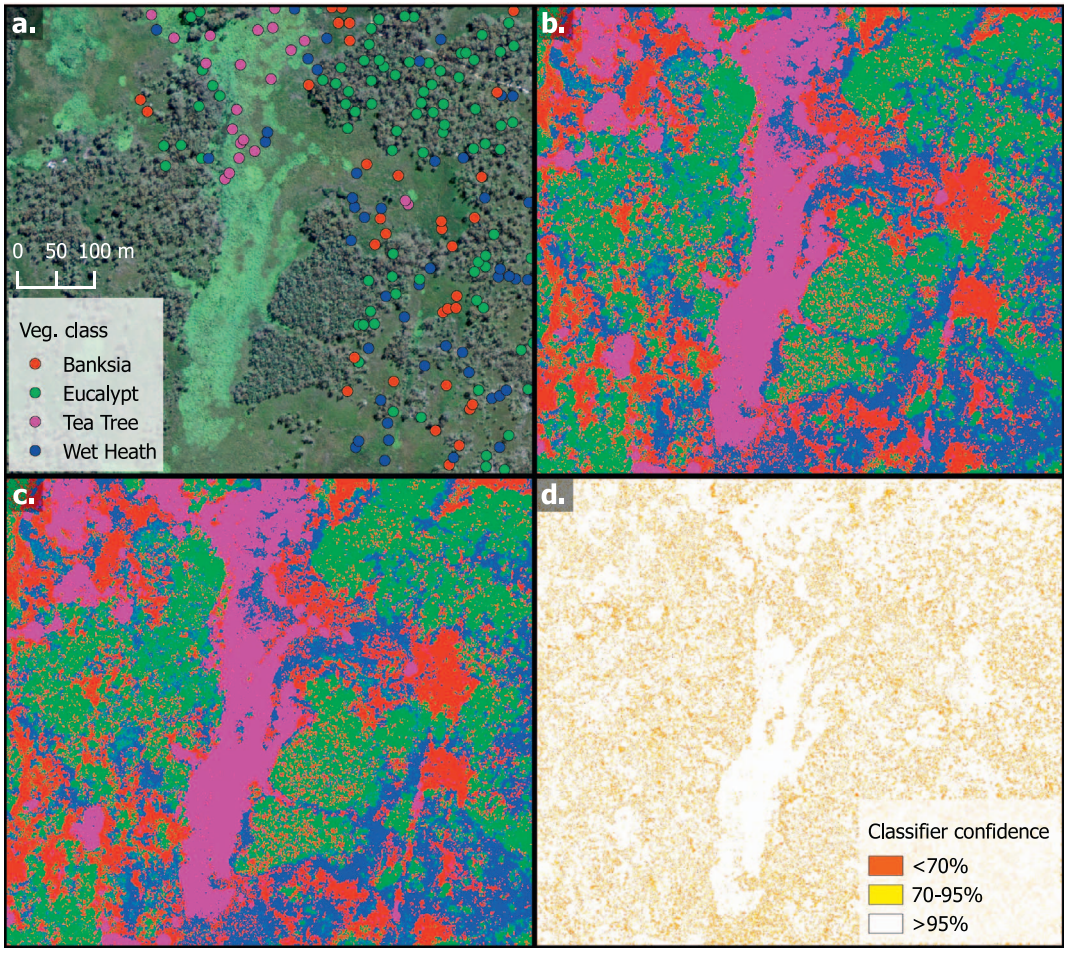

Fig. 4. Mapping example using a resampling framework. **Panel a.** shows the image data and field observations of vegetation classes. **Panel b** shows an example iteration of the resampling and **panel c** shows the mode of the sampling distribution (i.e. most frequent class for each pixel). **Panel d** shows the percentage of the classification distribution that was the same as the mode. The resampling-based accuracy (and 90% interval) was 70 (63–77), and the estimate based on standard error (± 95%) for panel b was 64 (± 5) - see Sections 2.2.6 and 3.4 for resampling parameterisation and results.

perverse consequence of our findings is that we could manipulate the sampling process to produce a overall accuracy statistic anywhere within the discussed ranges, whilst still adhering to various "best practice" sample designs. However, with a transparent resampling design, it is unlikely such a manipulation would be possible, either by accident or on purpose.

Resampling methods have been most commonly employed to quantify accuracy of individual mapping classes of interest (e.g. Brenning 2009; Hsiao and Cheng 2016); individual class metrics were enlightening for our data too. Variance in user and producer accuracies and class area estimates suggested that the Tea Tree Thicket class was composed of some sites that were easily mapped as well as sites that had little relationship to the image data (Fig. 3, Figs. S5–6). Inducing stratification alleviated this issue to some degree, since the range of representative spectral signatures is increased. Importantly, any *single* split in the data between training and test might not have indicated this

Table 1
Accuracy and area estimate results for an example mapping scenario (see Section 2.2.6) comparing the resampling approach from this paper to estimates from more traditional accuracy assessment approaches for two selected iterations. Class proportions are based on the population error matrix (sensu Olofsson et al. 2014). 'Standard' estimates were based on standard error from the full error matrix and 'bootstrap' estimates were calculated based on resampling (1000 iterations) of the error matrix.

	Resampling approach	Single-run estimate close to median			Single-run estimate distant to median		
	Median (95% interval)	Pixel/ sample count	Standard estimate (± 95%)	Bootstrap estimate (± 95%)	Pixel/ sample count	Standard estimate (± 95%)	Bootstrap estimate (± 95%)
Overall accuracy	69 (63–76)	58	61 (± 5)	58 (± 5)	69	67 (± 5)	69 (± 8)
Estimated vegetation class proportions (% mapped area)							
Banksia	18 (13 − 22)	29	16 (± 5)	16 (± 5)	27	16 (± 6)	16 (± 5)
Eucalypt	48 (43–53)	31	47 (± 6)	47 (± 2)	31	47 (± 6)	47 (± 1)
Tea tree	16 (13–17)	14	16 (± 2)	16 (± 0)	16	15 (± 2)	15 (± 2)
Wet heath	19 (14–23)	26	22 (± 6)	22 (± 3)	27	22 (± 6)	22 (± 4)

issue with Tea tree thicket. Indeed, the estimates (and confidence intervals) from either the population or sample count error matrices may have indicated that it was a well mapped class. On the other hand, Eucalypt woodland was well mapped regardless of resampling or stratification design. If we were pursuing a map of the vegetation classes, the variance measures (opposed to just the user/producer accuracy values) would be a key tool for refining our methodological choices.

We found that variance in accuracy due to the differing partitioning of data covered intervals greater than the magnitudes of increase or decrease that the literature has often attributed to choices in data (e.g. different sensors), algorithms (e.g. different classification algorithms) and even sampling design. As advocated for some time now (e.g. Brenning 2009), a resampling approach offers a more robust approach for testing whether differences in design or methodological choices are justified. This also includes inference about differences between map accuracies. Despite the attention given to the effect sampling design can have on accuracy estimates (sensu Zhen et al. 2013), resampling frameworks are not commonly utilised, and the accuracy estimates still inform decision making. Indeed, the authors here have routinely used accuracy metrics without considering variance to make methodological choices (Roelfsema et al. 2015), to inform subsequent modelling exercises (Lyons et al. 2015) and to compare between maps (Lyons et al. 2011). Looking to the future, this paper could provide an option for both map producers and users to abandon some of the traditional pitfalls that have vexed the remote sensing community since at least 2002 (Foody 2002). Moreover, we suggest that at least some of these issues might be completely reconciled by using accuracy estimates (and variance) obtained via a resampling approach.

In this paper we had the luxury of a large field data set from which we could draw samples, using the remaining field data as an additional test set to calculate accuracy metrics. Encouragingly, the median values for accuracy calculated on the full field data set matched the median values calculated on the sub-sample test sets (Figs. S3–4). If we consider the median from the larger field data set as a closer approximation of the truth, then we can conclude that even with the much smaller test sets, the resampling procedures provided quite accurate results. Two vegetation classes dominated sample numbers in our designs, so as the test set size increased, the variance of the estimates decreased. If overabundance or rarity of classes were considered to be an issue, a more balanced sampling design could be used or adjusting accuracy metrics by sample or map proportions would also be appropriate (Lyons et al. 2012; Olofsson et al. 2014). Future work also will involve altering the way the sub-sample is drawn the original field data, to test the effect of field data sampling design and issues of spatial dependence structures. Many of our resampling strategies did not mimic the original sampling design, so exploration of this too should be included in future research.

4.2. Considerations when implementing a resampling framework

There are two important questions to consider for implementing resampling: (1) which resampling approach is most fit for purpose and; (2) how should the final map be presented to maximise usability and interpretability?

4.2.1. Choice of resampling and stratification design

For our observations and image data, all resampling approaches gave similar results given adequate replication. Bootstrapping and Monte Carlo cross-validation with a 67:33 split gave consistently smaller variance for all accuracy metrics, but required more iterations to provide an accurate estimate. As noted, the larger test set reduced variance, but more iterations were needed due to the reduced size of the training set. This is consistent with theory (sensu Efron and Tibshirani 1997, Hastie et al. 2009), and both cross-validation approaches have advantages. *k*-fold cross-validation guarantees all data are tested once, whereas Monte Carlo cross-validation and bootstrapping have no such constraint. Conventional wisdom suggests that, for a limited number of iterations, *k*-fold is essentially unbiased but with large variance, whereas Monte Carlo cross-validation is less variable but is biased (Hastie et al. 2009). In practice, it depends heavily on the size of the full dataset, the fraction of data assigned to the training sets, if and how the split is stratified, and how many times the cross-validation is repeated. Our results confirm this idea; the *k*-fold cross-validation gave accurate but more variable estimates at smaller iteration numbers. Regardless of sampling strategy, the number of iterations required should be guided by stabilisation of the resample estimates (e.g. as demonstrated in Fig. S8).

The block hold-out design consistently produced more variable results, for both accuracy and class area, and had slightly poorer predictive performance on average (e.g. overall accuracy medians ~1% less). Unbalanced training and test samples have been one contributor, but this behaviour may also be indicative of spatial dependence. Indeed, indicator variograms suggested a small amount of spatial autocorrelation (Fig. S2). Stratification and the block hold-out designs changed our results, so we know the data are not independent and identically distributed (iid). This can be problematic since stratified designs still assume data are iid across space. If a spatial dependence structure is present, which is often unavoidable for categorical land cover observations, estimated predictive performance and variance can be incorrect (often optimistically). How strong the effect is will depend on the nature of the correlation. There was only minimal correlation in our observations, so it had minimal effect on our results, as evidenced by only a small effect of the spatial stratification/block hold-out designs. However, it may be more serious for other data sets. A block hold-out (or similar) design is the preferable approach to account for this type of correlation structure, as it ensures test sets are far from training sets, but implementation can present challenges (Roberts et al. 2017). The effect of spatial dependence in remote sensing mapping applications will be a key area of future research.

In terms of implementation, resampling approaches do require more work. Programmatic solutions are the most popular; we provide the code for this paper as an example, and there are also existing implementations that may be more user friendly (e.g. 'sperrorest' in R, Brenning 2012).

4.2.2. Presenting the final map and statistics

Guidance on how to present a final map from a resampling approach is lacking in the literature. This is because literature to date has mostly focused on the use of resampling approaches to make or compare methodological choices and estimate error or population parameters for particular map classes (e.g. Brenning 2009; Hsiao and Cheng 2016). We present four options (Table 2) and suggest that the first two are the most broadly applicable. These two are also the common options more generally when considering final predictions within a resampling framework (e.g. Roberts et al. 2017). Fig. 4 shows an example of how to present a final map. It also demonstrates how the accuracy value of one individual iteration can be misleading (see similarity between panels b & c). Either the mean or median (they will be the same for large numbers of iterations; the mean might be more appropriate for low iterations) can be reported as the single 'map accuracy' value, and then users can then decide what to present in terms of variance/uncertainty. Any estimator appropriate to a sample distribution could be used. Here we used the percentile method to construct confidence intervals, but other methods are available for both estimation of the median and construction of confidence intervals, which is an advantage of a resampling approach. One could also take the empirical variance and covariance if desired (e.g. for standard error and more traditional 95% confidence intervals). Of note is that our results suggested confidence intervals can often be asymmetric, which is important as other common approaches necessarily produce symmetric confidence intervals.

Since a map can be created for the whole image for each resampling iteration, prediction uncertainty can also be represented in a spatially continuous manner. This could be a single map of classification

Table 2
Options for presenting the final map when using a resampling procedure.

Final map option	Justification	Drawbacks
(1) Fit model to all available data. Seen in some of the species distribution modelling literature (e.g. Roberts et al. 2017)	Natural, and produces a map based on the most information possible. Assume more data = better results	Error estimates and variance do not directly relate to the final model. "Problematic" samples accounted for in resampling estimates may have undue influence on the final model
(2) Predict the full output map for every resampling iteration and take the mode as the final map. Analogous to taking the mean of the sampling distribution for continuous predictions (e.g. Roberts et al. 2017).	The mean and variance are calculated on the resampling distribution, so the map should be as well	Not using full amount of data available – data may be variable enough that any partitioning loses salient features, and might produce pessimistic performance metrics
(3) Use the individual iteration that most closely matches sample distribution mean/median. Suggested by Champagne et al. (2014)	Choosing a single model is simple and makes interpretation simpler	Same drawbacks as above two options, except even less data are utilised
(4) Geographical distribution of the probability of occurrence (i.e. percentage of iterations, with confidence intervals if desired) for each class. As presented in Lyons et al. (2017)	Hard classifications hide uncertainty, so viewing each class individually gives a more realistic picture of distribution	Difficult to interpret multiple maps, and many applications require a hard interpretation anyway

uncertainty (as in Fig. 4, panel d), or uncertainty and confidence intervals for each individual mapping class (option 4 in Table 2). This would provide an additional source of information that could inform about classifier performance into new data, as well as feed into decision-making processes that have capacity to incorporate spatially variable uncertainty.

5. Conclusions

In this paper we compared a range of resampling methodologies for performing and assessing image classification. We found that all resampling approaches provided accurate estimates of mapping accuracy, and provided a robust procedure for generating useful and interpretable confidence intervals. There was a large variance in the estimates for any given iteration using a single hold-out test set, and these estimates were often misleading. Existing approaches for confidence interval estimation often failed to account for this variance. We have shown that a resampling approach also provides a robust method for estimating class areas and can easily accommodate any procedure for calculating mapped area. Resampling approaches also enable maps to be presented such that accuracy estimates are explicitly linked to the mapped pixel estimates, including spatially continuous estimates of mapping uncertainty. This paper has provided a number of resampling options that users can adopt for their own mapping activities, along with implementation options in the open source programming language R.

Acknowledgments

The authors would like to acknowledge funding and support from an Australian Research Council Linkage grant (LP150100972) and Centre of Excellence (CE11001000104), and the Long Term Ecological Research Network, as well as the efforts of contributing vegetation scientists and some insightful comments from David Warton. Three anonymous reviewers also provided substantial and valuable insight.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2018.02.026.

References

Brenning, A., 2009. Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection. Remote Sens. Environ. 113 (1), 239–247.
Brenning, A., 2012. In: Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package sperrorest. Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International. IEEE, pp. 5372–5375 (July).
Burgman, M.A., Lindenmayer, D.B., Elith, J., 2005. Managing landscapes for conservation under uncertainty. Ecology 86 (8), 2007–2017.
Champagne, C., McNairn, H., Daneshfar, B., Shang, J., 2014. A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada. Int. J. Appl. Earth Obs. Geoinf. 29, 44–52.
Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560.
Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80 (1), 185–201.
Foody, G.M., 2004. Thematic map comparison. Photogramm. Eng. Remote. Sens. 70 (5), 627–633.
Foody, G.M., 2015. Valuing map validation: the need for rigorous land cover map accuracy assessment in economic valuations of ecosystem services. Ecol. Econ. 111, 23–28.
Gallaun, H., Steinegger, M., Wack, R., Schardt, M., Kornberger, B., Schmitt, U., 2015. Remote sensing based two-stage sampling for accuracy assessment and area estimation of land cover changes. Remote Sens. 7 (9), 11992–12008.
Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., ... Martin, T.G., 2013. Predicting species distributions for conservation decisions. Ecol. Lett. 16 (12), 1424–1435.
Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2nd edn. Springer-Verlag, New York.
Hess, G.R., Bay, J.M., 1997. Generating confidence intervals for composition-based landscape indexes. Landsc. Ecol. 12 (5), 309–320.
Hsiao, L.H., Cheng, K.S., 2016. Assessing uncertainty in LULC classification accuracy by using bootstrap resampling. Remote Sens. 8 (9), 705.
Lyons, M., Phinn, S., Roelfsema, C., 2011. Integrating Quickbird multi-spectral satellite and field data: mapping bathymetry, seagrass cover, seagrass species and change in Moreton Bay, Australia in 2004 and 2007. Remote Sens. 3, 42–64.
Lyons, M.B., Phinn, S.R., Roelfsema, C.M., 2012. Long term land cover and seagrass mapping using Landsat and object-based image analysis from 1972 to 2010 in the coastal environment of South East Queensland, Australia. ISPRS J. Photogramm. Remote Sens. 71, 34–46.
Lyons, M., Roelfsema, C., Kovacs, E., Samper-Villarreal, J., Saunders, M., Maxwell, P., Phinn, S., 2015. Rapid monitoring of seagrass biomass using a simple linear modelling approach, in the field and from space. Mar. Ecol. Prog. Ser. 530, 1–14.
Lyons, M., Foster, S., Keith, D., 2017. Simultaneous vegetation classification and mapping at large spatial scales. J. Biogeogr. http://dx.doi.org/10.1111/jbi.13088.
Mason, T.J., Keith, D.A., Letten, A.D., 2017. Detecting state changes for ecosystem conservation with long-term monitoring of species composition. Ecol. Appl. 27 (2), 458–468.
McKenzie, D.P., Mackinnon, A.J., Péladeau, N., Onghena, P., Bruce, P.C., Clarke, D.M., ... McGorry, P.D., 1996. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? J. Psychiatr. Res. 30 (6), 483–492.
McRoberts, R.E., 2014. Post-classification approaches to estimating change in forest area using remotely sensed auxiliary data. Remote Sens. Environ. 151, 149–156.
McRoberts, R.E., Magnussen, S., Tomppo, E.O., Chirici, G., 2011. Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sens. Environ. 115 (12), 3165–3174.
Pontius Jr., R.G., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. 32 (15), 4407–4429.
Nascimento, H.E., Laurance, W.F., 2002. Total aboveground biomass in central Amazonian rainforests: a landscape-scale study. For. Ecol. Manag. 168 (1), 311–321.
Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., ... Warton, D.I., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–925.
Roelfsema, C., Lyons, M., Dunbabin, M., Kovacs, E.M., Phinn, S., 2015. Integrating field survey data with satellite image data to improve shallow water seagrass maps: the role of AUV and snorkeller surveys? Remote Sens. Lett. 6, 135–144.
Stehman, S.V., Wickham, J.D., Wade, T.G., Smith, J.H., 2008. Designing a multi-objective, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the conterminous United States. Photogramm. Eng. Remote. Sens. 74 (12), 1561–1571.
Weber, K.T., Langille, J., 2007. Improving classification accuracy assessments with statistical bootstrap resampling techniques. Sci. Remote. Sens. 44 (3), 237–250.
Zhen, Z., Quackenbush, L.J., Stehman, S.V., Zhang, L., 2013. Impact of training and validation sample selection on classification accuracy and accuracy assessment when using reference polygons in object-based classification. Int. J. Remote Sens. 34 (19), 6914–6930.