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

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.