Remote Sensing of Environment 177 (2016) 89–100

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research

Reza Khatami [a], Giorgos Mountrakis [a,*], Stephen V. Stehman [b]

[a] Department of Environmental Resources Engineering, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, United States
[b] Department of Forest and Natural Resources Management, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, United States

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form 6 January 2016
Accepted 12 February 2016
Available online xxxx

Keywords:
Remote sensing
Machine learning
Environmental monitoring
Land cover mapping
Multi-time/angle imagery
Texture
Indices
Support vector machines
Classification accuracy

ABSTRACT

Classification of remotely sensed imagery for land-cover mapping purposes has attracted significant attention from researchers and practitioners. Numerous studies conducted over several decades have investigated a broad array of input data and classification methods. However, this vast assemblage of research results has not been synthesized to provide coherent guidance on the relative performance of different classification processes for generating land cover products. To address this problem, we completed a statistical meta-analysis of the past 15 years of research on supervised per-pixel image classification published in five high-impact remote sensing journals. The two general factors evaluated were classification algorithms and input data manipulation as these are factors that can be controlled by analysts to improve classification accuracy. The meta-analysis revealed that inclusion of texture information yielded the greatest improvement in overall accuracy of land-cover classification with an average increase of 12.1%. This increase in accuracy can be attributed to the additional spatial context information provided by including texture. Inclusion of ancillary data, multi-angle and time images also provided significant improvement in classification overall accuracy, with 8.5%, 8.0%, and 6.9% of average improvements, respectively. In contrast, other manipulation of spectral information such as index creation (e.g. Normalized Difference Vegetation Index) and feature extraction (e.g. Principal Components Analysis) offered much smaller improvements in accuracy. In terms of classification algorithms, support vector machines achieved the greatest accuracy, followed by neural network methods. The random forest classifier performed considerably better than the traditional decision tree classifier. Maximum likelihood classifiers, often used as benchmarking algorithms, offered low accuracy. Our findings will help guide practitioners to decide which classification to implement and also provide direction to researchers regarding comparative studies that will further solidify our understanding of different classification processes. However, these general guidelines do not preclude an analyst from incorporating personal preferences or considering specific algorithmic benefits that may be pertinent to a particular application.

1. Introduction

Remote sensing science offers a unique environmental monitoring capability that covers extensive geographical areas in a cost efficient manner while capturing irreplaceable information on the Earth's land, atmosphere and oceans. Remote sensing products play an integral role in numerous applications, for example carbon emission monitoring (Birdsey et al., 2013; DeFries et al., 2002; Myneni et al., 2001; Schwalm et al., 2012), forest monitoring (Asner et al., 2006; Gong et al., 2013; Hansen et al., 2008, 2013; Myneni et al., 2007; Potapov et al., 2015; Townshend et al., 2012), medical science and epidemiology studies (Evans et al., 2013; Gilbert et al., 2008; Liu & Weng, 2012; Lobitz et al., 2000) land change detection (Giustarini et al., 2013; Grekousis, Mountrakis, & Kavouras, 2015; Hussain, Chen, Cheng, Wei, & Stanley, 2013; Lambin & Meyfroidt, 2011; Rindfuss, Walsh, Turner, Fox, & Mishra, 2004), natural hazard assessment (Fialko, Sandwell, Simons, & Rosen, 2005; Khatami & Mountrakis, 2012), agriculture and water/wetland monitoring (Alcantara, Kuemmerle, Prishchepov, & Radeloff, 2012; Anderson, Allen, Morse, & Kustas, 2012; Hong et al., 2012; Ogilvie et al., 2015), climate dynamics (Keegan, Albert, McConnell, & Baker, 2014; Knyazikhin et al., 2013; McMenamin, Hadly, & Wright, 2008; Syed, Famiglietti, Chambers, Willis, & Hilburn, 2010), and biodiversity studies (Asner et al., 2009; Mendenhall, Sekercioglu, Brenes, Ehrlich, & Daily, 2011; Nagendra & Gadgil, 1999; Skidmore et al., 2015).

* Corresponding author.
E-mail addresses: sgkhatam@syr.edu (R. Khatami), gmountrakis@esf.edu (G. Mountrakis), svstehma@syr.edu (S.V. Stehman).

http://dx.doi.org/10.1016/j.rse.2016.02.028

Remote sensing image classification is the process that converts remotely sensed imagery to usable products. In this manuscript we focus on classification processes for creating land-cover maps. Land-cover mapping using satellite or airborne imagery has increased exponentially over the past decades, partially due to improved data availability and accessibility (Yu et al., 2014). Land-cover mapping is a complicated process with numerous factors influencing the quality of the final product. An image analyst has to select from a plethora of options including image type, classification algorithm, training/validation data, input features, pre- and post-processing techniques, ancillary data, and target classes. To make these decisions image analysts are typically drawing on their individual experience and expertise as opposed to the collective knowledge of the field.

The remote sensing community has undertaken considerable efforts to improve land-cover map accuracy. The majority of published research papers demonstrate the validity of their suggested improvements by comparing the accuracy of the proposed classification processes with that of an existing process. Due to the considerable work associated with reference data creation and the limited scope of most studies, the accuracy results reported in these studies are commonly limited to single sites with testing performed on reference data from a single image. Such comparisons are too limited to infer general guidelines for selecting a suitable process to produce highly accurate maps (Stehman, 2006). Moreover, in many cases different studies report conflicting results even when comparing similar classification methods, and inferring general recommendations from these individual studies in isolation is challenging. Consequently, questions such as "*Which classification process is the most promising among a set of processes*?" and "*What is the expected improvement in accuracy*?" have not been answered despite extensive work on classification methods. The objective of our research is to synthesize the collective knowledge of the remote sensing community, as represented by results in peer-reviewed journal articles, to identify which classification processes offer the most promising improvements in accuracy of supervised pixel-based land-cover classification. The analysis is focused on two general factors that can be controlled by analysts to improve classification accuracy, classification algorithms and input data manipulation.

Past review articles have provided useful descriptive summaries of methods and procedures of image classification. For example, Lu and Weng (2007) and Weng (2012) discussed details of major image classification approaches and their main steps, classification accuracy improvement techniques and issues affecting classification performance. Cihlar (2000) and Franklin and Wulder (2002) investigated mapping strategies used in large area land-cover classification and related issues such as multi-time/angle and multiple sensors, geometric processing, and radiometric scaling. Smits, Dellepiane, and Schowengerdt (1999) introduced a quality assessment protocol for land-cover mapping in the context of project requirements and economic cost of error. Wilkinson (2005) utilized scatter plots to investigate the relationship between classification accuracy and date of publication, number of classes, dimensions of feature space, spatial resolution, size of study area and neural network classifiers. Yu et al. (2014) also used scatter plots to investigate the relationship between classification accuracy and date of publication, size of study area, and classification system complexity and report the estimated average overall accuracy and its corresponding standard error for different sensors and classification algorithms. These previous review studies focused on descriptive results and direct comparisons were not targeted to attribute accuracy improvements to individual features of the classification process. A distinguishing feature of the meta-analysis that we implemented is that we synthesized the results of those studies that provided direct one-to-one comparisons of different classification processes. Consequently, our matched-pairs analysis controls for potential confounding factors such as different sites, legends, landscape complexities and reference data that would complicate our ability to fairly compare performance of different classification processes.

2. Protocol for selecting sample of articles

The comparisons among classification processes presented in this work were extracted from peer-reviewed articles published between 1998 and 2012 in five high-impact remote sensing journals: Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing, International Journal of Remote Sensing, and Photogrammetric Engineering and Remote Sensing. To identify recent findings while keeping the workload at manageable levels, our search focused on articles within the fifteen-year period, from 1998 to 2012. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009) was followed for article selection. Fig. 1 describes the selection process (see appendix Table S1 for a detailed PRISMA statement). The following criteria were applied to select relevant articles:

(1) Articles were limited to land-cover mapping. Articles using non-spatial images, images not covering the earth's surface, or simulated data were not included.
(2) Only articles containing supervised per-pixel classification techniques were included. "Soft" classification techniques, where land-cover proportions were estimated for each pixel, with hardened results (i.e. assigning a single label to each pixel) were included.
(3) Articles were required to contain two or more classification processes of the same image(s) using the same training dataset where the only differentiating factor was that either two different classification algorithms were used or an input data enhancement method was added to the first classification process. This allowed isolation of any effect in overall accuracy to a single contributing factor. This is a very important point that was critical in article selection process. Differences in classification tasks including different sites or images, target classes, landscape complexities, and reference data can affect performance of the classification processes. If the classifiers were not applied to the same case study it would be difficult to determine if the differences in overall accuracies were due to the performance of the classification processes or because they had been applied to two different classification scenarios with different levels of difficulty. Consequently, the selected articles and analyses were limited to those studies that compared two or more classification processes based on the same case study (i.e., same image(s), training and test data, and target classes).
(4) Articles included a quantitative accuracy assessment that reported overall accuracy (OA). OA was selected over other accuracy measures because it is most frequently reported and thus would result in a larger sample size of articles.
(5) Accuracy assessment results were based on reference data that were independent of data used in the training phase of the classification.
(6) Accuracy assessment results were based on per-pixel comparisons between the map labels and the reference labels.
The Scopus database reported 15,913 articles published by the five aforementioned journals over the 1998–2012 year period. An automated general query was designed to remove most of the unrelated articles and to extract articles that were more likely to satisfy the selection criteria. Multiple queries were tested by trial and error on some randomly sampled journal issues. Queries were applied on article title, abstract, and keywords. Recall of queries and number of articles returned were considered to determine the appropriate query. Recall was defined as the percent of the articles that could be used in the research returned by the query. The final query was as shown in Table 1 where "OR" operator was applied among expressions inside each column.

The automated query scanned the 15,913 articles and returned 2410 articles. These 2410 articles were then manually examined and 266