Remote Sensing of Environment 137 (2013) 112–123

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

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

Balancing misclassification errors of land cover classification maps using support vector machines and Landsat imagery in the Maipo river basin (Central Chile, 1975–2010)

Olga Lucia Puertas [a,b,*], Alexander Brenning [c], Francisco Javier Meza [a,b]

[a] Centro Interdisciplinario de Cambio Global, Pontificia Universidad Católica de Chile, Santiago, Chile
[b] Departamento de Ecosistemas y Medio Ambiente, Facultad de Agronomía e Ingeniería Forestal, Pontificia Universidad Católica de Chile, Santiago, Chile
[c] Department of Geography and Environmental Management, University of Waterloo, Waterloo, Ontario, Canada

ARTICLE INFO

Article history:
Received 31 January 2013
Received in revised form 16 May 2013
Accepted 1 June 2013
Available online 17 July 2013

Keywords:
Land cover maps
Unbiased area estimates
Mapping methods
Bias removal
Classifier adjustment

ABSTRACT

The ability to carry out land cover change analyses based on Land Use/Cover Classification (LUCC) maps from remote sensing data depends on the quality of the mapping method. Land cover areas obtained from unadjusted classifiers with unbalanced misclassification between different classes could result in erroneously identifying trends. The aim of our work is to describe a novel approach to obtaining LUCC maps with balanced misclassification errors and therefore unbiased predicted areas for each class. We achieve this by numerically minimizing the differences between area proportions obtained with unbiased statistical reference estimates, which is measured by a quantity we refer to as the Sum of Squared Class Unbalancedness (SSCU). We assess the proposed methods in the context of land cover classification with support vector machine classifiers at four points in time between 1975 and 2010 in the Maipo river basin (Central Chile) based on Landsat imagery. In this study, the optimization reduced the $SSCU$ (θ) by 94% on average compared to unadjusted classification. The classifier adjustment also slightly increased the accuracy of the resulting LUCC maps. The amount of bias in classified land cover area and the degree of unbalancedness of misclassification errors differed among the land cover classes. Agricultural land showed the largest reduction in mean relative differences from 27% to 2% compared to the unbiased statistical area estimates. The greatest increase in User's Accuracy was obtained for urban land cover in 1999, where an increase from 56% to 85% was achieved. Qualitative improvements in the classification were visible in difficult classification areas such as dry floodplains. The proposed method is especially recommended for studies that aim to provide multitemporal comparisons.

1. Introduction

Land cover and land use change (LCLUC) have been recognized as key elements of global environmental change (Manandhar, Odeh, & Pontius, 2010). Land use/cover classification (LUCC) maps are important information that provides evidence of spatial LCLUC dynamics and intensity. The ability to carry out LCLUC analyses based on LUCC maps from remote sensing data depends on the quality of the mapping method, in particular the sampled reference data and the classification algorithm applied. Additionally, map accuracy quantification is required to assess the utility of a map for an application (Stehman, 2001), such as detection of change trends (Bakr, Weindorf, Bahnassy, Marei, & El-badawi, 2010; Yuan, Sawaya, Loeffelholz, & Bauer, 2005) or spatial land use change modeling (Arsanjani, Helbich, Kainz, & Darvishi, 2013; Pontius et al., 2007).

Several statistical methods have been implemented and adapted to generate estimates of areas of land cover or land cover change. The objective of LUCC is not only to represent the area of land-cover classes in a map, but to estimate the geographic area of each class. Two common approaches are the use of a confusion matrix to adjust the area derived from pixel counting and the use of a survey sampling regression estimator that takes advantage of auxiliary variables to improve precision of the estimated area (Stehman, 2009). Remote sensing can be used in a variety of ways as support to area frame surveys, for example for stratification or to define sampling units (Gallego, 2004). Czaplewski and Catts (1992), Gallego (2005) and Gallego and Stibig (in press) suggested the use of stratified sampling designs for unbiased estimators. In these cases, the sampling probability is proportional to the area of the sampling unit within each stratum, and a weighted average can be used for area estimation. Subsequently, when LUCC mapping is required, it is possible to

* Corresponding author at: Pontificia Universidad Católica de Chile, Santiago, Chile. Tel./fax: +56 2 3541549.
E-mail address: olpuerta@uc.cl (O.L. Puertas).

0034-4257/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.rse.2013.06.003

apply several classification algorithms considering remotely-sensed and geographic attributes to obtain a spatial representation of land use/cover. Among the currently more widely used classification methods are linear and quadratic discriminant analysis, the support vector machine (SVM) and random forests (Otukei & Blaschke, 2010). Among these techniques, the SVM has proven to be a promising algorithm for LUCC studies (Mountrakis, Im, & Ogole, 2011).

Calibration of the classifier and validation of the predictions are required steps before using LUCC maps for research or to make policy recommendations. In an accuracy assessment process, data quality is quantified, so users can be able to evaluate the utility of a thematic map according to their specific objectives (Stehman, 2001). Primary components of an accuracy assessment are sampling design and reference data collection; these are used to select and obtain reference land-cover data sets. Finally accuracy parameters are calculated (Stehman & Czaplewski, 1998). Traditionally, research on the estimation of accuracy parameters has been focused on generating statistically valid estimates of accuracy rates and describing misclassification errors (Hammond & Verbyla, 1996; Nusser & Klaas, 2003; Stehman & Czaplewski, 1998; Stehman, Wickham, Smith, & Yang, 2003). From these primary studies, the confusion or error matrix is at the core of classification accuracy assessment and widely used in the domain of remote sensing (Congalton & Green, 2009). The main condition for the use of a confusion matrix is the correct application and interpretation required for the satisfaction of often untenable assumptions (e.g., perfect coregistration of data sets) and the provision of rarely conveyed information (e.g., sampling design) (Foody, 2002).

There are advantages when both land use/cover estimates of areas and classifier algorithm procedures are linked. Through the connection of these methodologies, land use/cover areas can be estimated by the unbiased procedure and evaluated according to the accuracy estimators from a probability sampling design. For instance, accuracy comparisons have been made to compare different sampling designs (e.g. Stehman et al., 2003), LUCC algorithms (e.g., Brenning, 2009; Brenning, Kaden, & Itzerott, 2006; Petropoulos, Kontoes, & Keramitsoglou, 2012) or the relevance of auxiliary variables in the LUCC model (e.g. Xie, Lin, & Ren, 2011). However, the issue of unbalanced misclassification errors in LUCC mapping has received little attention and requires further research. One possible approach is to apply optimization techniques that tune classifiers in order to ensure that the resulting thematic maps achieve balanced commission/omission errors and hence unbiased land cover areas. We therefore propose to identify classifier parameters that minimize a measure of unbalancedness of the confusion matrix, and to apply these optimized parameters to obtain LUCC maps with nearly balanced class proportions for each land cover class.

The aim of our work is to generate spatially explicit information of balanced land cover by development of algorithms balancing commission and omission errors in every class of LUCC maps. In this paper, we describe a methodology to obtain balanced LUCC maps based on the adjustment of predicted probabilities for each land cover class by an optimization process that reduces differences between land cover areas obtained by unbiased statistical estimation and by classifier prediction. We then illustrate the application of this method to SVM classification in a case study from Central Chile with several land cover classes of different proportions.

2. Data and methods

Land cover classification with balanced commission/omission errors was performed by applying pixel-based SVM classifiers using multispectral remote sensing and topographic data as prediction variables, and using the proposed method to adjust the SVM classifier based on unbiased statistical estimates of land cover areas. The classification and adjustment were performed independently for four time points between 1975 and 2010. An overview of the main processes applied is shown in Fig. 1, while a brief description of these steps is provided in the following sections. After introducing the collection of the reference land-cover dataset (Section 2.1) and outlining the area estimation procedure (Section 2.2), we describe the general SVM classification algorithm (Section 2.3) and the proposed procedure used for achieving balanced commission and omission errors of land cover maps (Section 2.4).

2.1. Reference land-cover data collection

We adopted a stratified sampling design to handle classes with small coverage (e.g., water and wetlands). We calculated the required sample size for each sampling stratum using the approach for multinomial proportions of Thompson (1987). In this method, which considers a worst-case scenario, the number of samples does not depend on the number categories of the population (Thompson, 2000). The main advantage is that the method is less conservative and does not require prior knowledge about largest land cover class size as compared to other approaches (Congalton & Green, 2009). Thus, Thompson's sample size rule depends only on two parameters for each stratum s: a confidence level α and the precision d_s. We chose the $\alpha = 10\%$ confidence level and d values between 0.05 and 0.10 as shown in Table 1 and explained below. With these parameters, the required sample size n_s for a stratum s can be determined from the tabulated $d_s^2\, n_s$ values of Thompson (1987). The resulting number of sample pixels was then randomly selected from the area corresponding to each stratum.

Our objective was to identify six land cover classes ($c = 1,\ldots,6$): *snow cover*, *water bodies*, *urban*, *agricultural*, *bare soil*, and *vegetation* (shrublands and grasslands). We stratified with respect to two parameters, slope and normalized difference vegetation index (NDVI; Rouse and Space (1973)). NDVI values were obtained from the visible (Red) and near infrared (NIR) reflectance as NDVI = (NIR − Red) / (NIR + Red). For stratification we used NDVI values calculated from Landsat 5 TM images of 1987, which showed higher quality due to low cloud cover, and does not suffer from gaps such as those resulting from the failure of the Scan Line Corrector (SLC) of Landsat 7 ETM+ in 2003.

For the stratification step, two slope and three NDVI classes were combined, resulting in six sampling strata (Table 1). We used a precision of $d = 0.1$ for all but one stratum and $d = 0.05$ in stratum E22. E22 (48.4% of the area) was given a better precision because of its larger contribution to the overall precision and because of the heterogeneity of this stratum. The same precision of 0.1 was used for each of the smaller strata regardless of stratum size in order to obtain precise area estimates for land cover classes such as *water bodies* that are prevalent only in small strata such as E11. The total sample size was 908 pixels, covering 14,833 km^2 (Table 1).

Reference sample points were classified by visual image interpretation based on three-band false-color composites of Landsat MSS, TM and ETM+ spectral images. To better distinguish the different land cover classes we used different three-band composites based on the available bands from B2 to B5 in the case of MSS, and from B1 to B5 and B7 for TM and ETM+.

2.2. Reference condition: land cover area estimation

Land cover area $\hat{P}_c$ estimates for the area proportion of P_c each reference land cover class c were estimated using standard procedures for stratified sampling designs (Thompson, 2000):

$$\hat{P}_c = \sum_{s=1}^{k} \frac{n_{cs}}{n_s} p_s \qquad (1)$$

where n_{cs} is the number of samples in class c within stratum s, n_s is the total sample size for this stratum, and P_s is the area proportion