Long-term land cover dynamics by multi-temporal classification across the Landsat-5 record

Joseph O. Sexton a,b,⁎, Dean L. Urban a, Michael J. Donohue a, Conghe Song c

a Nicholas School of the Environment, Duke University, Durham, NC, United States
b Global Land Cover Facility, Department of Geographical Sciences, University of Maryland, College Park, MD, United States
c Department of Geography, University of North Carolina, Chapel Hill, NC, United States

A R T I C L E   I N F O

Article history:
Received 5 August 2009
Received in revised form 12 October 2012
Accepted 13 October 2012
Available online 10 November 2012

Keywords:
Landcover
Landsat
Multi-temporal
Land cover change
Supervised classification
Validation
Land cover dynamics
Sampling bias

A B S T R A C T

Supervised classification of land cover across space and time is a long-standing goal of the Earth Science community. Although most past and current analyses focus on detecting changes between two or more times, the opening of the USGS Landsat archive in 2009 has enabled exploration of methods for higher-frequency, time-serial monitoring of land cover dynamics. Modifying the protocols used to develop the 2001 National Land Cover Database (NLCD 2001), we fit a single classifier to a spatio-temporally distributed reference sample and applied the model to 55 Landsat-5 images covering a section of the North Carolina Piedmont Plateau from 1984 to 2007. A generalized classification scheme, multi-temporal sampling design, supervised classification based on intra-annual spectral indices, and design-based accuracy assessment yielded a time-series of 16 land cover maps from 1985 to 2006 with a spatial extent of 1.7 × 106 ha, minimum mapping unit of 1 ha, and mean temporal interval of ~2 years. Comparable to accuracy of the NLCD 2001 Land Cover Layer for the region, overall accuracy for a spatio-temporally independent test sample was 75%, with κ = 0.7. When weighted by class proportions, percent correctly classified and kappa rose to 88% and 0.84, respectively. The resulting map series shows spatially and temporally complex changes in water, urban, forest, and herbaceous cover resulting from natural and anthropogenic processes that would not be observable in either uni- or bi-temporal maps. Agricultural crop area dropped from ~45% in the 1980s to ~36% in the 1990s and then rose slightly to ~38% at the end of the period. Forest area increased to a maximum of ~55% in the 1990s and then dropped to ~53% in 2005. Urban growth appeared to be most rapid in the 1980s and 1990s and slowed thereafter. With continued focus on the semantics, causation, sampling, and uncertainty underlying spectral land cover classification, long-term series of Landsat images will provide increasingly robust, reliable records for a growing scientific user community. These multi-temporal datasets will be indispensable for understanding past land cover dynamics and predicting the implications of future change on the provision and management of ecosystem services.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

The Landsat satellites have imaged Earth’s terrestrial and near-shore marine surfaces for over thirty years. These images support numerous scientific applications, providing assessments of ecosystem quality, resources, services, and changes due to natural and anthropogenic forcing (National Academy of Science, 2005). Such a long and consistent record allows scientists to study current phenomena, relating satellite data to ground-truth or other references before interpreting patterns, but also to retrieve information from the past, over periods for which little or no reference data are available.

Among the most widely used, the 2001 National Land Cover Database (“NLCD 2001”: Homer et al., 2004) consists of fractional tree canopy and impervious surface cover and categorical land-cover/land-use layers covering the United States. Datasets such as NLCD 2001 are increasingly accurate and consistent sources of information for each dataset’s reference year. However, no such data are available spanning multiple years over the long term. To capture the complexities of Earth’s changing surface, time-serial landcover maps are needed.

Monitoring changes over time requires consistency among maps, both ontologically and statistically. That is, the maps must have identical semantics between real-world ecosystem types and thematic classes; and, unless accuracy is extremely high, they must also be based on similar correlations to the image data on which they are based. Without these two properties, it is impossible to determine whether changes observed are due to actual landscape changes or to ontological and statistical artifacts from the underlying data model.
Comparing inconsistent maps leads to unreliable conclusions. Consider the changing proportions of four land cover classes over time (Fig. 1) for a 1.7 Mha region of North Carolina (Fig. 2). The data were obtained from three publicly available and widely used maps, and the 4-class scheme was generalized from the maps’ original schema to maximize comparability. In this depiction, land cover proportions fluctuated erratically, with impossible rates and reversals of change within a decade. Clearly more representative of underlying model differences than of real changes, these measurements would not be useful for further scientific analyses or policy decisions.

A single, robust model applied consistently over time could avoid these ambiguities. As the basis for land cover classification and change detection over space, this “signature extension” was an active topic of research early in the history of Landsat (e.g., Minter, 1978), and remains so today (Oltbrof et al., 2005; Woodcock et al., 2001). The process of signature extension requires training a classification model of the spectral signature on a relatively small number of land cover observations matched with spectral measurements in time and space and then extrapolating the classifier on image data collected at other times and/or locations for which no such reference is available (Botkin et al., 1984; Cihlar, 2000; Jensen, 1983; Muller, 1988). Whether in space or in time, signature extension relies on invariance of the relationship between cover and its spectral signature—a consistency that can be compromised by numerous factors, including atmospheric contamination, the interaction of illumination and viewing angles with the bidirectional reflectance distribution function of the surface, and plant phenology (Song & Woodcock, 2003). Because images from many dates must often be composited into large regional mosaics, much research has gone into calibrations to remove this unwanted variation.

All of these effects can be accommodated, but to varying degrees (Huang et al., 2009; Masek et al., 2006). Sensor radiometry is calibrated systematically (Chander & Markham, 2003; Chander et al., 2007), and bidirectional and atmospheric correction can be reasonably accomplished through models of varying complexity (Song et al., 2001). Even phenological variation—which is affected by vegetation type and environmental characteristics (Rathcke & Lacey, 1985; Schwartz, 2003)—has been used empirically to interpolate “synthetic” Landsat images between the satellite’s 16-day orbital repeat (Hilker et al., 2009). However, remnant effects from both modeled and unmodeled...
factors continue to contribute significant noise to models of land cover signatures (Moisen et al., 2000).

Despite these difficulties, monitoring landscape change is possible through the satellite record. To date, success has been greatest in ecosystem systems defined by a small number of types. Pax-Lenney et al. (2001) devised an innovative generalization approach to extend spectral signatures of conifer forest across time and space based on fuzzy ARTMAP neural networks; Masek et al. (2008) used relative changes in the Disturbance Index (DI) (Healey et al., 2005) to monitor continuous, sub-pixel forest disturbance and recovery across North America; and the Vegetation Change Tracker of Huang et al. (2010) automated forest cover monitoring using dense time series of Landsat images.

These studies have focused solely on forest changes, but potential for multi-temporal monitoring has been demonstrated in multiple-class land cover systems as well. Latifovic and Pouliot (2005) generated a 1-km resolution land cover series for Canada at 5-year increments by updating a single reference map through change detection and local heuristics, the National Land Cover Database 2006 Fry et al. (2011) quantified changes between 2001 and 2006 by following change detection with unsupervised classification, and Fraser et al. (2009) extended this approach to Landsat data with empirical normalization to a “master” image and temporal extension of spectral signatures derived from unsupervised classification.

The National Land Cover Database is the most popular land cover data product in use within the United States, although spatially coarser products at annual frequency are used for broad-scale assessments. The NLCD 2001 and 2006 share a similar classification scheme, but the 1992 and 2001 NLCD were based on different classification schema and models, severely limiting their comparability, and the land cover change product derived from the NLCD 1992 and 2001 layers requires “retrofitting” to cross-reference the classifications (MRLC, 2008). The MODIS land cover products (Friedl et al., 2002; Hansen et al., 2005), are each based on a consistent model but are intended for global assessments, lacking either the temporal extent or spatial resolution of the Landsat mission.

2. Objectives

Adequately capturing the spatio-temporal dynamics of human-impacted ecosystems requires consistent land-cover map series that combine the time span and spatial resolution of the Landsat mission with ~5-yr temporal interval (Lunetta et al., 2004). We present a signature extension approach for dense time-series of Landsat Thematic Mapper images that is based on a single supervised classification trained over a spatially and temporally distributed reference sample. For a study area typical of north-temperate land cover dynamics, we focus on a portion of the Piedmont Plateau in North Carolina, USA, where changing land-use practices are resulting in a complex exchange of forested, agricultural, and urban cover. We report the image processing, sampling, and signature-modeling methods used to maximize the classification’s robustness to phenological and other sources of image-to-image variation. From standard probabilistic validation statistics, we derive methods for removing sampling bias to allow comparisons of accuracy across multiple independent test samples. We conclude by discussing the general requirements and some potential avenues for further improvement to multi-temporal land cover classification and analysis.

3. Methods

3.1. Study area

Extending from New York City to central Alabama, the Piedmont Plateau is America’s most populous region. On the North Carolina Piedmont, extensive pine forests established in farm fields abandoned early in the twentieth century are succeeding naturally to dominance by hardwood forests, but are also rapidly being converted to suburbs around metropolitan centers (McDonald & Urban, 2006; Taverna et al., 2005). Piedmont soils are spatially complex, but the region’s topography is characterized by gentle relief, with broad river valleys draining extensive uplands. Phenologically, the area’s agricultural fields and pastures begin vernation in middle- to late March, with the deciduous forests following throughout the month of April. Responding variously to drought, storms, day-length, temperature (and ultimately harvest for herbaceous crops), senescence ranges from early September to early December.

The 1.67-Mha study area is defined by the intersection of the World Reference System 2 (WRS-2) Path 16/Row 35 scene and the North Carolina counties lying completely on the Piedmont Plateau (Fig. 2). From east to west, the study area extends from the Triangle region (framed by the cities of Raleigh, Durham, and Chapel Hill) to the Triad region (framed by Greensboro, High Point, and Winston-Salem). The area’s combination of: (1) rapid and complex changes in recent years, (2) high spatial edaphic and hydrological variability, and (3) low topographic relief makes the region ideal for studying mixed natural and anthropogenic land cover dynamics.

3.2. Data

3.2.1. Landsat

Fifty-five Landsat-5 Thematic Mapper (TM) images (Table 1) from the WRS-2 Path 16, Row 35 scene were acquired and geometrically rectified to 1-m resolution digital ortho-photographs by 1st- or 2nd-order polynomial transformation. Spatial misregistration was less than 20 m RMSE for most images, and every image had RMSE ~25 m. Clouds and their shadows, snow, and radiometrically “bad” pixels were identified visually and removed by on-screen digitizing. The reflective bands of the screened images were then converted to radiance following Chander et al. (2007) and atmospherically corrected to surface reflectance by the DOS3 dark-object subtraction method (Song et al., 2001).

The surface reflectance images were masked to the study area extent and gathered into composite datasets for training, extrapolating, and validating the classification. Following Homer et al. (2004), images were divided into seasonal triplets: early growing season ("spring") (April 24–July 2), late growing season ("fall") (September 3–October 22), and dormant season ("winter") (November 21–February 2). Tasseled-Cap "brightness" (B) and "wetness" (W) (Crist, 1985; Crist & Cicone, 1984; Kauth & Thomas, 1976) and the Normalized Difference

<table>
<thead>
<tr>
<th>Year</th>
<th>Early growing season</th>
<th>Late growing season</th>
<th>Dormant season</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>July 1, 1984</td>
<td>September 19, 1984</td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td>May 1, 1985</td>
<td>September 6, 1985</td>
<td>January 9, 1985</td>
</tr>
<tr>
<td>1993</td>
<td>May 7, 1993</td>
<td>September 28, 1993</td>
<td>December 1, 1993</td>
</tr>
<tr>
<td>1994</td>
<td>April 24, 1994</td>
<td>September 15, 1994</td>
<td>December 12, 1993</td>
</tr>
<tr>
<td>1995</td>
<td>June 14, 1995</td>
<td>September 15, 1994</td>
<td>February 6, 1995</td>
</tr>
<tr>
<td>1999</td>
<td>June 9, 1999</td>
<td>October 15, 1999</td>
<td>December 2, 1999</td>
</tr>
<tr>
<td>2004</td>
<td>May 21, 2004</td>
<td>September 24, 2004</td>
<td>January 1, 2004</td>
</tr>
<tr>
<td>2006</td>
<td>April 25, 2006</td>
<td>October 13, 2006</td>
<td>December 5, 2006</td>
</tr>
<tr>
<td>2007</td>
<td>May 14, 2007</td>
<td>August 18, 2007</td>
<td></td>
</tr>
</tbody>
</table>
Vegetation Index (NDVI) (Rouse et al., 1973) were computed for each image. When more than one acceptable image was available for any season within a year, surface reflectances were estimated for all available dates and averaged within the season before calculating indices. Idiosyncrasies of phenology and image acquisition date frequently cause confusion of herbaceous types with urban and forest classes (Seto et al., 2002; Yuan et al., 2005). To minimize errors due to phenological variation within the growing season, spring and fall brightness and NDVI were combined into growing-season averages. Mean and maximum annual wetness were likewise computed from each annual triplet’s Tasseled-Cap wetnesses to minimize the impact of individual precipitation events. The final dataset thus consisted of four seasonal and two annual metrics: mean growing-season (i.e., “summer”) brightness and NDVI ($B_{\text{sum}}$ and $\text{NDVI}_{\text{swavg}}$), winter brightness and NDVI ($B_{\text{w}}$ and $\text{NDVI}_{\text{w}}$), and annual mean and maximum wetness ($W_{\text{wavg}}$ and $W_{\text{wmax}}$):

$$B_{\text{swavg}} = \frac{B_{\text{spring}} + B_{\text{fall}}}{2}.$$  

$$\text{NDVI}_{\text{swavg}} = \frac{\text{NDVI}_{\text{spring}} + \text{NDVI}_{\text{fall}}}{2}.$$  

$$W_{\text{wavg}} = \frac{W_{\text{spring}} + W_{\text{fall}} + W_{\text{winter}}}{3}.$$  

$$W_{\text{wmax}} = \max(W_{\text{spring}}, W_{\text{fall}}, W_{\text{winter}}).$$

### 3.2.2. Landcover

A 13-class scheme was adapted from the NLCD 2001 Land Cover Layer (Table 2). The NLCD “grassland” class (71) was disregarded because natural grasslands are an extreme rarity in this region. The “developed open space” class (21) was also disregarded because its identification is contextual, requiring municipal boundaries, the changes of which are not available over the entire study span. The NLCD “shrubland” class (52) was simply interpreted as short, young forests, or “scrub.”

Reference land cover data were identified by a single observer, using mosaics of ortho-rectified high-resolution digital images taken over Guilford County in 1995, Randolph County in 2004, and Wake and Granville Counties in 2005 (Fig. 2). The sample was located spatially by stratified random sampling, using NLCD 2001 for strata. The NLCD 2001 Land Cover Layer was clipped to the boundaries of each high-resolution image mosaic and filtered to exclude spatial heterogeneity of land cover within a 3 x 3-pixel window to avoid misclassifications due to spatial misregistration. A random sample of 150 pixels of each NLCD 2001 class (excluding classes 21 and 71) was then selected from the masked pixels. These points were overlaid on the 1995, 2004, and 2005 high-resolution images and then labeled with their co-incident image’s year of acquisition and the cover type identified in that year. During this final identification phase, points were discarded if their cover type was visibly mixed with other types within a 30-m buffer or could not be determined due to image quality.

The reference sample was then randomly divided into two subsamples for model training and validation. Poor image quality (especially in 1995) and differences between NLCD 2001 and the high-resolution images led to large differences in abundance among the classes, and so allocation of data between the training and validation samples was determined by overall abundance in the multi-year sample (Table 3). Classes that were abundant in the sample were represented by approximately 150 observations in the training subsample, those with intermediate abundance by ~100 observations, and rare classes by ~50 observations. The remaining observations from each class were allocated to the test sample, to which 1087 observations made by a second observer over Wake County in 1999 were added. Thus, approximately half of the observations upon which model validation was based were

### Table 2

<table>
<thead>
<tr>
<th>Label</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Water</td>
<td>Open water (100% cover)</td>
</tr>
<tr>
<td>22</td>
<td>Urban, low-density</td>
<td>Mix of built and natural cover, &lt;50% impervious</td>
</tr>
<tr>
<td>23</td>
<td>Urban, medium-density</td>
<td>Mix of built and natural cover, 50–75% impervious</td>
</tr>
<tr>
<td>24</td>
<td>Urban, high-density</td>
<td>Mostly built cover, &gt;75% impervious</td>
</tr>
<tr>
<td>31</td>
<td>Bare</td>
<td>Cleared soil or rock, with &lt;25% vegetation cover</td>
</tr>
<tr>
<td>41</td>
<td>Forest, deciduous</td>
<td>Trees &gt;5 m tall and &gt;50% cover, &lt;25% evergreen</td>
</tr>
<tr>
<td>42</td>
<td>Forest, evergreen</td>
<td>Trees &gt;5 m tall and &gt;50% cover, &gt;75% evergreen</td>
</tr>
<tr>
<td>43</td>
<td>Forest, mixed</td>
<td>Trees &gt;5 m tall and &gt;50% cover, 25–75% evergreen</td>
</tr>
<tr>
<td>52</td>
<td>Scrub</td>
<td>Mix of trees (&lt;5 m tall) and other natural cover, &gt;25% vegetated</td>
</tr>
<tr>
<td>81</td>
<td>Pasture/hay</td>
<td>Fields with &gt;25% persistent herbaceous cover, dominated by non-cereal grasses</td>
</tr>
<tr>
<td>82</td>
<td>Row crops</td>
<td>Fields with &gt;25% herbaceous cover, periodically harvested</td>
</tr>
<tr>
<td>90</td>
<td>Wetland, woody (swamp)</td>
<td>Mix of water, herbaceous plants, and bare soil</td>
</tr>
<tr>
<td>95</td>
<td>Wetland, herbaceous (marsh)</td>
<td>Mix of water, trees, and bare soil</td>
</tr>
</tbody>
</table>

### Table 3

Landcover sample sizes ($n$) and proportions in the reference sample and study area (taken from NLCD 2001), and model priors derived from NLCD 2001 by splitting probabilities for class 71 equally between classes 81 and 82, and class 21 proportionally among non-urban land classes (i.e., excluding classes 11, 23, and 24). Years in bold were used for both model training and testing, whereas 1999 was reserved solely for model testing.
spatially independent from the training data, and another half was unique in terms of space, time, and also observer.

3.3. Analysis

3.3.1. Landcover classification

Prior to classification, separability of classes was assessed using dendrograms of Euclidean distance between class centroids in the domain defined by the seasonal indices. Land cover was then classified by a hybrid of Classification and Regression Trees (CART) and Quadratic Discriminant Analysis (QDA) followed by temporal and spatial filtering, executed in three passes. CART (Breiman et al., 1984) is a nonparametric algorithm that recursively partitions the spectral domain into regions of maximally homogeneous cover, and QDA – the basis for the “Maximum Likelihood Classifier” – estimates the probability of each class across the spectral domain based on Mahalanobis distance from the class centroid (Jensen, 2005). Based on the strengths of each, a hybrid of these two approaches was used to avoid overfitting the training sample and retain probabilistic information further in the classification process. Preliminary analyses showed that CART provided superior land cover extrapolations for phenologically invariant cover types but overfitted the training data in phenologically variable classes. Conversely, QDA provided a less precise classification, but the continuous probability density functions it provided were more robust to image-to-image variation. Further, both approaches can be used to generate probabilistic maps, which are more amenable to adjustment by prior probabilities and temporal filtering than categorical maps of the classes alone.

During the first pass, CART was used to determine a threshold of maximum annual wetness for discriminating water from land, based on the three-year training sample. This threshold was then applied to the image series to create a water mask for each year, assigning the user’s accuracy (UA) calculated for the independent test sample as the probability of water in pixels meeting the criterion and 1-UA as the probability of water for all other pixels (see Appendix A for derivation of all accuracy metrics). In the second pass, QDA was used to estimate spectral signatures and assign probabilities of all classes, including water. Prior probabilities of the classes were obtained from the NLCD spectral signatures and assign probabilities of all classes, including pasture/hay (81) and row crops (82). Classes, and the prior probability of grassland (71) was distributed opened open space) was distributed equally among all remaining land classes. The spatial “Clump” and “Eliminate” operations were then performed in ERDAS IMAGINE on the categorical maps to further reduce speckle and spurious changes due to spatial misregistration. The clump function identifies each pixel with a group of equally valued neighbors, and the eliminate function recodes the land cover of pixels in groups whose area is below a given threshold. Pixels were clumped using an 8-neighbor rule, and clumps with area <1 ha (–11 pixels) were eliminated.

3.3.2. Accuracy assessment

Accuracy assessment was based on confusion matrices, using producer’s accuracy (PA), user’s accuracy (UA), percent correctly classified (PCC), and the Kappa statistic (κ) as validation metrics. Class proportions in the sample were neither uniform nor representative of the study area at any time, so it was necessary to adjust PCC and Kappa estimates to neutralize sampling bias (Byr et al., 1993; Congalton, 1991, Appendix A). To correct this bias, a vector of class weights:

\[ w_i = \frac{p_i^*}{p_i} \]

in which \( p_i \) is the proportion of class \( i \) in the reference sample and \( p_i^* \) is the probability of \( i \) from a distribution chosen a priori as a standard, was applied to calculations of PCC and Kappa. As in model training, \( p_i^* \) was derived from NLCD 2001 (Table 2), although these proportions could be derived from any reliably accurate and consistent land cover map. The standardized Percent Correctly Classified (PCC), and Kappa (κ) metrics were thus calculated from the joint class frequencies \( f_{ij} \) of the confusion matrix as:

\[ PCC_j = \frac{1}{n} \sum_{i,j} w_i f_{ii} \]

and

\[ \kappa_i = \frac{\sum_i w_i f_{ij} - \sum_i w_i f_j}{n^2 - \sum_i w_i f_j} \]

All validations were based on the independent test sample (Table 3).

4. Results

4.1. Land cover classification

Spectral similarities reflected the semantic hierarchy among the classes (Fig. 3). Water (11) was spectrally distinct from the land classes. Forests formed a cohesive group, but the various gradations of deciduous forest - i.e., deciduous forest (41), mixed forest (43), and (predominantly deciduous) swamps (90) - were more similar to scrub (52) and herbaceous marshes (95) than they were to evergreen pine forests (42). This broad forest group was distinct from the urban and field classes, within which high-density urban (24) was spectrally unique. The remaining urban, field, and bare classes were spectrally...
similar to one another, with medium-density urban (23) and bare ground (31) exhibiting the greatest similarity and the field classes – i.e., pasture/hay (81) and row crops (82) – also similar to one another.

Above the optimal threshold of maximum annual wetness estimated by CART for discriminating water from land (0.01588), the probability of water (i.e., user’s accuracy, UA) was 0.97. Among the individual years, user’s accuracies for water ranged from 84% (2004) to 98% (1995). Most water misclassifications by CART were made in forested and herbaceous wetlands, but a small minority was also committed in urban types, likely due to shadows cast by large buildings.

The test sample was classified with PCC = 75% and κ = 0.70 (Table 4). After adjusting for the distribution of classes in the sample, the standardized (i.e., map) accuracy rose to PCC = 88% and κ = 0.84. Water was accurately classified without bias, with producer’s accuracy (PA) and UA both equaling 98%. Within urban types, classification was biased toward lower densities, with high-density urban (24) more often labeled as medium-density (23), which was in turn more often labeled as low-density (22). Low-density urban was under-predicted, mostly due to confusion with herbaceous fields (80). Fields themselves were over-predicted at the expense of all types except water, but were especially greedy of the low-density urban type. Both evergreen and deciduous forests were classified with very high accuracy. The rarest types – bare (31), scrub (52), and marsh (95) – were classified poorly.

4.2. Land cover change

Change was unevenly distributed across the study area, with greatest change near cities (Fig. 4). Development was most pronounced in the Triangle region, where the urban cover of Raleigh and Durham and smaller towns of Cary and Apex expanded predominantly into the surrounding agricultural fields. Urban area also expanded significantly in the Triad, but in comparison to the larger agglomerations of the Triangle and the Triad, the relatively isolated city of Burlington expanded only slightly. Overall, the region exhibited a broad initial conversion of agricultural fields to forests from 1985 to 1995, followed by partial recovery of fields over the next decade; but these rural changes were diffuse, with no discernible spatial pattern.

The general types – forest, field, urban, and water – exchanged between 5 and 20% of their 1985 areas, with greater proportional changes in the rarer classes (Fig. 5). As a proportion of the study area, forest cover increased from ~50 to ~53% over the period, expanding mostly at the expense of herbaceous fields, which decreased from 42 to 46% in the late 1980s to 38–39% around 2005. Urban area increased from 3 to 5% in the beginning of the period to 6–7% at the end of the period. The coverage of water increased from 1.9% of the landscape in 1985 to 2.1% in 2006, with the greatest gain associated with the filling of reservoirs between 1987 and 1989.

Higher-frequency temporal dynamics were blurred by errors in individual years. Three years – 1991, 1993, and 2001 – especially exhibited transient cover changes that were more likely due to phenological or other sources of noise than to true changes. However, misclassifications were not consistent across the outlier years. Urban area increased erratically in both 1991 and 1993 — in 1991 at the expense of both forests and fields but in 1993 at the expense of fields alone. In 2001, fields increased erratically at the expense of forest area, with urban area mostly unaffected.

Ignoring the outlier years, the region’s various land cover changes exhibited slight accelerations or decelerations over time. The area of agricultural fields dropped from its maximum near 45% in the 1980s to a minimum around 36% in the 1990s, then rose again slightly to ~38% at the end of the period. Forest area changed complementarily, increasing from a minimum in the 1980s to a maximum of ~55% in the 1990s and then dropping slightly to ~53% in 2005. Acceleration of urbanization was more difficult to discern due to imprecision in the low-density urban class, but it appears that urban growth was most rapid in the 1980s and 1990s and slowed thereafter.

The effect of classification error is shown by comparison of three classes with different distributions of error (Fig. 6). The imposition of a single set of priors over time resulted in constant predictions of the upper margin of error for each class, but lower margins varied along

---

Table 4
Confusion matrix for landcover classification validated on a 4-year test sample, with producer’s accuracy (PA), user’s accuracy (UA), Kappa (κ), percent correctly classified (PCC) and Kappa (κ), as well as standardized percent correctly classified (PCCs) and standardized Kappa (κs). Standardizations were computed based on priors from NLCD 2001 (Table 2). Values along the matrix diagonal (correct classifications) are in emphasized in bold font.

<table>
<thead>
<tr>
<th>Class</th>
<th>Prediction</th>
<th>11</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>31</th>
<th>41</th>
<th>42</th>
<th>52</th>
<th>80</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>11</td>
<td>317</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>0</td>
<td>97</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>11</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>0</td>
<td>51</td>
<td>31</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>0</td>
<td>7</td>
<td>110</td>
<td>121</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>351</td>
<td>16</td>
<td>2</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>209</td>
<td>0</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>7</td>
<td>46</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>554</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>95</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>UA (%)</td>
<td>98.4</td>
<td>99.2</td>
<td>19.7</td>
<td>94.8</td>
<td>78.0</td>
<td>85.3</td>
<td>95.8</td>
<td>70.7</td>
<td>50.0</td>
<td>77.7</td>
<td>7.7</td>
</tr>
<tr>
<td>PA (%)</td>
<td>PCC = 75.1</td>
<td>PCCs = 88.1</td>
<td>κ = 0.70</td>
<td>κs = 0.85</td>
<td>a = 2301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 4. Regional cover of the dominant landcover classes for three selected years, with detail of the Triangle Region (inset). (A) Reservoirs built around the city of Burlington between 1987 and 1989 that accounted for the region’s greatest increase in water area over the period; (B) Urban expansion in the Triangle region from the towns of Cary and Apex into the surrounding agricultural fields; (C) The Triad region, consisting of the cities of Greensboro, High Point, and Winston-Salem (off map). Vertical axes of barplots are in units of percent cover.

Fig. 5. Landcover change among generalized classes within the study area from 1985 to 2006.
with the respective posterior class probabilities. Thus, the lower prediction interval for each class expanded in proportion to its predicted cover. Strong asymmetries between the upper and lower bounds of class 80 resulted from biased classification of fields (high PA, low UA), with over-prediction leading to negatively biased prediction intervals. Due to this over-prediction, the observed changes in field cover were smaller than the class’ margin of classification error. The cover of low-density suburbs at any time was also surrounded by broad intervals relative to the magnitude of predicted cover, but roughly equal errors of omission and commission led to more equal distribution of uncertainty in this class. The high precision and minimal bias in classifying evergreen forests resulted in narrow, symmetric prediction intervals of this class over time.

5. Discussion

Ecologists and other scientists are increasingly embracing the challenge of understanding and predicting the complex dynamics of coupled human and natural systems (Colwell, 1998; Holling, 2001; Michener et al., 2001; Pickett et al., 2005). In this context, land cover datasets are relied upon primarily as indicators of underlying ecosystem conditions such as habitat value or carbon sequestration potential (Heinz Center, 2002; MEA, 2005; NRC, 2000), and are employed predominantly as uni- or bi-temporal maps. However, ecosystems exhibit nonlinearities, time lags, legacy effects, and threshold responses (Burke et al., 1990; Foster, 1992; Green et al., 2005, 2006), complexities which cannot be observed between two snapshots in time. Analyzing coupled human-natural ecosystems requires time-series of land cover data that are both sufficiently dense and extensive in time to adequately capture complex system dynamics (e.g., Lunetta et al., 2004). The reference observations needed to fit and validate land cover classifications are scarce, and so providing these spatially extensive, multi-temporal datasets requires spatial and temporal extrapolation.

Although multi-temporal signature extension adds another dimension to purely spatial land cover classification, the added complexity does not fundamentally change the statistical nature of the signature extension problem as long as sampling is adequate in both spatial and temporal domains. Variation of spectral signatures is increased when models are extrapolated over both space and time, but the robustness of multi-temporal land cover classifications can be maximized despite this uncertainty. To do so, four fundamental aspects of modeling itself must be addressed:

(1) semantic relationships between the model and reality,
(2) causal relationships between land cover and the data used to predict it,
(3) representation of spatial and temporal variation, and
(4) uncertainty accumulated in the modeling process.

5.1. Semantics

The semantic stage of classification is unavoidably subjective. Defining the set of classes to be assumed a priori by all subsequent analyses requires a great deal of induction and anticipation of the data’s later use, as well as compromise between the classes desired and those that are discernible in the data. This effort is well spent however, as loose definitions can lead to spurious conclusions and bad decisions (Comber et al., 2005; DeFries & Los, 1999).

The most effective solution to this tradeoff is to collect more data, augmenting the database by denser, more extensive, or altogether different measurements (e.g., lidar, radar, hyperspectral, or environmental data). However, this solution is only feasible where and when the required data are available, and today’s measurements are not available for the past. Extrapolation requires limiting the database to measurements that are both available and consistent over time.

Reliance is thus placed on robust features of the spectral, spatial, and temporal signatures of the cover classes. Our results show that slight modifications to the NLCD 2001 classification scheme yield high accuracy at a generalized schematic level, but confusion existed among specific classes within the genera. Among herbaceous classes, confusion was due to two reasons: (1) the variable practice of planting secondary “cover” crops (i.e., alfalfa, clover, etc.) in harvested row-crop fields over the winter and (2) poor visual discrimination between mowed hay and harvested row-crop fields in the aerial photographs used as reference images. Among forest classes, confusion was due to the arbitrary partitioning of this continuous gradient into three classes—i.e., deciduous, evergreen, and mixed forest. In urban types, confusion was also due to arbitrary partitioning, as well as to the inherently mixed nature of the low-density urban class.

Semantic errors can thus arise from three sources: (1) class variability beneath the spatial or temporal resolution of the data; (2) poor visual discrimination of classes in reference images; and (3) arbitrary binning of continuous gradients into discrete partitions, which can be either between gradients of similar (e.g., forest) or different types (e.g., the mixes of forested, herbaceous, and impervious cover that characterize developed classes). To the degree possible, classes should be defined such that they are distinct at the temporal and spatial grain of analysis, clearly identifiable in reference datasets, and based on discrete types in reality (Kelley et al., 1999). Hierarchical schemas can help to trap errors between specific classes within genera or broader levels of association (Di Gregorio and Jansen, 1998), and methods must be developed to aggregate uncertainty when merging classes.
5.2. Causation

Land cover change is affected by its ecological and economic environment (Lambin et al., 2001). Based on these relationships, “ancillary” variables have long been used to increase accuracy in land cover mapping (Jensen, 2005) and, more recently, in change detection as well (Rogan et al., 2002). However, despite the obvious benefits to accuracy, incorporating the variables that cause land cover changes into the classification logic complicates their use in subsequent analyses. In the worst case, the circular logic imposed by mixing of spectral and environmental data can completely invalidate some applications of the data.

This is especially true for dynamic variables measured once and for either static or dynamic variables whose effect on land cover changes over time. For example, municipal boundaries for a single time might well improve classification accuracy, even in years far from that of the data’s origin, but treating the boundary’s dynamic effect as a constant would inflate the probability of some changes while diminishing others over time. Subsequent analyses would be unable to distinguish the real effect of the municipality from the artifact imposed by the classification algorithm. To avoid circular logic, classifications intended for subsequent analysis should be based on spectral reflectance alone; and when environmental data are deemed necessary, their effects should be reported so that subsequent researchers can avoid spurious or circular interpretations.

5.3. Sampling

Whether over space or time, model extrapolation requires adequate representation of the variability to be encountered. Reference observations are typically sampled widely over study areas to capture the spatial variation of spectral signatures, based on the Law of Large Numbers (Bernoulli, 1713). In spatio-temporal applications, it is likewise necessary to distribute reference samples widely over the spatial and temporal domain. Collecting these spatio-temporal reference samples is more difficult than purely spatial sampling because of the sparser availability of the necessary high-quality historical images (as well as the possibility of retroactive ground-truth). However, growing archives of high-resolution aerial imagery promise to ease this limitation and will likely eventually provide sufficient coverage to fully capture temporal as well as spatial variations.

5.4. Uncertainty

Despite their uncertainties, land cover maps are analyzed as data – if not as truth itself – in a wide variety of scientific applications. Accuracy assessment has become common practice in the production of maps (Foody, 2002; Stehman, 1997), and propagation of uncertainty has long been an important focus of geographic information systems research (Hunsaker et al., 2001). Due diligence toward model and data validation promotes long-term methodological advancement as well as continued adoption by the user community.

However, accuracy assessments are rarely directly applicable to the various samples drawn by users (Appendix A). This is due in large part to biases between the data drawn for validation versus those drawn for subsequent analysis. Adjusting validation metrics to reflect class proportions can neutralize these biases; and when weighted appropriately, estimates of uncertainty can be extrapolated from validation samples to any other sample, providing relevant accuracy estimates for different study areas and times.

Because models cannot capture every detail even in well-specified systems, it is necessary to model errors alongside the values of the response variables themselves. Biases due to sampling intensity are only one of the many sources of discrepancy, and so much research will be needed to understand, model, and ultimately remove artifacts from other sources. Although many studies have been published on atmospheric correction (Vermote et al., 2002, 2006), phenological processes are especially important, yet not yet sufficiently understood to remove their effects. In addition to operational studies of sampling adequacy, fundamental research on ecosystem phenology will be required to maximize multi-temporal mapping accuracy.

In the meantime, robust analysis of land cover maps by the growing user community requires knowledge of the uncertainty in the data, and so continued research is needed to advance methods for estimating uncertainty in multi-temporal maps. Earnest appraisal and reporting of uncertainty can minimize the downstream impact of classification error, and possibly its magnitude as well.

6. Future research

Multi-temporal land cover classification presents advantages and disadvantages that purely spatial classifications do not. Given the rapidly increasing availability and quality of satellite images, ongoing efforts will continue to improve multi-temporal land cover classifications. Townshend et al. (2012) have reviewed many of the challenges and opportunities of long-term land cover monitoring with Landsat. Based on this study, we suggest a few specific areas where gains may be close at hand.

6.1. Robust classification schemes

With increasing temporal extension of land cover datasets, continued research is needed to define classification schemes that are robust despite phenological, atmospheric, and other variation across both space and time (Strahler, 1980). Given the difficulty of discriminating them, special attention should be paid to urban and herbaceous classes. Archives of high-resolution images over time, exploratory cluster analyses (e.g., Kaufman & Rousseeuw, 1990) on multi-temporal data, and hierarchical approaches such as the United Nations’ Land Cover Classification System (Di Gregorio & Jansen, 1998) will be critical to these studies.

6.2. Multi-model and adaptive parameterization

Collection of a spatio-temporal reference dataset is necessary to fully capture the variability in spectral signatures, but fitting the classifier simultaneously to all years’ training data sacrifices accuracy in any given year for that of the whole. Given sufficient sample sizes, accuracy may be further improved by fitting individual models for each reference year, and either selecting the one most similar to each extrapolation year or combining land cover probabilities from multiple models as weighted averages. Model selection or averaging could be based entirely upon the images themselves, through distributional similarities between the reference and extrapolation images.

Alternatively, hierarchical models capable of incorporating systematic variations in data and/or parameters will likely lead to further improvements in data quality and error characterization (Clark, 2005). Due to the large sample sizes afforded by satellite datasets, remote sensing is often immune to controversy over the importance of model priors (reviewed in Clark et al., 2004). Further, given the many sources of “true” variation (e.g., phenology, terrain, atmosphere, illumination geometry, and land cover change itself) as well as statistical artifacts from sampling biases and errors due to model imprecision or misspecification, multi-temporal remote sensing seems an appropriate application of the hierarchical modeling framework.

6.3. Temporal filters

Relying on horizontal texture, spatial filters – including pre-classification “high-pass” and “low-pass” convolutions (Jensen, 2005) as well as post-classification “clump-and-eliminate” (Leica Geosystems, Inc., 2006) and “boundary-clean” (ESRI, 2008) operations – are applied routinely to increase map accuracy (Jensen, 2005). Post-classification
filters typically assume spatial homogeneity of cover, recoding isolated pixels to match the locally dominant type. Compared to variations in space however, land cover change in time is a relatively rare phenomenon, and even rarer are changes back and forth between types over short intervals. This constancy over time provides a complementary approach to spatial filters for removing erroneous cover changes. By comparing each location’s cover classification within short time intervals, the temporal stability of cover can be used to filter out erroneous changes and increase the accuracy of each map in the series. Although this approach was initially suggested over three decades ago (e.g., the “cascade classifier”, Swain, 1978), and similar, heuristic approaches have been successful more recently, (e.g., Fraser et al., 2009; Latifovic & Pouliot, 2005), the breadth of possibilities in this area remains largely unexplored and unexploited.

Similarly, pre-classification temporal filters based on plant phenology – including approximately weekly, monthly, and annual integrations – have been applied to high temporal frequency images to improve classification accuracy in coarse resolution land cover datasets (e.g., Belward et al., 1999; Latifovic & Pouliot, 2005). This technique shifts reliance from phenological similarity between the dates of the reference and each extrapolation image to a more general constancy of class summary statistics (e.g., maximum, minimum, median NDVI). Landsat’s 16-day orbit cycle constraints the choice of metrics compared to those supported by sensors with daily overpasses, and so research is needed to determine the best sets of indices for any given region. Also, the sample available within each temporal window depends on the number of cloud-free observations for each location. Multi-date MODIS composites rely on automated cloud-removal algorithms (King et al., 1997), and so improvements in cloud-detection from the Landsat bands would likewise increase the utility of partially cloudy Landsat images for multi-date composites, thereby increasing the number of observations available for temporal composites.

6.4. Land cover dynamics

The increased availability of the Landsat archive has the potential to change not only the way regional land cover maps are produced, but also how they are analyzed. Provided with dense, long-term datasets, users will be able to analyze nonlinear changes over time, including the accelerations and thresholds hypothesized for complex systems. Larger sample sizes in time will also lighten the burden on each map, allowing users to minimize inference based on outliers and other errors. With proper accounting for uncertainty, this promotes a shift in the analytic framework from change detection to a more general, regression-based approach. In short, ecologists and other Earth-scientists will be liberated to reduce their reliance on individual maps as absolute truth and begin treating entire series as data. This will in turn promote a shift in focus from detecting land cover changes to understanding land cover dynamics.

7. Conclusions

Adequately representing the complex dynamics of coupled human-natural systems requires consistent land cover maps spanning multiple time periods. With proper image calibration and representative spatio-temporal sampling, these datasets can be produced by supervised classification of time-serial Landsat images. Due to the added uncertainty of temporal extrapolation and the scarcity of historical reference data, methods must be applied to estimate uncertainty along with predictions over time. The resulting multi-temporal maps can depict spatio-temporal complexities of change that are not observable either in sparser time series or by comparing maps from varying provenances. Increases in data availability will continue to facilitate such studies, and attention to the semantics, causation, representation, and uncertainty in the classification process will maximize reliability of time-serial maps for a receptive and growing user community of scientists and managers. Continued progress in this area will allow robust monitoring and analysis of the land cover dynamics of Earth’s coupled human and natural ecosystems.

Acknowledgments

This research was supported by the NASA Earth Systems Science Fellowship Program, and was endorsed by the Global Land Project, a joint project of the International Geosphere–Biosphere Programme and International Human Dimensions Programme on Global Environmental Change. The work was performed in the Duke University Landscape Ecology Laboratory. Landsat images were contributed by the Multi-Resolution Land Characteristics Consortium, the Global Land Cover Facility, the Duke University Social Science Research Institute, and the libraries of Duke University, the University of North Carolina, and North Carolina State University. High-resolution reference images and supporting GIS layers were provided by the State of North Carolina. Technical assistance was generously given by Ben Best, John Fay, Pete Harrell, Gyanesh Chandar, and Scott Loarie. Technicians Nathan Emery, Maggie Goldman, Neoma Lavalle, Ian Varley, Phil Schutt, and Christine Cope processed data and performed preliminary analyses. Norm Christensen and Jim Clark (Duke University), Tammy Wilson (Utah State University), Sean McMahon (Smithsonian Environmental Research Center), Fernando Colchero (Max Planck Demographic Institute), and three anonymous reviewers provided helpful editorial comments on the manuscript. Analyses were performed in ArcGIS 9.3 (ESRI, Inc.) using the Geoprocessor object in Python 2.7 (Python Software Foundation, 2008), ERDAS IMAGINE (Leica Geosystems, LLC.), and the rpart library of the statistical software R version 2.6.1 (R Development Core Team, 2007; Therneau & Atkinson, 2007).

Appendix A. Distributional standardization of classification accuracy metrics

Accuracy assessment for categorical land cover classifications is based on the i × j confusion matrix, wherein the columns (j) tally reference, or “truth” class assignments and the rows (i) tally “model” assignments resulting from applying the classifier to a dataset of predictor variables (Congalton & Green, 2009; Congalton et al., 1983). The confusion matrix can be expressed either in terms of frequencies $f_{ij}$ or in terms of proportions $p_{ij}$, by dividing each frequency by the total sample size, n:

$$p_{ij} = \frac{f_{ij}}{n}$$

(A1)

Because the reference and model frequencies (and therefore proportions) are both based on the same sample, $n = n_i = n_j$.

The vector of proportional column sums:

$$p_j = \sum_{i=1}^{k} p_{ij} = \frac{1}{n} \sum_{i=1}^{k} f_{ij}$$

(A2)

gives the reference probability distribution of the classes in the sample, and the vector of row sums:

$$p_i = \sum_{j=1}^{k} p_{ij} = \frac{1}{n} \sum_{j=1}^{k} f_{ij}$$

(A3)

gives the model probability distribution of classes in the sample.
As an estimate of the overall accuracy of the model, the Percent Correctly Classified metric, PCC, is calculated as the sum of the diagonal elements of the proportional confusion matrix, rescaled to percentages:

$$PCC = 100 \times \sum_{i} p_{ij} = \frac{100}{n} \sum_{i} f_{ij}$$  \hspace{1cm} (A4)

Conversely estimating the accuracy of the model specifically with respect to the reference class assignments, “producer’s” accuracy for each class j is calculated as the proportion of reference assignments to class j also assigned by the model:

$$PA_j = \frac{p_{0j}}{\sum_j p_{ij}}.$$  \hspace{1cm} (A5)

Producer’s accuracy is so named because it informs the map producer of the accuracy of his or her classifier based on information to which a user of the map does not have access—i.e., the reference, or “truth” data. The inverse of producer’s accuracy (1-PA) is interpreted as errors of omission. Conversely, “user’s” accuracy quantifies model accuracy relative to the predictions themselves, and is calculated as the proportion of model assignments to class i in agreement with the reference assignments:

$$UA_i = \frac{p_{i0}}{\sum_j p_{ij}}.$$  \hspace{1cm} (A6)

The inverse of user’s accuracy (1-UA) is interpreted as errors of commission.

The Kappa coefficient (Cohen, 1968) is the proportion of agreement between a reference and model classification, adjusted for chance:

$$\kappa = \frac{p_c - p_e}{1 - p_e},$$  \hspace{1cm} (A7)

where $p_c$ is the observed proportion of agreement between model and reference and $p_e$ is the proportion of agreement expected from a random assignment of classes within the sample. Originally developed to measure agreement between observers in psychometric studies (Cohen, 1968), Kappa has been widely adopted by the remote sensing community for accuracy assessment in land cover classifications (Foody, 2002; Liu et al., 2007).

The observed proportion of agreement $p_o$ is equivalent to PCC on the unit scale:

$$p_o = \sum_{i} p_{ij},$$  \hspace{1cm} (A8)

and the chance proportion of agreement $p_e$ is calculated as the sum of the row x column (i.e., marginal) products of the confusion matrix:

$$p_e = \sum_{i,j} p_{ij} \times p_i.$$  \hspace{1cm} (A9)

The Kappa coefficient is therefore calculated as:

$$\kappa = \frac{\sum_{i,j} p_{ij} - \sum_{i,j} p_{ij} p_{ij}}{1 - \sum_{i,j} p_{ij} p_{ij}} - \frac{\sum_{i} f_{ij}}{n^2 - \sum_{i} f_{ij}}.$$  \hspace{1cm} (A10)

Assuming that changes to the sample distribution affect model errors proportionally (i.e., that observations added or removed bear the same covariance structure as the original sample, or that changes to the sample impart no relational bias), PA and UA are unaffected by the class frequency distribution of the sample. But because PCC and Kappa summarize accuracy across the distribution of classes, these metrics are affected by the relative class frequencies within the sample. This is known as the prevalence problem (Byrt et al., 1993) — leading to different accuracy estimates from various samplings of the population, prevalence bias greatly diminishes relevance of accuracy estimates based to populations of which the sample is not representative.

Several modifications have been proposed to adjust Kappa (Cohen, 1968; Foody, 1992; Ma & Redmond, 1995). Of these, weighted Kappa $\kappa_w$ (Cohen, 1968) can be defined to remove the effect of sampling bias by standardizing the metric to a reference distribution. This is accomplished by defining the weights used to compute $\kappa_w$ as vector $w$ of class weights, with each element $w_j$ equal to the ratio of the population proportion to the sample (reference) proportion for each class (Stehman & Czaplewski, 1998):

$$w_j = \frac{N_j}{n_j} = \frac{n \times N_j}{N \times n_j}.$$  \hspace{1cm} (A12)

The standardized Kappa coefficient is thus calculated as:

$$\kappa_s = \frac{\sum_{i,j} w_{ij} p_{ij} - \sum_{i,j} w_{ij} p_{ij} p_{ij}}{1 - \sum_{i,j} w_{ij} p_{ij} p_{ij}} - \frac{n \sum_{ij} f_{ij} - \sum_{ij} w_{ij} f_{ij}}{n^2 - \sum_{ij} w_{ij} f_{ij}}.$$  \hspace{1cm} (A14)

Similarly, the vector $w$ can be applied to the diagonal of the confusion matrix to remove sampling bias from PCC:

$$PCC_s = \sum_{i,j} w_{ij} p_{ij} = \frac{n \sum_{ij} w_{ij} f_{ij}}{n^2}.$$  \hspace{1cm} (A15)

This distribution-based weighting scheme has the effect of standardizing PCC and the Kappa coefficient computed from any sample to some a priori distribution of classes, thus allowing consistent comparisons regardless of sampling idiosyncrasies. Further, any distribution may be chosen as the standard, including uniform (i.e., $n_i = n_j$ for all i,j) or some accepted empirical distribution such as the U.S. National Land Cover Database 2001 (Homer et al., 2004).

UA and PA can also be weighted based on a class frequency distribution to estimate errors for a class over a predicted land cover map:

$$PA_{map}(i) = p_i \times PA_i,$$  \hspace{1cm} (A16)

and

$$UA_{map}(i) = p_i \times UA_i,$$  \hspace{1cm} (A17)

where $p_i$ is the probability or proportion of class i over the area of interest. However, the two metrics must use different distributions as their standards. As a metric of the errors of omission, 1-PA_{map} must be based on some reference distribution that is accepted as truth—identical to the adjustment of Kappa and PCC above. But as a metric of the map’s errors of commission, 1-UA_{map} is based on the distribution of the predictions themselves.

Based on these two metrics, the range of frequencies (or proportions) can be predicted for a land cover map. The expected upper bound for the frequency of class i on the map, $f^u_i$, is:

$$f^u_i = f_i + (p^r_i \times (1-PA_i)),$$  \hspace{1cm} (A18)
where $p_i$ is the proportion of $i$ in the reference distribution. The expected lower bound for the frequency of $i$ on the map is:

$$ f_i^- = f_i - \left( \bar{p} (1 - UA) \right) $$

(A19)

where $\bar{p}$ is the posterior probability of $i$ i.e., its proportion on the predicted map.

This method corrects only for prevalence bias in the validation sample, but does not account for phenological, atmospheric, or other changes in the cover–reflectance relationship between the reference and extrapolation datasets (i.e., relational, or reflectance bias). Thus it assumes that the sample is adequate and that the error-generating phenomena are constant over space and time. Atmospheric and especially phenological conditions vary considerably in time, and so effort should be spent maximizing the sample size, with special attention to temporal coverage. However, even with adequate samples, variations in the reflectance properties of the land cover and atmosphere will not be accounted for, and research into the underlying processes necessary to incorporate this variation as well.

References


