Modeling the Effects of Horizontal Positional Error on Classification Accuracy Statistics

Henry B. Glick, Devin Routh, Charlie Bettigole, Lindsi Seegmiller, Catherine Kuhn, and Chadwick D. Oliver

Abstract

Using a concept proposed by Stehman and Czaplewski (1997), we implemented spatially-explicit Monte Carlo simulations to test the effects of manually introduced horizontal positional error on standard inter-rater statistics derived from twelve classified high-resolution images. Through simulations we found that both overall and kappa accuracies decrease markedly with increasing error distance, varying greatly across distances relevant to practical application. The use of ground reference sites falling solely in homogeneous patches significantly improves inter-rater statistics and calls into question the use of kernel-smoothed data in one-time accuracy assessments. Our simulations offer insight into the scale of both structural and cover type heterogeneity across our landscapes, and support a new method for minimizing the effects of positional error on map accuracy. We recommend that analysts use caution when applying traditional accuracy assessment strategies to categorical maps, particularly when working with high-resolution imagery.

Introduction

The assessment of thematic map accuracy, defined as the degree to which classified or categorical mapped feature labels correspond with true feature labels, has been central to the field of remote sensing for approximately 40 years (Congalton and Green, 2009). During this time, practitioners have adopted increasingly rigorous assessment strategies. The use of error matrices and inter-rater agreement statistics (e.g., overall accuracy, Cohen's (1960) kappa coefficient, errors of omission and commission, user's and producer's accuracies) has been common in recent decades (Congalton, 1994; Congalton and Mead, 1983; Congalton *et al.*, 1983; Liu et *al.*, 2007; Story and Congalton, 1983), although the remote sensing community has not reached consensus on general standards, statistical reporting, or target accuracies (Foody, 2002; Liu *et al.*, 2007; Stehman, 1997). That thematic accuracy assessment has not been standardized in the way that positional accuracy assessment has (i.e., ASPRS, 2015; FGDC, 1998) may be due to this lack of consensus, its shorter history (40 years versus over 70 years; Congalton and Green, 2009), and/or the rapidity with which remote sensing technology has changed (Campbell and Wynne, 2011; Graham, 1999). As the resolutions (i.e., spatial, spectral, radiometric, and temporal) of remotely sensed data sets become continually finer, we face the ongoing challenge of how best to evaluate our spatial models.

Error-matrix based accuracy assessment relies on the ability to relate known information from ground reference sites to the predicted information for those sites. However, for the comparison to take place, the locations must be described in theoretical space using coordinate systems (usually a function of GPS or survey-based geometry), datums, and geospatial projections that rely on transformation functions and generalized models of the Earth's shape. If we were to trust that two datasets with the same coordinate systems and no positional error were perfectly co-registered (a questionable assumption given satellite orbiting speeds, changes in satellite altitude, and rounding error, among others) then we could be confident that our comparison of the datasets makes sense at all spatial scales. Unfortunately, we can rarely, if ever, satisfy these conditions.

The most spatially limiting sampling unit size occurs when treating the imagery as point-sampled data in which each point has the spatial extent of a single pixel (Janssen and van der Wel, 1994), and where a single pixel is used to represent a ground control plot. To accurately compare ground and image data using this format, one needs a portable Global Positioning System (GPS) receiver and a georeferenced image whose combined root mean squared horizontal georeferencing error is less than one-half the length of the shortest side of a pixel. These hypothetical circumstances are seldom achievable. Many analysts often ignore one form of georeferencing error, limiting registration hurdles to either (a) obtaining a highly accurate image, or (b) using a GPS receiver whose horizontal error is less than one-half the length of the shortest side of a pixel (Weber, 2006). Recreational-grade GPS technology has improved markedly since its early years, and many affordable units can now achieve locational accuracies of ±5 m or less. For moderate to coarse resolution datasets (e.g., Landsat, EO-1, MODIS), such GPS units work well. However, as Weber (2006) points out, they are generally insufficient to assess the accuracy of high spatial resolution datasets (here defined as imagery with a ground resolved distance of 5 m or less), and failure to understand GPS unit receiver-specific accuracy and error propagation may lead analysts astray.

High spatial resolution satellite imagery has been available on the commercial market for roughly 15 years. Though it offers a unique and often alluring level of detail, as image resolution increases the analyst must contend with potentially less accurate co-registration between image and ground reference sites, as well as a concomitant decrease in class separability that may reduce classification accuracy (Carleer and Wolff, 2005). When measured as a function of linear

Henry B. Glick, Devin Routh, Charlie Bettigole, and Chadwick D. Oliver are with the Ucross High Plains Stewardship Initiative, Yale School of Forestry and Environmental Studies, 195 Prospect Street, New Haven, CT, 06511, USA (henry.glick@yale.edu; henry.glick@gmail.com).

Lindsi Seegmiller is an independent geospatial consultant, 3319 North Stone Creek Circle, Madison, WI, 53719; and formerly with the Ucross High Plains Stewardship Initiative, Yale School of Forestry and Environmental Studies, Yale University.

Catherine Kuhn is with the School of Environmental and Forest Sciences, University of Washington, Box 325100, Seattle, WA, 98195; and formerly with the Ucross High Plains Stewardship Initiative, Yale School of Forestry and Environmental Studies, Yale University.

Photogrammetric Engineering & Remote Sensing
Vol. 82, No. 10, October 2016, pp. 789–802.
0099-1112/16/789–802

doi: 10.14358/PERS.82.10.789

distance, co-registration error may generally be lower for high resolution imagery. In this context however, we are referring to co-registration error as a function of the number of pixels, which increases with increasing resolution if linear error is held constant. Significant attention has been devoted to maximizing the potential of high resolution datasets, particularly as they relate to land cover and land use classification (e.g., Elsharkawy *et al.*, 2012; Immitzer *et al.*, 2012; Kiang, 2002; Liu and Yamazaki, 2012; Ünsalan and Boyer, 2004; Wolf, 2012). To date, the primary application of high-resolution imagery has been on urban and infrastructure-laden environments, with less work focused on highly vegetated landscapes (Carter, 2013; Kiang, 2002; Yu, *et al.* 2006; Zhang and Huang, 2010) whose heterogeneity may exacerbate inaccuracies resulting from a variety of sources, including poor co-registration (Gu *et al.* 2015).

Accuracy assessment of high-resolution datasets parallels the assessment of moderate and coarse resolution data prior to the removal of "Selective Availability" of civilian GPS in 2000 (Grewal *et al.*, 2001). In both cases, GPS equipment exceeds tolerances for locational error. This calls into question the use of single pixels as ground control plots, and so discussions of kernel smoothing (i.e., applying a filter across a multi-cellular area to contextualize the center cell) and optimal sampling unit size become increasingly important (Congalton and Green, 2009; Plourde and Congalton, 2003; Stehman and Czaplewski, 1998; Stehman and Foody 2009). In an attempt to remove bias introduced by poor co-registration (also known as horizontal locational error, horizontal positional error, and locational misalignment, among other terms), many analysts have adopted an approach wherein accuracy assessment statistics are computed solely from ground reference sites falling within homogeneous multi-cellular patches whose size(s) exceeds the horizontal error of their GPS receivers (Hammond and Verbyla, 1996; Plourde and Congalton, 2003; Stehman and Foody, 2009). This approach is thought to largely remove the effects of locational error by ensuring that the vast majority of GPS-surveyed sites fall *somewhere* within uniform (predicted) cover classes. Homogenous patches from which to sample are either inherent in the classified dataset, derived by applying assignment operators to fuzzy classifications (Gopal and Woodcock, 1994; Stehman and Czaplewski, 1998), or derived by application of a moving window analysis (i.e., kernel function) that helps to homogenize heterogeneous class structures (Stehman and Foody, 2009). With traditional error matrix-based accuracy assessment methods, this sampling scheme is often considered acceptable even though it may artificially inflate accuracy statistics (Stehman and Czaplewski, 1998; Plourde and Congalton, 2003; Stehman and Foody, 2009).

To better understand the effects of horizontal locational error on common accuracy assessment statistics, we explored a theoretical concept proposed by Stehman and Czaplewski (1997, p. 550) in which we manually introduced horizontal error between digitally fixed ground reference locations and their perceived locations on twelve high-resolution classified maps through a series of spatially-explicit Monte Carlo simulations. The primary objectives were to understand how overall and kappa accuracies change with increasing locational error under three scenarios: (1) using a random selection of ground reference sites, (2) using ground reference sites falling solely in homogeneous patches, and (3) using ground reference sites falling solely in heterogeneous patches. Since small homogeneous patches may be associated with larger homogeneous landscape structures, we postulated that not only would inter-rater statistics improve when performing accuracy assessment using a sub-sample from homogeneous patches, but that spatial autocorrelation may reduce the value of inter-rater statistics derived from homogeneous patches at multiple scales of horizontal displacement. A secondary objective of this study was to examine the variability of overall and kappa accuracies across distances to better understand how spatial autocorrelation and landscape structure affect accuracy assessment statistics.

We present details on twelve high-resolution categorical maps used in testing, followed by findings from our simulations. While locational error is a central challenge to those working with high-resolution imagery specifically, this report will provide useful guidance to those working with categorical maps, their production, and the assessment of their accuracies.

Study Areas

Our analysis focused on twelve high spatial resolution (≤5 m per pixel side) categorical maps from across the U.S. and southern Canada (Figure 1). These maps, or portions thereof, were selected from a slightly larger pool of available data to enable analysis of different types of landscapes (e.g., urban, rural, coastal, agricultural) and different landscape structures (i.e., the spatial configuration of cover types). The maps used in this study came from private collections in addition to the "Tree Canopy Assessment" distribution portal (*http://gis.w3.uvm.edu/utc/*) hosted by the Spatial Analysis Lab at the University of Vermont. Many of the Tree Canopy Assessment datasets were produced as part of the US Department of Agriculture Forest Service's Urban Tree Canopy (UTC) assessment program, and details on their methodological underpinnings can be found in O'Neil-Dunne *et al.* (2013 and 2014). Broadly speaking, the UTC datasets were produced using geographic object-based image analysis (e.g., feature extraction) and data fusion methods, drawing on high-resolution multispectral imagery, thematic data, and lidar datasets. O'Neil-Dunne et al. (2014) report that all datasets achieved >90 percent overall accuracy, with some as high as 98 percent. This bears relevance because the present studied necessitated categorical maps with realistic landscape structures, and the high accuracies reflect real-world abundances and distributions of cover types. Landscape structure aside, the present study did not require high accuracy maps and could theoretically have been conducted on artificial, computer-generated matrices.

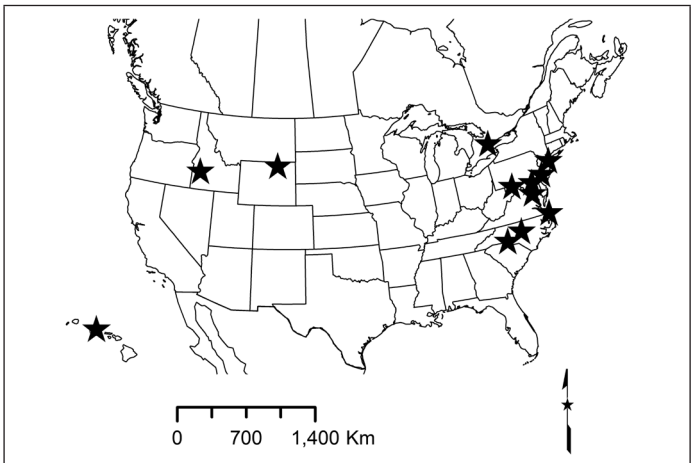

Figure 1. The locations of mapped landscapes used in analysis.

Table 1 provides general information on the categorical maps we used, including two metrics that quantify landscape structure: patch density, and the interspersion and juxtaposition index (IJI; McGarigal and Marks 1995), both of which were computed at the landscape level using Fragstats v. 4.2.1.603 (McGarigal *et al*, 2012). A large number of metrics exist for evaluating the structure of landscapes (McGarigal,

2015). Patch density and IJI were chosen because they constitute measures of spatial aggregation that are interpretable and, in the case of IJI, robust to changes in spatial resolution (Corry and Lafortezza, 2007; McGarigal, 2015).

Patch density is computed as the number of patches of any cover type within 1 km^2, constrained by 0 on the low end and controlled by the spatial resolution of the map at the high end (i.e., a finer grained map would permit a greater number of theoretical patches to exist). Patch density is sensitive to grain size; as cell size increases patch density decreases since there are fewer pixels per unit area. We have included this metric because it captures the "clumpiness" of a landscape as one might think of it in casual terms.

The IJI aims to quantify the mixing of land cover/land use classes by considering adjacencies between the classes. Unlike contagion, a similar metric that targets dispersion and interspersion, the Interspersion and Juxtaposition Index is based on patch adjacencies instead of cell adjacencies. The index "isolates the interspersion aspect of aggregation; it increases in value as patches tend to be more evenly interspersed in a 'salt and pepper' mixture" (McGarigal, 2015; p. 129; original emphasis). The IJI is expressed as a percentage (0 to 100 scale) of the maximum interspersion of patches that is possible given the number of different patch types.

Methods

The Ucross, Wyoming map (See Table 1) was the only dataset used in this study that we produced. Using field-derived training regions we performed a guided clustering, unbounded, maximum-likelihood supervised classification (Bauer *et al.*, 1994; Lillesand and Kiefer, 1999) on a single atmospherically corrected (Perkins *et al.*, 2012) and orthorectified WorldView-2 image. A low-pass majority smoothing filter was employed because the positional error of our GPS units (95 percent circular error probable of ±2.38 m) exceeded one-half the pixel size (2 m) of our imagery. The 74 percent overall accuracy (kappa = 0.73) reported in Table 1 was achieved through lab-based photointerpretation (e.g., Klöditz *et al.*, 1998; Seto, 2002; Zhang and Huang, 2010), whose reliability was independently evaluated using 80 field sites from select cover types that were more challenging to photointerpret.

For each of the twelve landscapes noted in Table 1, we projected the raster into the location-appropriate Universal Transverse Mercator (UTM) coordinate system with a metric linear unit. Then, as a starting point, we generated 50 ground reference sites per land cover class using a stratified random sampling design (Brioch et *al.*, 2009; Gregoire and Valentine, 2008). Each ground reference site was treated as a point-sample, such that each site had a specific location in two-dimensional space with no inherent spatial area (i.e., each point was dimensionless).

Simulating Locational Error

To conduct a thorough locational error assessment the ground reference sites (n = 50 * number of classes) were first divided into three datasets. The first set included only those locations where the land cover at those points was homogeneous across a 3 × 3 cellular (i.e., pixel-based) area surrounding each point; the second set consisted of the remaining heterogeneous locations; and the third set consisted of the full combination of the two. From these three datasets we then performed both a basic and a more comprehensive assessment of horizontal locational error.

For the basic assessment we contrasted the figures from homogeneous sites with those from the complete dataset to obtain a single metric for how much error is attributable to horizontal displacement. The way in which this comparison isolates locational error is described below. To evaluate the effects of horizontal misalignment at a more comprehensive level, we used the three datasets noted above as the base files

TABLE 1. DESCRIPTIVE STATISTICS ON THE TWELVE LANDSCAPES USED IN ANALYSIS

Location	Initial Res. (m)	Processed Res. (m)	Num. of Classes	Patch Density	IJI	Overall Accuracy	Area km^2	Description
Baltimore, MD	0.91	3	7	2106.1	74.7	94%	238.0	City of Baltimore including water of Patapsco River/Chesapeake Bay
Bolton, ON	0.6	0.6	7	5829.7	58.8	≥90%	16.8	Small town, dominated by suburban housing developments and industrial/commercial
Cumberland, MD	1.04	1.04	7	1697.4	72.6	≥90%	26.1	Small city surrounded by large tracts of forest
Garden City, ID	1	1	9	919.6	58.0	≥90%	11.0	Mixed development with minimal open space
Honolulu, HI	0.5	1	8	1291.2	64.6	≥90%	7.3	Subset of city, dominated by agricultural land; small contributions from development and water
Mecklenburg County, NC	1	1	7	2059.3	69.4	≥90%	73.9	Subset of county, dominated by suburban housing and developed open space
New York City, NY	0.91	0.91	7	2873.0	62.1	98%	26.4	Subset of city, dominated by dense urban, central park, Hudson and East Rivers
Philadelphia, PA	0.3	0.3	7	1024.2[a]	68.1[a]	95%	11.9	Subset of city, dominated by industrial, developed open space, water
Prince George's County, MD	0.91	3	7	919.6	58.0	≥90%	645.7	Subset of county (southern half), extensive forest/undeveloped, agriculture, with minor developed components
Ucross, WY	2	2	12	2178.9	40.3	77%	111.7	Undeveloped working rangeland and agriculture
Virginia Beach, VA	0.61	5	7	245.2	61.6	≥90%	404.5	Large coastal complex, agriculture, and forest
Wake County, NC	2.4	2.4	5	7284.8[a]	60.0[a]	N/A	449.8	Mixed rural landscape
Range	0.3–2.4	0.3–5	5–12	245.2–5829.7	40.3–74.7	77–98%	7.3–645.7	

[a]Due to computational limitations, this value represents the average of two tiles covering the study site.

for three independent Monte Carlo simulations. The simulations were scripted, implemented, and visualized using R v. 3.1.3 (R Development Core Team, 2015).

In each simulation, all target ground reference sites (homogeneous, heterogeneous, or combined) were sequentially buffered to create a series of annuli extending from 0 m to 1,000 m (Table 2, Figure 2). Then, for each set of annuli (where a set contains a single distance range, e.g., 10 m to 11 m), a single spatially random point was generated within each confining annulus, and the land cover value denoted by our classified image was recorded for that location. For each reference point in each of the three datasets (homogeneous, heterogeneous, or combined) this process was repeated 100 times at 1 m increments from 1 m to 50 m, and 100 times at tiered distances from 75 m to 1,000 m. Separate error matrices were used to compute the overall and kappa accuracies for each iteration (n = 5,900) of each simulation (n = 3) of each landscape (n = 12) by pairing the extracted values with the original reference cover types. Multivariate analysis of variance (MANOVA) was used to evaluate the effects of dataset group assignment on both overall and kappa accuracies. Post-hoc univariate contrasts in the form of Tukey's Honest Significant Difference (HSD) tests were applied to both overall and kappa accuracies to assess pairwise differences between datasets.

In addition to the above-mentioned assessments, the Monte Carlo simulations we conducted allowed us to indirectly visualize and quantify the degree of spatial autocorrelation across the landscape. We plotted the variability in accuracies across distance to examine the interplay between autocorrelation and classification accuracy.

Several details deserve elaboration. First, it may initially appear that when using these approaches, the effects of locational error cannot be teased from issues of class separability, thematic (classification) error, and/or sampling error. The confounding effects of class separability are intrinsically tied to thematic accuracy, wherein greater spectral confusion may contribute to lower accuracy when using certain classifiers. As mentioned above (see the Study Areas section), thematic accuracy is relevant to our results only in the degree to which the map accurately represents the spatial distribution of cover classes. Our simulations required only that the analyzed landscapes reflected true, real-world structures. Beyond patch type dispersion and interspersion, map accuracy plays no role in governing changes in accuracy that result from introduced locational error. The reason for this is that, through our approach, each landscape is evaluated with respect to itself. This means that the mapped class value at each control site is considered the true value (100 percent accurate) against which changes are measured. Where each control site is dimensionless (through point-sampling), accuracy values are not directly influenced by adjacent cells, as they would be when using patch-based assessment methods. Using this framework, the analyzed landscapes need not be "real", and could have been generated artificially, provided that their structures resembled structures in the real world.

With respect to sampling error, our basic approach remains largely unbiased (excepting the inherent use of reference sites falling in homogeneous patches), though it does not eliminate the potential confounding effects of an imprecise sample, which is precisely why we explored the more comprehensive approach. Using Monte Carlo simulations we capitalize on the central limit theorem and generally reduce the effects of sampling error through repeated sampling along with the isolation of more accurate mean values.

Finally, it is important to note that in practice, our more comprehensive strategy is not the same as moving a classified image while holding the ground reference points constant, which is another simpler approach mentioned by Stehman and Czaplewski (1997) and minimally examined by Verbyla and Hammond (1995), and Carmel *et al.* (2001). Instead, our approach enabled each point to be spatially relocated independently of all other points. For instance, one point could be moved 5.13 m at 184° while a neighboring point could simultaneously be move 5.75 m at 14°. Our analysis parallels Gu *et al.*'s (2015) recent work, focusing on maps with "hard" classification instead of fuzzy, or "soft" classification. We extend their methods by evaluating trends associated with real landscapes; by exploring datasets with numerous cover classes; by exploring a large number of horizontal displacement distances; by using Monte Carlo methods to obtain more accurate and more precise mean accuracy estimates; and by evaluating the effects of different subsets of ground control sites.

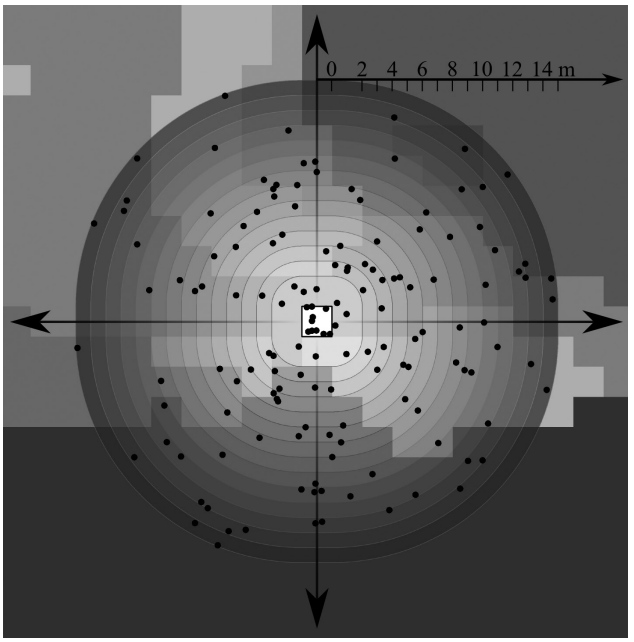

Figure 2. An example of the sampling scheme applied to each ground control site, here represented as a cell, used in error simulation. Actual sites were dimensionless points in two-dimensional space. Sampling annuli extended from 1 m to 1,000 m as shown in Table 2, and mapped classes were recorded at 100 sites per annulus (only 10 shown).

Table 2. Horizontal Displacement Distances Used in Monte Carlo Simulations

Displacement Distance (m)	Number of Iterations
0	1
1	100
2	100
3	100
⋮	⋮
49	100
50	100
75	100
100	100
150	100
200	100
275	100
375	100
500	100
750	100
1000	100

Results

Quantifying Horizontal Locational Error

When averaged across twelve landscapes, a raw comparison of one-time inter-rater statistics between homogeneous

reference sites and combined reference sites reveals an 8.1 percent change in overall accuracy (98.7 percent versus 90.6 percent) and a change of .094 in kappa accuracy (0.985 versus 0.891) attributable to potential fine-scale locational misalignment between our imagery and our ground reference sites (see Figure 3 when distance = 0 m). These values are limited, however, in that they reflect a single instance of ground assessment, shedding little light on the variability of accuracy statistics over time (e.g., repetitions) or distance.

Overall and kappa accuracies derived through simulations decreased with distance in a highly predictable fashion (Figure 3) and are best modeled using non-linear least squares regression (Table 3). Power and log-based mean models produced exceptionally high coefficients of determination

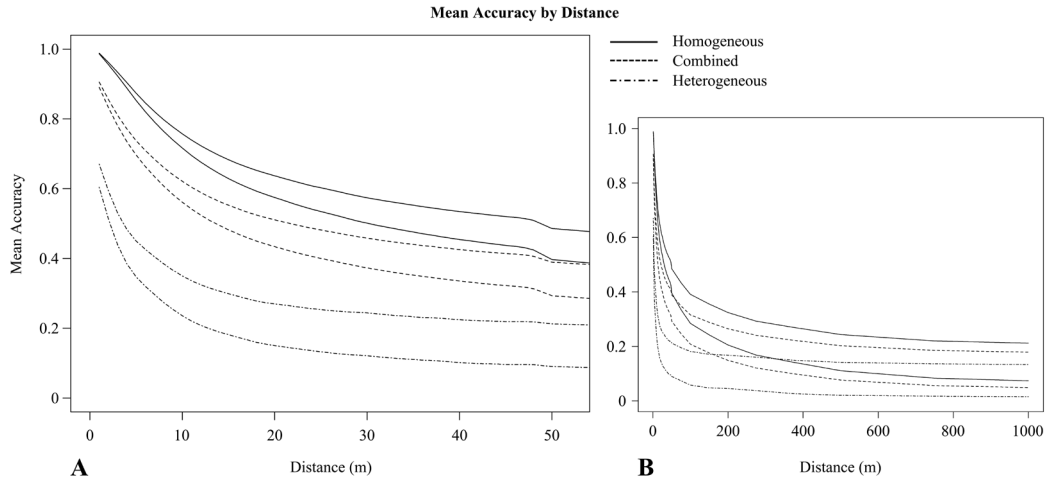

Figure 3. Mean overall (upper line of each pair) and kappa (lower line of each pair) accuracies derived from Monte Carlo simulations applied to ground reference plots originating in homogeneous, heterogeneous, or combined 3 × 3 cellular patches for error distances 0 to 50 m (A) and 0 to 1,000 m (B). Averages derived from twelve landscapes used in analysis.

TABLE 3. REGRESSION MODELS AND COEFFICIENTS OF DETERMINATION (r^2) FOR OVERALL AND KAPPA ACCURACIES, MODELED AS A FUNCTION OF DISTANCE (≤1,000 M)

		Combined		Homogeneous		Heterogeneous		Comb. Ov. - Comb. Ka.[a]	
	Location	Model	r^2	Model	r^2	Model	r^2	Model	r^2
Models for Mean Values by Distance (n = 58) — Overall Accuracy	Baltimore	$y = 1.0385x^{-0.209}$	1.00	$y = -0.126\ln(x) + 1.1399$	0.97	$y = 0.621x^{-0.222}$	0.85	$y = 0.0029x + 0.0233$	0.88
	Bolton	$y = 0.9792x^{-0.332}$	0.92	$y = -0.133\ln(x) + 0.833$	0.81	$y = 0.573x^{-0.31}$	0.83	$y = 0.0045x + 0.0275$	0.95
	Cumberland	$y = 1.2181x^{-0.292}$	0.98	$y = 1.3992x^{-0.299}$	0.97	$y = 0.5766x^{-0.248}$	0.92	$y = 0.0033x + 0.0207$	0.94
	Garden City	$y = 1.0082x^{-0.347}$	0.98	$y = 1.3092x^{-0.379}$	0.97	$y = 0.505x^{-0.255}$	0.96	$y = 0.0029x + 0.0291$	0.93
	Honolulu	$y = -0.12\ln(x) + 0.9253$	0.98	$y = -0.128\ln(x) + 0.9774$	0.98	$y = 0.5687x^{-0.173}$	0.97	$y = 0.0027x + 0.0172$	0.93
	Mecklenburg	$y = 1.2234x^{-0.306}$	0.98	$y = 1.3714x^{-0.308}$	0.97	$y = 0.5664x^{-0.294}$	0.91	$y = 0.0036x + 0.0197$	0.94
	NYC	$y = -0.122\ln(x) + 0.9169$	0.98	$y = -0.135\ln(x) + 1.0143$	0.98	$y = 0.5729x^{-0.228}$	0.95	$y = 0.0027x + 0.0181$	0.94
	Philadelphia	$y = -0.137\ln(x) + 1.0275$	0.98	$y = -0.141\ln(x) + 1.0578$	0.98	$y = 0.562x^{-0.271}$	0.90	$y = 0.0027x + 0.0143$	0.96
	Prince George	$y = -0.113\ln(x) + 0.938$	0.99	$y = -0.126\ln(x) + 1.1697$	0.96	$y = 0.7264x^{-0.309}$	0.95	$y = 0.0027x + 0.0207$	0.91
	Ucross	$y = 0.8681x^{-0.216}$	0.99	$y = -0.106\ln(x) + 0.8909$	0.99	$y = 0.5692x^{-0.192}$	0.99	$y = 0.0027x + 0.0683$	0.94
	Virginia Beach	$y = 0.9838x^{-0.189}$	0.99	$y = -0.093\ln(x) + 1.1677$	0.90	$y = 0.7262x^{-0.325}$	0.90	$y = 0.0028x + 0.0279$[b]	0.83[b]
	Wake County	$y = -0.088\ln(x) + 0.8287$	0.99	$y = -0.128\ln(x) + 1.1334$	0.97	$y = 0.6149x^{-0.139}$	0.95	$y = 0.0033x + 0.055$	0.89
Kappa Accuracy	Baltimore	$y = 1.1969x^{-0.309}$	0.98	$y = -0.149\ln(x) + 1.1612$	0.97	$y = 0.8744x^{-0.589}$	0.97		
	Bolton	$y = -0.142\ln(x) + 0.7279$	0.80	$y = -0.157\ln(x) + 0.8055$	0.81	$y = -0.081\ln(x) + 0.3927$	0.73		
	Cumberland	$y = -0.147\ln(x) + 0.8971$	0.96	$y = -0.166\ln(x) + 1.0273$	0.97	$y = 0.9665x^{-0.667}$	0.92		
	Garden City	$y = 2.2992x^{-0.737}$	0.93	$y = 2.8251x^{-0.726}$	0.93	$y = -0.067\ln(x) + 0.3554$	0.83		
	Honolulu	$y = -0.137\ln(x) + 0.9147$	0.98	$y = -0.146\ln(x) + 0.9737$	0.98	$y = 0.6038x^{-0.359}$	0.93		
	Mecklenburg	$y = -0.151\ln(x) + 0.8904$	0.94	$y = -0.168\ln(x) + 1.0053$	0.95	$y = -0.075\ln(x) + 0.3724$	0.77		
	NYC	$y = -0.139\ln(x) + 0.9051$	0.98	$y = -0.154\ln(x) + 1.0145$	0.98	$y = 0.9703x^{-0.652}$	0.93		
	Philadelphia	$y = -0.16\ln(x) + 1.0321$	0.98	$y = -0.165\ln(x) + 1.0682$	0.98	$y = 0.5679x^{-0.551}$	0.84		
	Prince George	$y = -0.131\ln(x) + 0.9277$	0.99	$y = -0.152\ln(x) + 1.2061$	0.96	$y = 1.2978x^{-0.696}$	0.97		
	Ucross	$y = 0.111\ln(x) + 0.6797$	0.98	$y = -0.137\ln(x) + 0.8849$	0.99	$y = 0.076\ln(x) + 0.4388$	0.89		
	Virginia Beach	$y = 1.0875x^{-0.274}$	0.99	$y = -0.124\ln(x) + 1.2258$	0.90	$y = 0.7066x^{-0.467}$	0.92		
	Wake County	$y = -0.111\ln(x) + 0.7859$	0.99	$y = -0.158\ln(x) + 1.1443$	0.98	$y = 0.7417x^{-0.43}$	0.94		
	Range		0.8–1		0.81–0.99		0.73–0.99		0.88–0.96
	Mean		0.97		0.95		0.90	$y = 0.0030x + 0.0285$	0.92

[a]Linear least-squares regression models for the differences between combined overall and combined kappa accuracies by distance, averaged across landscapes.

[b]Excludes calculations based on 1 m of displacement.

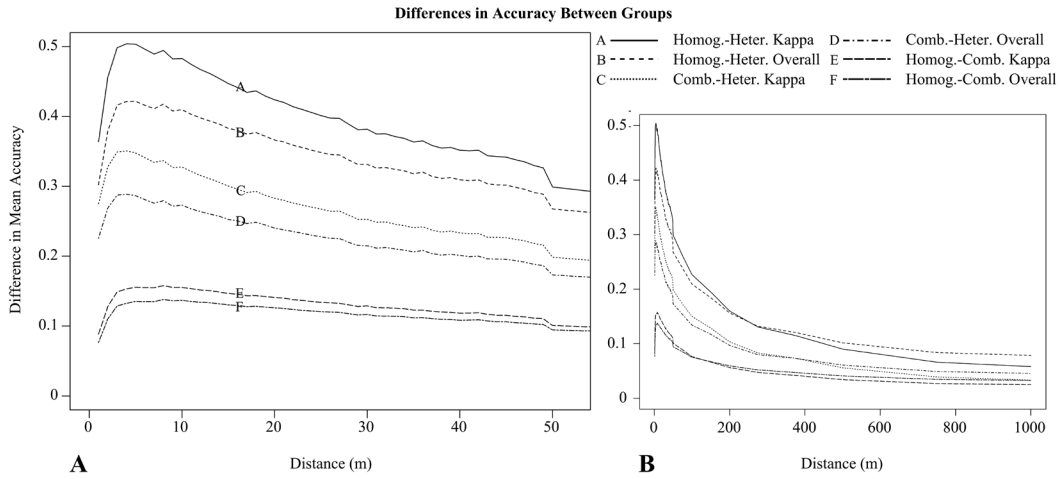

Figure 4. Differences between mean overall and kappa accuracies (across all landscapes) derived from Monte Carlo simulations applied to ground reference plots originating in homogeneous, heterogeneous, or combined 3 × 3 cellular patches for distances 0 to 50 m (A) and 0 to 1,000 m (B).

(r^2). Mean r^2 values (across all landscapes) for the combined dataset were 0.98 (range = 0.92 to 1) and 0.96 (0.80 to 1) for overall and kappa accuracies respectively, and no lower than 0.73 for all models. Importantly, a notable strength of these models (excluding the initial one-time accuracy figures) is that they allow us to invert the original independent and dependent variables so as to estimate our initial, unknown locational error as a function of accuracy. This strategy aims to minimize the effects of locational error and is presented in the discussion.

Relative to our one-time assessment, simulations revealed that when averaging across all landscapes, the homogeneous datasets produced overall accuracies between 13.8 percent (0.158 kappa) and 3.3 percent (0.025) greater than the combined datasets, and between 42.2 percent (0.504) and 7.9 percent (0.058) greater than the heterogeneous datasets (Figure 4). The combined datasets produced overall accuracies between 28.9 percent (0.351) and 4.6 percent (0.033) greater than the heterogeneous datasets. These differences are reflected in the results of MANOVA, which indicated that dataset grouping had a significant effect on accuracy in every case (Figure 5). Post-hoc univariate Tukey HSD contrasts (Table 4) revealed that for both overall and kappa accuracies the heterogeneous datasets had significantly lower accuracies than the combined datasets (–0.208 mean-centered accuracy units (MCAU) for overall; –0.241 MCAU for kappa) and the homogeneous datasets (–0.310 MCAU for overall; –0.353 MCAU for kappa). Similarly, the combined datasets had significantly lower overall and kappa accuracies than the homogeneous datasets (–0.102 MCAU for overall; –0.112 MCAU for kappa). All pairwise comparisons were highly significant when tested at 95 percent confidence (p = <0.0001).

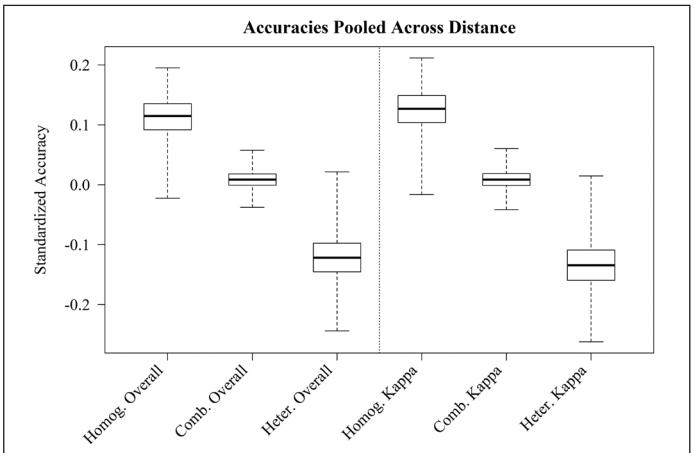

Figure 5. A visual example of MANOVA results from the Ucross, Wyoming study site illustrating mean-centered overall and kappa accuracies for homogeneous, heterogeneous, and combined datasets. Boxes denote the interquartile range, and whiskers cover the full range. Since mean-centering relied on the mean from all groups' overall or kappa accuracies, a direct comparison should not be made between overall and kappa plots. Other landscapes showed similar patterns and all landscapes showed significant differences between groups. See Table 4.

Quantifying Spatial Autocorrelation in Landscape Structure and Composition

When comparing homogeneous and heterogeneous datasets for insight into landscape structure, the results were as predicted. Over short distances (≤50 m) and across all twelve landscapes, the variability in overall and kappa accuracies increased rapidly for homogeneous and, to a lesser degree, combined reference sites (Figure 6). Heterogeneous sites showed a minor increase in variability over distances less than 5 m, followed by a steady decline after 5 m. When considering only short distances (≤50 m), the variability in accuracy for all three types of reference sites appears to asymptote by roughly 25 m, suggesting that across the landscapes that were evaluated, the distribution of landscape structure is relatively uniform at resolutions greater than ≈25 m. When considering the entire range of displacement distances (Figure 7), the variability of all three datasets more clearly asymptotes by roughly 150 m, though that of the heterogeneous sites does not plateau as consistently as do the homogeneous or combined sites. Beyond 150 m a level of dynamic equilibrium is reached, and all six accuracy measures (overall or kappa × 3 datasets) vacillate within a range of 1 percent accuracy through to 1,000 m. With the limited number of distances at which simulations were run beyond 50 m, these longer trends are less clear than those associated with the distances applicable to most research.

With respect to spatial autocorrelation, the results suggest that the vast majority of the landscapes contain similar degrees of interspersion and patch-type adjacencies. Where the variability in overall and kappa accuracies is a function of different error matrices, and the error matrices are a function

Table 4. Univariate Analysis of Variance Contrasts (Tukey Multiple Comparisons of Means Using a 95 Percent Family-Wise Confidence Level) *(continued on next page)*

	Contrast	Accuracy Type	Difference[a]	Lower Bound[a,b]	Upper Bound[a,b]	p (adj)[b]
Baltimore, MD	Heterogeneous-Combined	Overall	-0.220	-0.223	-0.218	0
	Homogeneous-Combined	Overall	0.184	0.181	0.186	0
	Homogeneous-Heterogeneous	Overall	0.404	0.402	0.406	0
	Heterogeneous-Combined	Kappa	-0.291	-0.294	-0.289	0
	Homogeneous-Combined	Kappa	0.207	0.204	0.209	0
	Homogeneous-Heterogeneous	Kappa	0.498	0.495	0.500	0
Bolton, ON	Heterogeneous-Combined	Overall	-0.132	-0.134	-0.130	0
	Homogeneous-Combined	Overall	0.030	0.028	0.032	0
	Homogeneous-Heterogeneous	Overall	0.162	0.160	0.164	0
	Heterogeneous-Combined	Kappa	-0.127	-0.129	-0.125	0
	Homogeneous-Combined	Kappa	0.029	0.027	0.031	0
	Homogeneous-Heterogeneous	Kappa	0.156	0.154	0.158	0
Cumberland, MD	Heterogeneous-Combined	Overall	-0.220	-0.223	-0.218	0
	Homogeneous-Combined	Overall	0.060	0.058	0.062	0
	Homogeneous-Heterogeneous	Overall	0.280	0.278	0.283	0
	Heterogeneous-Combined	Kappa	-0.261	-0.264	-0.258	0
	Homogeneous-Combined	Kappa	0.067	0.064	0.069	0
	Homogeneous-Heterogeneous	Kappa	0.328	0.325	0.330	0
Garden City, ID	Heterogeneous-Combined	Overall	-0.119	-0.121	-0.117	0
	Homogeneous-Combined	Overall	0.064	0.062	0.066	0
	Homogeneous-Heterogeneous	Overall	0.183	0.181	0.185	0
	Heterogeneous-Combined	Kappa	-0.135	-0.138	-0.133	0
	Homogeneous-Combined	Kappa	0.071	0.068	0.073	0
	Homogeneous-Heterogeneous	Kappa	0.206	0.203	0.208	0
Honolulu, HI	Heterogeneous-Combined	Overall	-0.197	-0.199	-0.194	0
	Homogeneous-Combined	Overall	0.024	0.022	0.027	0
	Homogeneous-Heterogeneous	Overall	0.221	0.219	0.224	0
	Heterogeneous-Combined	Kappa	-0.256	-0.259	-0.253	0
	Homogeneous-Combined	Kappa	0.028	0.025	0.031	0
	Homogeneous-Heterogeneous	Kappa	0.284	0.281	0.287	0
Mecklenburg, NC	Heterogeneous-Combined	Overall	-0.240	-0.242	-0.237	0
	Homogeneous-Combined	Overall	0.054	0.051	0.056	0
	Homogeneous-Heterogeneous	Overall	0.293	0.291	0.296	0
	Heterogeneous-Combined	Kappa	-0.259	-0.262	-0.257	0
	Homogeneous-Combined	Kappa	0.059	0.056	0.062	0
	Homogeneous-Heterogeneous	Kappa	0.318	0.316	0.321	0
NYC	Heterogeneous-Combined	Overall	-0.227	-0.229	-0.225	0
	Homogeneous-Combined	Overall	0.053	0.051	0.055	0
	Homogeneous-Heterogeneous	Overall	0.281	0.278	0.283	0
	Heterogeneous-Combined	Kappa	-0.291	-0.293	-0.288	0
	Homogeneous-Combined	Kappa	0.061	0.059	0.064	0
	Homogeneous-Heterogeneous	Kappa	0.352	0.349	0.354	0
Philadelphia, PA	Heterogeneous-Combined	Overall	-0.323	-0.327	-0.320	0
	Homogeneous-Combined	Overall	0.018	0.014	0.021	0
	Homogeneous-Heterogeneous	Overall	0.341	0.338	0.344	0
	Heterogeneous-Combined	Kappa	-0.377	-0.381	-0.373	0
	Homogeneous-Combined	Kappa	0.019	0.015	0.023	0
	Homogeneous-Heterogeneous	Kappa	0.396	0.392	0.400	0
Prince George's County, MD	Heterogeneous-Combined	Overall	-0.278	-0.280	-0.275	0
	Homogeneous-Combined	Overall	0.187	0.185	0.189	0
	Homogeneous-Heterogeneous	Overall	0.465	0.463	0.467	0
	Heterogeneous-Combined	Kappa	-0.311	-0.314	-0.309	0
	Homogeneous-Combined	Kappa	0.208	0.205	0.211	0
	Homogeneous-Heterogeneous	Kappa	0.519	0.517	0.522	0
Ucross, WY	Heterogeneous-Combined	Overall	-0.150	-0.152	-0.148	0
	Homogeneous-Combined	Overall	0.144	0.142	0.146	0
	Homogeneous-Heterogeneous	Overall	0.294	0.292	0.296	0
	Heterogeneous-Combined	Kappa	-0.145	-0.147	-0.143	0
	Homogeneous-Combined	Kappa	0.144	0.142	0.146	0
	Homogeneous-Heterogeneous	Kappa	0.289	0.287	0.291	0

[a]Values are in mean-centered accuracy units.
[b]Based on a 95 percent family-wise confidence level. All 0 values returned during computation as 3.40E-08.

TABLE 4. *(CONTINUED)*

	Contrast	Accuracy Type	Difference[a]	Lower Bound[a,b]	Upper Bound[a,b]	p (adj)[b]
Virginia Beach, VA	Heterogeneous-Combined	Overall	-0.263	-0.265	-0.260	0
	Homogeneous-Combined	Overall	0.320	0.317	0.322	0
	Homogeneous-Heterogeneous	Overall	0.582	0.580	0.585	0
	Heterogeneous-Combined	Kappa	-0.272	-0.275	-0.269	0
	Homogeneous-Combined	Kappa	0.351	0.347	0.354	0
	Homogeneous-Heterogeneous	Kappa	0.623	0.620	0.626	0
Wake County, NC	Heterogeneous-Combined	Overall	-0.139	-0.141	-0.136	0
	Homogeneous-Combined	Overall	0.171	0.168	0.173	0
	Homogeneous-Heterogeneous	Overall	0.309	0.307	0.312	0
	Heterogeneous-Combined	Kappa	-0.214	-0.217	-0.211	0
	Homogeneous-Combined	Kappa	0.198	0.195	0.201	0
	Homogeneous-Heterogeneous	Kappa	0.412	0.409	0.414	0
Summary Statistics (Range)	Heterogeneous-Combined	Overall	-0.32 - -0.12	-0.33 - -0.12	-0.32 - -0.12	0 - 0
	Homogeneous-Combined	Overall	0.02 - 0.32	0.01 - 0.32	0.02 - 0.32	0 - 0
	Homogeneous-Heterogeneous	Overall	0.16 - 0.58	0.16 - 0.58	0.16 - 0.58	0 - 0
	Heterogeneous-Combined	Kappa	-0.38 - -0.13	-0.38 - -0.13	-0.37 - -0.12	0 - 0
	Homogeneous-Combined	Kappa	0.02 - 0.35	0.02 - 0.35	0.02 - 0.35	0 - 0
	Homogeneous-Heterogeneous	Kappa	0.16 - 0.62	0.15 - 0.62	0.16 - 0.63	0 - 0
Summary Statistics (Mean)	Heterogeneous-Combined	Overall	-0.208	-0.210	-0.206	0
	Homogeneous-Combined	Overall	0.102	0.100	0.105	0
	Homogeneous-Heterogeneous	Overall	0.310	0.308	0.312	0
	Heterogeneous-Combined	Kappa	-0.241	-0.244	-0.238	0
	Homogeneous-Combined	Kappa	0.112	0.109	0.115	0
	Homogeneous-Heterogeneous	Kappa	0.353	0.350	0.356	0

[a]Values are in mean-centered accuracy units.
[b]Based on a 95 percent family-wise confidence level. All 0 values returned during computation as 3.40E-08.

of land cover types, it might initially appear that the relative stabilization of variability in map accuracy statistics reflects a stabilization of land cover types. However, variability could remain the same in cases where mis-matched cover types (that contributed to misclassification at a given location) changed relative to other mis-matched cover types across distances. Figure 3 helps to show that the spatial autocorrelation in land cover (not structure) stabilizes initially at ≈200 m and more completely by ≈800 m, though accuracy statistics do not converge to the degree we had expected.

As noted above, the landscapes used in this study were quantified with two structural metrics, of which the Interspersion and Juxtaposition Index (IJI) is robust to changes in spatial resolution. Figure 8 visually illustrates the relationship between overall accuracies for each of the modeled landscapes and their relative IJI values. While these were not evaluated statistically, in our limited sample of 12 landscapes no clear relationship or pattern emerged between IJI values and changes in map accuracy. Several landscapes with lower IJI values exhibited both lower accuracies overall and more rapid declines in accuracy with respect to other landscapes. This might be expected, as lower IJI values imply that patches of a given cover type are themselves more closely positioned in two-dimensional space (i.e., more poorly interspersed). With these conditions we would expect accuracy to decline rapidly as the amount of locational error introduced exceeds the mean patch size (or mean patch size ±1 standard deviation) and creates a pronounced increase in the proportion of sample sites falling in patches of the incorrect cover type. This does not agree with Gu *et al.*'s (2015) findings, which suggest that positional error has a greater effect on accuracy as fragmentation increases and patch size decreases. However, in that our data shows no definitive patterns for a single landscape metric, the interplay between structure, locational error, and accuracy needs further exploration. The relationship between patch density and accuracy is not meaningful since patch density is strongly affected by spatial resolution, though here too no clear patterns emerged.

Discussion

There has been much discussion on the meaning and efficacy of accuracy assessment statistics (Liu *et al.*, 2007; Stehman, 1997). Contrary to what some may believe, the values of common metrics (e.g., overall accuracy Cohen's kappa, [1960; Congalton and Mead, 1983; Congalton *et al.*, 1983], Brennan and Prediger's kappa [1981, Foody, 1992], Ma and Redmond's Tau [1995]) are by no means fixed and are instead a function of what the map producer hopes to assess, how the map user will apply the mapped results, how many land cover classes exist in a given map, and how similar compared landscapes may be, among other variables. Overall, user's and producer's accuracies seem to be useful in most, if not all, circumstances since they have direct probabilistic interpretations with respect to the true landscape cover type structure (Stehman, 1997). Views on Cohen's kappa (1960) and its variants are,

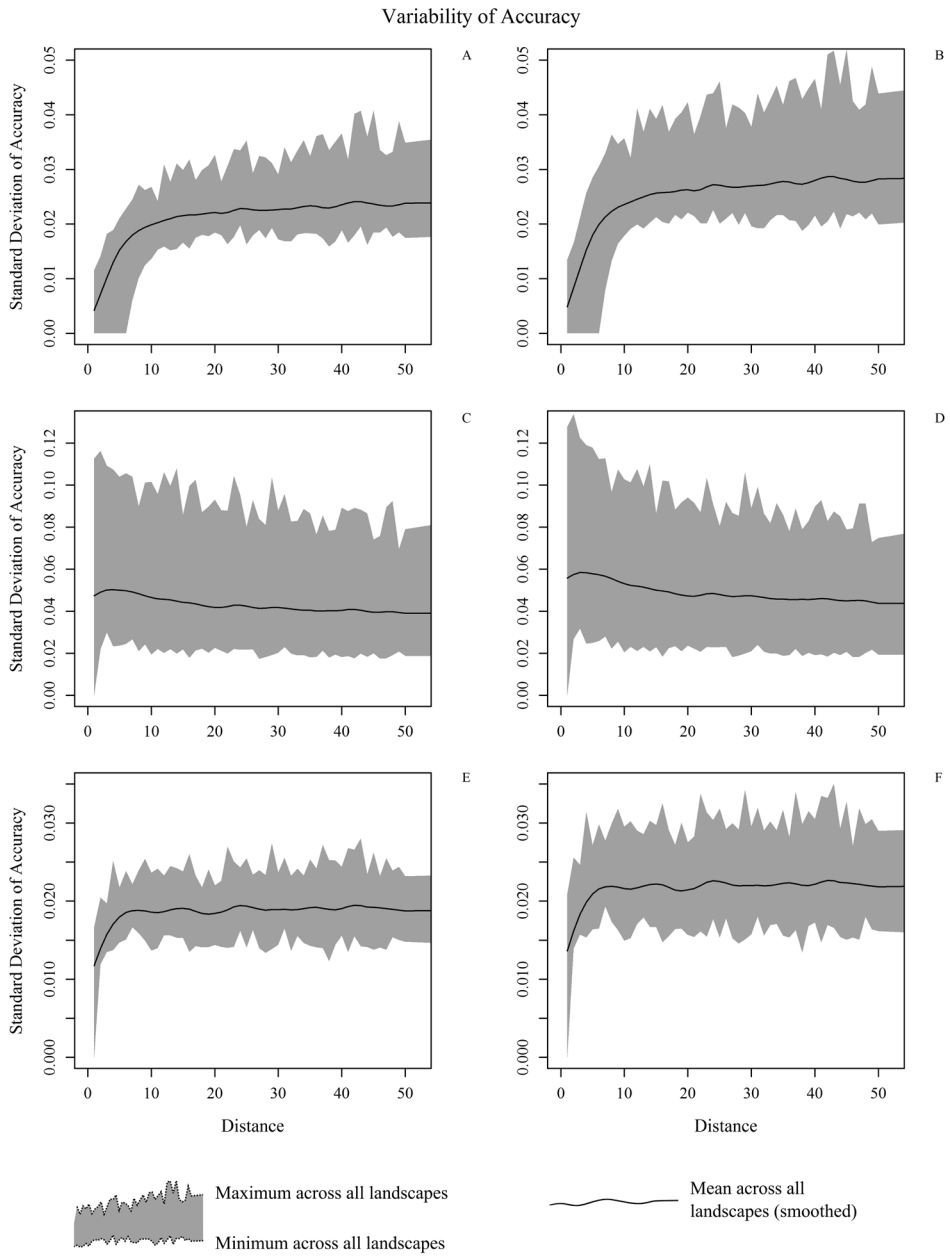

Figure 6. Variability in overall (left column) and kappa (right column) accuracies derived from Monte Carlo simulations applied to ground reference plots originating in homogeneous (A, B), heterogeneous (C, D), or combined (E, F) 3 × 3 cellular patches for distances 0 to 50 m.

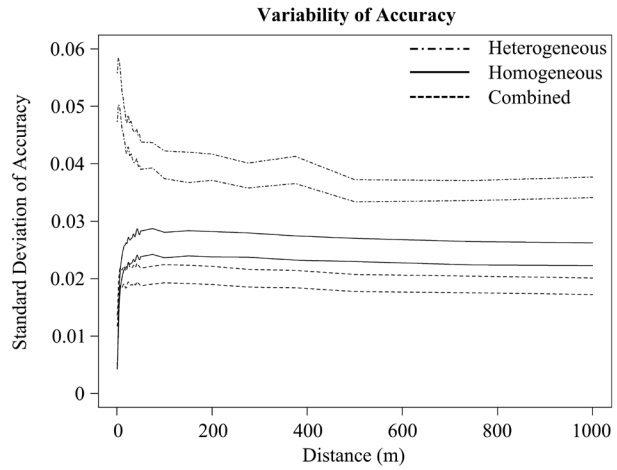

Figure 7. Mean variability in overall (upper line of each pair) and kappa (lower line of each pair) accuracies of all landscapes, as derived from Monte Carlo simulations applied to ground reference plots originating in homogeneous, heterogeneous, or combined 3 × 3 cellular patches for distances 0 to 1,000 m.

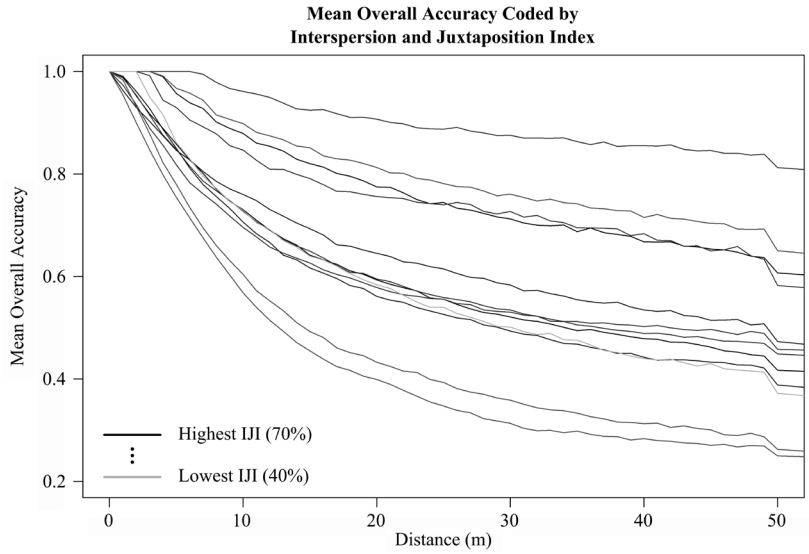

Figure 8. Mean overall accuracy for 12 landscapes as derived from Monte Carlo simulations for distances 0-50 m. Landscape trends are symbolized according to their respective Interspersion and Juxtaposition Index values (0 to 100 percent).

however, more varied (Stehman, 1997; Congalton and Mead, 1983; Congalton *et al.*, 1983; Pontius and Millones, 2011).

The results we present here help to substantiate claims that accuracy assessment metrics may not offer the surety we hope they do. That mean accuracies declined markedly with distance agrees with common sense, the conservative bias Verbyla and Hammond (1995) found when manually imposing a single cell's worth of locational error in each of the cardinal directions, and the trends presented by Gu *et al.* (2015), who performed fuzzy classification on a series of simulated landscapes that varied according to co-registration error and landscape structure. Practically, the analyst may suffer ill effects from horizontal misalignment across all realistically pertinent distances (≤50 m). Where analysts and field technicians are aware of possible bias attributable to the inaccuracies of GPS technology or image georeferencing, they might be inclined to believe that the true accuracy of their categorical map (which remains theoretical since it can only be estimated in light of locational error) is greater than that estimated by standard accuracy assessment. While we consider this circumstance largely true, when holding the unknown co-registration error constant (e.g., 2 m horizontal displacement) the difference in accuracy between the theoretical true accuracy of a map and that which can be obtained during conventional accuracy assessment likely increases with true classification accuracy. That is, the overall and kappa accuracies of a highly accurate categorical map will appear proportionally lower (i.e., worse accuracy) under a fixed amount of co-registration error compared to an inaccurate version of that same map with the same co-registration error. This is because a reduction in true classification accuracy implies increased randomization of cell assignment to informational classes during the classification process, and under increasingly random map structures the rate of accuracy decline slows.

It is obvious that a single comparison of a categorical map with reference information is incapable of shedding light on the variability inherent in the map, and the producer has little sense for what degree of error is attributable to locational error unless he or she tests for it. This is also highlighted by Gu *et al.* (2015), whose simulations and "soft" classification are analogous to the "hard" classification presented here. While a single comparison of statistics derived from homogeneous and combined datasets provides a simple metric for the degree of locational error, the minimum and maximum bounding intervals (across all landscapes) in Figure 6 illustrate how similar figures are possible across wide ranges of misalignment distances. Holding other factors (e.g., true classification accuracy) constant, as overall and kappa accuracies decline we become less confident in the degree of positional error.

The interpretation of both overall and kappa accuracies becomes difficult under the simulated conditions we present here. What does it really mean to have 70 percent overall accuracy when every predicted data point had a spatially randomly allocated cover type? Are we to interpret this to mean that, for a given distance, the baseline randomness which kappa professes to eliminate may be as high as 70 percent? The interpretation of kappa accuracy is further complicated. Where the standard kappa (Cohen 1960) "attempts to account for the expected agreement due to random spatial reallocation of the categories in the comparison map, given the proportions of the categories in the comparison and reference maps, regardless of the size of the quantity disagreement" (Pontius and Millones 2011: p. 4414), we might expect the simulated kappa accuracies to hover around overall accuracies since, at first glance, it appears that each iteration of our simulations performed the spatial reallocation kappa corrects for.

However, this disagreement between the expected accuracy inherent in kappa's computation and the expected accuracy that is generated by random sampling at various locational error distances is, in part, due to the fact that the former allows for spatial reallocation of classes from the entire map while the latter effectively reduces the map to contain only those classes and class proportions contained by a given set of annuli, such that each simulation of our study was not truly random across the entire mapped domain. More fundamentally, kappa is a function of overall accuracy and mathematically the two can only unify when overall accuracy reaches 100 percent or when the expected accuracy in kappa's computation is 0 percent, at which point both will be 0 percent. The latter can occur when the base error matrix is populated such that row and column marginals produce no values in the trace cells of kappa's cross-product matrix. That a random sample for a given locational error distance can produce potentially meaningful levels of accuracy is a function of landscape structure and makes cross-map accuracy comparisons suspect.

Locational Error-Corrected Accuracy Statistics

At all distances, ground reference plots falling in homogeneous 3 × 3 cell patches yielded statistically greater accuracies. This result further confirms the optimistic bias about which other authors have issued warning (Hammond and Verbyla 1996; Plourde and Congalton 2003; Stehman and Czaplewski 1998; Stehman and Foody 2009). Homogeneous patches neither accurately reflect the landscape of study nor uphold core assumptions of probability sampling, and should be avoided if possible. In light of our findings, we offer several possible strategies to better estimate the accuracy of a classifier while mitigating the effects of locational error.

Patch-based assessment methods (e.g., Gopal and Woodcock 1994; Hagen 2003; Power et al 2001) provide one useful alternative for estimating map accuracy, given that while error is not entirely eliminated the cell-based assessment strategy (i.e., using a single cell as the primary sampling unit)

can be abandoned. These approaches usually rely on fuzzy set theory, which can often provide the analyst with a more intuitive means of categorizing real world gradations in cover types. However, at least some fuzzy classifications appear to be similarly sensitive to positional error as are the "hard" classifications presented here (Gu *et al.* 2015), and the effects of locational error are more masked than eliminated.

With very high resolution imagery available for many locations, photointerpretation (Klöditz *et al.* 1998; Seto 2002; Zhang and Huang 2010) is another promising route. Provided that the temporal periods of the analyzed and interpreted images coincide, and that the co-registration between images is not greater than the combined GPS and analyzed image error, photointerpretation of reference sites may offer some benefits over traditional field campaigns. Such a strategy can be particularly alluring since it has the potential to drastically reduce both time spent in the field and overall sampling error through the use of a greater number of reference sites than could be visited in-person. As image georeferencing error, orthorectification error, and return time improve, photointerpretation may see increased attention.

Ideally, technicians should employ the strategy noted above, of utilizing GPS technology and georeferenced imagery whose combined root mean squared horizontal error is less than one-half the length of the shortest side of a pixel. This guarantees accurate within-pixel field placement and eliminates potential locational error. However, where imagery is never perfectly accurate and sub-meter GPS technology is still expensive, we recognize that this may be an unrealistic strategy in many, if not most, circumstances.

The models we present above provide a new method for obtaining map accuracy statistics that are minimally influenced by the effects of locational error. Using overall accuracy and base data from Ucross, Wyoming as an example, the procedure is as follows:

1. A field campaign (or photointerpretation) is carried out for basic field assessment and to compute initial measures of accuracy (here, overall accuracy). This will be referred to as the "one-time" accuracy, since field campaigns are generally conducted only once, providing a single set of reference data for accuracy assessment. For Ucross, WY our one-time overall accuracy was 76.8 percent.
2. Independent of the one-time accuracy assessment, the classified map is subjected to the Monte Carlo simulated sampling protocols detailed above. Through the simulations we engage the central limit theorem and obtain a more accurate estimate of the mean overall accuracy at each introduced locational error distance.
3. The mean overall accuracy at each locational error distance is used to build a piece-wise function for modeling map accuracy. Presuming the map contains square pixels, the function (Equation 1) evaluates to an accuracy of 1 (100 percent accurate) for error distances ≤0.5 pixel width. At all other error distances, the function relies on a non-linear parametric model and evaluates to a landscape-dependent accuracy of <1 (Equation 2).

$$f(x) = \begin{cases} 1, if\, x \le h_w \\ f_2(x), if\, x > h_w \end{cases}$$ Where x = error distance, h_w = one-half the pixel width, and $f_2(x)$ = landscape-specific non-linear model. (1)

$$f_2(x) = 0.7857x^{-0.17}$$ for Ucross, WY, error distances ≤15 m (2)

4. One can utilize the non-linear portion of Equation 1 to obtain the maximum predicted accuracy that is possible, but heretofore unknown, with a given classified map. The overall accuracies of landscapes we analyzed (with the combined datasets) in the present work produced models with exceptional predictive strength (Table 3) with a mean r^2 of 0.98. The model for overall accuracy built from simulations applied to the Ucross, WY landscape (Equation 2) produced an r^2 of 0.99. It is this model that is used to estimate the actual locational error inherent in the one-time accuracy figures.
5. Referring to Figure 9, the maximum accuracy that is attainable with a given map will always be infinitesimally close to where the error distance = 0.5 pixel width (Equation 3). This is largest distance for which an accuracy assessment will always compute to 100 percent accurate, since for all distances ≤0.5 pixel width the

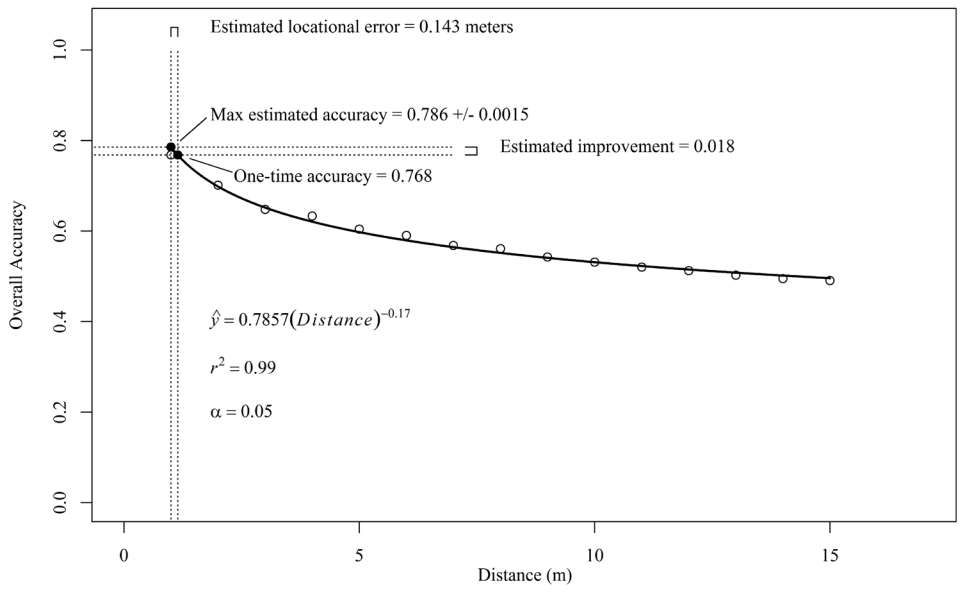

Figure 9. Maximum estimated overall accuracy with 95 percent confidence interval and associated locational error for the Ucross, WY landscape. Values are estimated using a non-linear function (Equation 2) that captures the decline in accuracy as locational error increases (up to 15 m).

accuracy assessment compares the pixel with itself. In the example shown, the maximum predicted overall accuracy for Ucross, WY occurs at 1m (given 2 m pixels) and was 0.786 ±0.0015 where $\alpha = 0.05$ and $r^2 = 0.99$. This range of values (78.5 percent to 78.8 percent) contains the "true" accuracy of our classifier in the absence of locational error in 95 percent of theoretical samples, contingent on predictive model strength (i.e., $r^2 = 0.99$). The difference between 100 percent accuracy and ~79 percent accuracy is attributable to poor classification.

$$maximum\ accuracy = \lim_{x \to h_w^+} f(x) \quad \text{where } f(x) = \text{Equation 1} \qquad (3)$$

6. Next, the one-time accuracy is fit to the modeled trend using the non-linear model (Equation 2 for Ucross, WY). For this value there is no margin of error or confidence interval since it is the result of a one-time accuracy assessment.
7. The theoretical improvement in accuracy that is possible with a given classification can be computed as the maximum predicted accuracy less the one-time accuracy. In the example, this is an estimated improvement of 0.018. This value is subject to the margin of error on the maximum predicted accuracy, making the possible improvement in accuracy that was possible in the Ucross, WY example 0.018 ± 0.0015.
8. In addition to estimating the maximum accuracy or improvement in accuracy that is attainable with a given classification, one can also estimate the linear measure of mean locational error that biases the one-time accuracy assessment. This is obtained by first inverting Equation 2 and substituting in the one-time accuracy to solve for the covariate value (error distance). Then, one subtracts the error distance associated with 0.5 pixel width from the predicted error distance of the one-time accuracy. In the example given, this distance is 0.143 meters. In other words, for the Ucross, WY landscape we had an average of ~15 cm of locational error that could theoretically have been removed to improve our accuracy by another ~2 percent.

Using this strategy, the estimates of map accuracy and locational error are contingent upon (a) the explanatory strength of the non-linear model, and (b) upon the range of error distances over which changes in accuracy are modeled. With respect to (a), as noted above, the non-linear models across the range of landscapes we have evaluated demonstrate r^2 values very close to 1. In the empirical example illustrated, we are highly confident in the maximum potential accuracy we could have obtained and the locational error correction that would have been needed to correct the bias inherent in the field campaign, as $r^2 = 0.99$. In our example, the maximum potential accuracy is not corrected for random chance, though the approach we detail here is just as easily applied to any confusion matrix-based metric, kappa or otherwise.

With respect to (b), the range of error distances over which a curve is fitted influences the estimated maximum potential accuracy and locational error for a given map. As the modeled distance increases, the (quasi) asymptotic right tail exerts proportionally greater influence over the curve, forcing the trend line to overestimate mean accuracy at low (i.e., meaningful) error distances relative to comparable models relying on fewer error distances. Conceptually, this consequence shifts our initial uncertainty in map accuracy away from the classification and towards the equipment or model used to collect the field data, effectively raising a new question: over what error distances should we build our models? While the model fitted to the full complement of tested error distances (here: 1,000 m) is strong enough for inference, in practice a much smaller range of distances should be used. In the example presented above, we use error distances ≤15 m, reflecting the typical ±15 m positional accuracy of our civilian-grade Garmin GPS equipment in the absence of differential GPS and Wide Area Augmentation System correction. We recommend that field campaign managers explicitly measure the circular error probable (CEP) of their field equipment at numerous locations on the landscape. The maximum and minimum 95 percent CEP values reflect the range of error distances over which predictive models should be built. These two values produce the most liberal (higher) and conservative (lower) estimates of maximum potential accuracy and locational error, respectively. In this way, the researcher obtains a realistic range of potential maximum accuracies and their implicit locational errors, providing a mechanism for understanding the magnitude of potential accuracy improvement.

As this approach is expanded in the future, we recommend performing the Monte Carlo simulations across a continuous range of error distances. Instead of sampling in discrete annuli as we have done here, one might create a single sampling domain bounded by (a) half pixel width, and (b) a single maximum distance based on the largest 95 percent CEP for a given landscape and the technology used during the field campaign. This approach will ensure that there is a sufficient quantity of data for relevant horizontal displacement distances to permit robust model fitting.

The accuracy statistics and non-linear trends we present are tied to the modeled landscape structures and do not necessarily apply to any other datasets. They speak more broadly, however, to the importance of landscape structure in accuracy assessment and call for an explicit integration of structure into accuracy assessment metrics. Pontius and Millones (2011) provide two metrics (quantity disagreement and allocation disagreement) that provide an easy means of taking both quantity and structure into account when using error matrices for accuracy assessment. It may be worth pursuing the integration of other spatially explicit structural metrics into accuracy assessment, for there are a large number (e.g., McGarigal, 2014) that could help the "age of the error matrix" (Congalton and Green, 2009) continue to mature. In the meantime, the approach we detail here can be applied to historical data to re-evaluate old map accuracies.

Conclusions

High-resolution imagery promises extraordinary levels of detail on land cover and landscape structure but poses unique challenges relative to coarser grained counterparts. Per-pixel accuracy assessment of high-resolution data may suffer from locational misalignment between real world locations and where we perceive those locations to be in digital space. Even with highly accurate global positioning system equipment, there is likely to be some degree of horizontal displacement between ground referenced samples and their corresponding locations on imagery and classified maps. Our Monte Carlo simulations show that this was the case for twelve datasets from across the U.S. and southern Canada, with substantial changes in mean inter-rater accuracy statistics over small distances relative to our pixel size. The use of reference samples falling solely in small homogeneous land cover patches to remove the effects of locational misalignment is not a sound practice, and we recommend that it be avoided unless it is used to specifically quantify the possible loss in accuracy attributable to horizontal locational error. Homogeneous sites provide significantly greater accuracies compared with reference samples falling in heterogeneous or combined sites for distances up to 1,000m (Figures 3 and 4). Further,

the variability of inter-rater statistics for homogeneous sites changes with horizontal distance, providing additional indication that using a sub-sample of ground reference sites is inappropriate. With regard to variability in combined reference sites over short distances, it is clear that a one-time accuracy assessment may yield unpredictable results (Figure 6).

We offer a new strategy for obtaining locational error-corrected map accuracy statistics that depends on non-linear models derived through Monte Carlo-based sampling simulations. With respect to traditional pixel-based one-time accuracy assessment, our approach can help researchers better identify the maximum theoretical accuracy that is possible with a given classified map, as well as the horizontal locational error inherent in statistics from one-time map accuracy assessment.

As the field progresses, we see three potentially fruitful areas of research: (1) The direct inclusion of structural or contextual landscape metrics in accuracy assessment is both a logical and useful direction to explore (e.g., Pontius and Millones, 2011). (2) It is time to further consider the need for accuracy assessment metrics tailored to specific landscape properties. Unlike the comparison of two maps from precisely the same landscape, it may not make sense to compare categorical maps from different bioregions with different structural characteristics (e.g., patch densities), or constructed at different resolutions. The results we present here suggest that cross-landscape comparisons may be acceptable, but more detailed investigation is needed. (3) Where map production usually occurs in response to specific planning, research, or management objectives, we should lend thought to accuracy assessment strategies that embrace the underlying goals leading to their production. Maps are made for a wide variety of reasons, and their relative utility is governed less by a single overall or kappa accuracy value than by what the producers or end users initially set out to obtain. As we collectively move to explore these and other facets of thematic accuracy assessment, we encourage analysts to use caution when treating high-resolution imagery as point sampled data and when applying error matrix-based accuracy assessment techniques to any categorical map.

Acknowledgments

We gratefully acknowledge the Ucross Foundation, Apache Foundation, and Bauer Land and Livestock for providing assistance during our field campaign; and R. Congalton, as well as several anonymous reviewers, for their constructive feedback on the manuscript. This research would not have been possible without the support of the Ucross High Plains Stewardship Initiative.

References

ASPRS - American Society for Photogrammetry and Remote Sensing, 2015. New standard for new era: Overview of the 2015 ASPRS Positional Accuracy Standards for Digital Geospatial Data, *Photogrammetric Engineering & Remote Sensing*, 81(3):173–176.

Bauer, M., T. Burk, A. Ek, P. Coppin, S. Lime, T. Walsh, D. Walters, W. Befort, and D. Heinzen, 1994. Satellite inventory of Minnesota forest resources, *Photogrammetric Engineering & Remote Sensing*, 60(3):287–298.

Brennan, R.L., and D.J. Prediger, 1981. Coefficient Kappa: Some uses, misuses, and alternatives, *Educational and Psychological Measurement*, 41:687–699.

Brioch, M., S. Stehman, M. Hansen, P. Potapov, and Y. Shimabukuro. 2009. A comparison of sampling designs for estimating deforestation from Landsat imagery: A case study of the Brazilian Legal Amazon, *Remote Sensing of Environment*, 113(11):2448–2454.

Campbell, J., and R. Wynne, 2011. *Introduction to Remote Sensing* , Fifth edition, The Guilford Press, New York, 667 p.

Carmel, Y., D.J. Dean, and C.H. Flather, 2001. Combining location and classification error sources for estimating multi-temporal database accuracy, *Photogrammetric Engineering & Remote Sensing*, 67(7):865–872.

Carleer, A.P., and E. Wolff, 2005. Assessment of very high spatial resolution satellite image segmentations, *Photogrammetric Engineering & Remote Sensing*, 71(11):1285–1294.

Carter, N., 2013. *An Assessment of WorldView-2 Imagery for Classification of a Mixed Deciduous Forest*, Master's thesis, Rochester Institute of Technology, Rochester, New York, ProQuest LLC: Ann Arbor, Michigan, 53 p.

Cohen, J., 1960. A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20:37–46.

Congalton, R.G., 1994. Accuracy assessment of remotely sensed data: Future needs and directions, *Proceedings of the Pecora 12 Symposium: Land Information from Space-Based Systems*, 24-26 August 1993, Sioux Falls, South Dakota, ASPRS, Bethesda, Maryland, pp. 383–388.

Congalton, R., and K. Green, 2009. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, Second edition, CRC Press, Boca Raton, Florida, 177 p.

Congalton, R.G., and R.A. Mead, 1983. A quantitative method to test for consistency and correctness in photo-interpretation, *Photogrammetric Engineering & Remote Sensing*, 49(1):69–74.

Congalton, R., R. Oderwald, and R. Mead, 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques, *Photogrammetric Engineering & Remote Sensing*, 49(12):1671–1678.

Corry, R.C., and R. Lafortezza, 2007. Sensitivity of landscape measurements to changing grain size for fine-scale design and management, *Landscape and Ecological Engineering*, 3:47–53.

Elsharkawy, A., M. Elhabiby, and N. El-Sheimy, 2012. Improvement in the detection of land cover classes using the WorldView-2 imagery, *Proceedings of the American Society for Photogrammetry and Remote Sensing 2012 Annual Conference*, 19-23 March, Sacramento, California, pp. 1–11.

FGDC (Federal Geographic Data Committee, Subcommittee for Base Cartographic Data), *1998. Geospatial Positioning Accuracy Standards, Part 3: National Standard for Spatial Data Accuracy, FGDC-STD_007.3-1998*, URL: *https://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3*, FGDC, Reston, Virginia (last date accessed: 22 August 2016).

Foody, G.M, 1992. On the compensation for chance agreement in image classification accuracy assessment, *Photogrammetric Engineering & Remote Sensing*, 58(10):1459–1460.

Foody. G.M, 2002. Status of land cover classification accuracy assessment, *Remote Sensing of Environment*, 80:185–201.

Gopal, S., and C. Woodcock, 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets, *Photogrammetric Engineering & Remote Sensing*, 60(2):181–188.

Graham, S., 1999, *Remote Sensing Accomplishments*, URL: *http://earthobservatory.nasa.gov/Features/RemoteSensing/*, NASA, Greenbelt, Maryland (last date accessed: 22 August 2016).

Gregoire, T., and H. Valentine, 2008. *Sampling Strategies for Natural Resources and the Environment*, Chapman and Hall/CRC, Boca Raton, Florida, 474 p.

Grewal, M.S., L.R. Weill, and A.P. Andrews, 2001. *Global Positioning Systems, Inertial Navigation, and Integration*, John Wiley and Sons, New York, 416 p.

Gu, J., R.G. Congalton, and Y. Pan, 2015. The impact of positional errors on soft classification accuracy assessment: A simulation analysis, *Remote Sensing*, 7:579–599.

Hagen, A. 2003. Fuzzy set approach to assessing similarity of categorical maps, *International Journal of Geographical Information Science*, 17(3): 235-249

Hammond, T.O., and D.L. Verbyla, 1996. Optimistic bias in classification accuracy assessment, *International Journal of Remote Sensing*, 17(6):1261–1266.

Immitzer, M., C. Atzberger, and T. Koukal, 2012. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 data, *Remote Sensing*, 4:2661–2693.

Janssen, L., and F. van der Wel, 1994. Accuracy assessment of satellite derived land-cover data: A review, *Photogrammetric Engineering & Remote Sensing*, 60(3)419–426.

Kiang, R, 2002. Utilizing spatial features in classifying high-resolution imagery data, Algorithms and technologies for multispectral, hyperspectral, and ultraspectral imagery VIII, *Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), Vol. 4725* (S. Shen and P. Lewis, editors), 01 April, Orlando, Florida, 622 p.

Klöditz, C., A. Boxtel, E. Carfagna, and W. van Deursen, 1998. Estimating the accuracy of coarse scale classification using high scale information, *Photogrammetric Engineering & Remote Sensing*, 64(2):127–133.

Lillesand, T.M., and R.W. Keifer, 1999. *Remote Sensing and Image Interpretation*, Fourth edition, John Wiley & Sons, New York, 760 p.

Liu, C., P. Frazier, and L. Kumar, 2007. Comparative assessment of the measures of thematic classification accuracy, *Remote Sensing of Environment*, 107:606–616.

Liu, W., and F. Yamazaki, 2012. Object-based shadow extraction and correction of high-resolution optical satellite images, *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(4):1296–1302.

Ma, Z., and R.L. Redmond, 1995. Tau coefficients for accuracy assessment of classification of remote sensing data, *Photogrammetric Engineering & Remote Sensing*, 61(4):435–439.

McGarigal, K. 2015. *FRAGSTATS Help*, URL: *http://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf*, (last date accessed: 22 August 2016).

McGarigal, K., and B.J. Marks, 1995. *FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure,* General Technical Report PNW-GTR-351, USDA Forest Service, Pacific Northwest Research Station, Portland, Oregon, 134 p.

McGarigal, K., S.A. Cushman, and E. Ene, 2012. *FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps*, Software produced at the University of Massachusetts, Amherst, URL: *http://www.umass.edu/landeco/research/fragstats/fragstats.html* (last date accessed: 22 August 2016).

Plourde, L., and R. Congalton, 2003. Sampling method and sample placement: How do they affect the accuracy of remotely sensed maps?, *Photogrammetric Engineering & Remote Sensing*, 69(3):289–297.

O'Neil-Dunne, J.P.M., S.W. MacFaden, A.R. Royar, and K.C. Pelletier, 2013. An object-based system for LiDAR data fusion and feature extraction, *Geocarto International*, 28:227– 242.

O'Neil-Dunne, J., S. MacFaden, and A.R. Royar, 2014. A versatile, production-oriented approach to high-resolution tree-canopy mapping in urban and suburban landscapes using GEOBIA and data fusion, *Remote Sensing*, 6:12837–12865.

Perkins, T., S.M. Adler-Golden, P. Cappelaere, and D. Mandl, 2012. High-speed atmospheric correction for spectral image processing, *Proceedings of SPIE: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII, Vol. 8390* (S.S. Shen, and P.E. Lewis, editors), SPIE, Baltimore, Maryland, pp. 1–7.

Pontius, G., Jr., and M. Millones, 2011. Death to Kappa: Birth to quantity disagreement and allocation disagreement for accuracy assessment, *International Journal of Remote Sensing*, 32(15):4407–4429.

Power, C., A. Simms, and R. White. 2001. Hierarchical fuzzy pattern matching for the regional comparison of land use maps, *International Journal of Geographical Information Science*, 15(1): 77-100.

Seto, K., C. Woodcock, C. Song, J. Huang, and R. Kaufmann, 2002. Monitoring land-use change in the Pearl River Delta using Landsat TM, *International Journal of Remote Sensing*, 23(10):1985–2004.

Stehman, S.V., 1996. Estimating the kappa coefficient and its variance under stratified random sampling, *Photogrammetric Engineering & Remote Sensing*, 62(4):401–402.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy, *Remote Sensing of Environment*, 62:77–89.

Stehman, S.V., and R.L. Czaplewski, 1997. Basic structures of a statistically rigorous thematic accuracy assessment, *Proceedings of the American Society of Photogrammetry & Remote Sensing Annual Conference*, 3:543–553.

Stehman, S.V., and R.L. Czaplewski, 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles, *Remote Sensing of Environment*, 64:331–344.

Stehman, S.V., and G.M. Foody, 2009. Accuracy assessment, *The SAGE Handbook of Remote Sensing* (T. Warner, M. Nellis, and G. Foody, editors), SAGE Publications, London, UK, pp. 297–311.

Story, M., and R.G. Congalton, 1986. Accuracy assessment: A user's perspective, *Photogrammetric Engineering & Remote Sensing*, 52(3):397–399.

Ünsalan, C., and K. Boyer, 2004. Classifying land development in high-resolution satellite imagery using hybrid structural-multispectral features, *Transactions on Geoscience and Remote Sensing*, 42(12):2840–2850.

Verbyla, D.L., and T.O. Hammond, 1995. Conservative bias in classification accuracy assessment due to pixel-by-pixel comparison of classified images with reference grids, *International Journal of Remote Sensing*, 16:581–587.

Weber, K., 2006. Challenges of integrating geospatial technologies into rangeland research and management, *Rangeland Ecology & Management*, 59:38–43.

Wolf, A., 2012. Using Worldview-2 Vis-NIR multispectral imagery to support land mapping and feature extraction using normalized difference index ratios, *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII, Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), Vol. 8390* (S. Shen and P. Lewis, editors), 23 April, Baltimore, Maryland, pp. 1–8.

Yu, Q., P. Gong, N. Clinton, G. Biging, M. Kelly, and D. Schirokauer, 2006. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery, *Photogrammetric Engineering & Remote Sensing*, 72(7):799–811.

Zhang, H., and W. Huang, 2010. Accuracy assessment of coastal zone remote sensing survey based on high-resolution remote sensing image, *Earth Resources and Environmental Remote Sensing/GIS Applications, Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE), Vol. 7831* (U. Michel, and D. Civco, editors), 20 September, Toulouse, France, pp.1–7.

(Received 26 February 2016; accepted 26 May 2016; final version 03 June 2016)