Remote Sensing of Environment 128 (2013) 176–185

Contents lists available at SciVerse ScienceDirect

Remote Sensing of Environment

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

An accuracy assessment of forest disturbance mapping in the western Great Lakes

P.L. Zimmerman [a,b], I.W. Housman [c], C.H. Perry [a,*], R.A. Chastain [c], J.B. Webb [c], M.V. Finco [c]

[a] USDA Forest Service, Northern Research Station, St. Paul, MN, USA
[b] University of Minnesota, School of Statistics, Minneapolis, MN, USA
[c] USDA Forest Service, Remote Sensing Applications Center, Salt Lake City, UT, USA

ARTICLE INFO

Article history:
Received 22 December 2011
Received in revised form 21 September 2012
Accepted 22 September 2012
Available online 8 November 2012

Keywords:
Accuracy assessment
Land cover map
Forest
Land cover change
Landsat time series stacks

ABSTRACT

The increasing availability of satellite imagery has spurred the production of thematic land cover maps based on satellite data. These maps are more valuable to the scientific community and land managers when the accuracy of their classifications has been assessed. Here, we assessed the accuracy of a map of forest disturbance in the watersheds of Lake Superior and Lake Michigan based on an improved version of the Vegetation Change Tracker algorithm (VCTw). We constructed a probability-based sampling design using two stages of sampling with stratification at each stage. Results are presented for the portion of the map within the U.S. as well as separately for the U.S. portion of Lake Superior's watershed and for Lake Michigan's watershed. We also present estimates and standard errors of the percent cover for each land cover class that incorporate both the map's data and our sample data. The overall accuracy for the U.S. portion of the map is estimated to be 91% with a standard error of 0.8%. We discuss the relative strengths of the VCTw algorithm as well as the dependence of such an algorithm's success on the characteristics of the landscape being mapped.

Published by Elsevier Inc.

1. Introduction

The Great Lakes Restoration Initiative (GLRI) is a multi-year and multi-agency investment in the improved health of the Great Lakes coordinated by the US Environmental Protection Agency (EPA). The Great Lakes watersheds are approximately 54% forested, and the wise management of forest in watersheds has long been identified as critical to the maintenance of high quality streamflow (Gregory et al., 1991; Karr & Schlosser, 1978; Naiman & Bilby, 1998; Peterjohn & Correll, 1984; Sweeney, 1992). The USDA Forest Service serves as a partner in the GLRI, with a particular focus on investigating the relationship between land management in watersheds and the health of the Great Lakes.

Recent publications provide information on forest land uses and land use change in Michigan (Pugh et al., 2009), Wisconsin (Perry et al., 2008), and Minnesota (Miles et al., 2011), focusing on states as analysis units, but actual management occurs on an owner-by-owner basis within states. Forest restoration and management activities will be more effective at achieving identified water quality objectives if implemented through a watershed perspective. Additionally, the existing data are not designed to provide insight on the watershed-level spatial patterns inherent in forest ecology and management.

As an alternative to state-level analysis, Landsat Time Series Stacks (LTSS) and the Vegetation Change Tracker (VCT) algorithm (Huang et al., 2010) can be used to map forest cover and disturbance at a spatial scale effective for informing watershed-level management. We have adapted the VCT process for the watersheds of Lake Superior and Lake Michigan using a surrogate source of winter images to reduce commission errors in the predictions of forest disturbance classes. The result is a product we call VCTw, which is used to map forest disturbance and described in detail by Stueve et al. (2011) (Fig. 1).

Basing land management decisions on a land-cover map will always be more defensible if a proper accuracy assessment has been performed. In fact, an accuracy assessment of land-cover classifications is considered to be a necessary part of the publication of any such map (Cihlar, 2000). References exist in the literature to help plan such accuracy assessments (Congalton & Green, 1999; Stehman, 2009a; Stehman & Czaplewski, 1998; Strahler et al., 2006), but tailoring an assessment to a particular map can still be challenging; what constitutes an effective assessment of one map may be a poor assessment of another. Some examples of specially-tailored assessments are those by Nusser and Klaas (2003) and Stehman et al. (2003). Here, we describe our assessment procedure and report results for the portion of the map within the U.S. of the forest disturbance map produced by VCTw.[1] Additionally, we present improved estimates of the percent cover of each disturbance class that take advantage of the sample data collected in the accuracy assessment.

* Corresponding author at: 1992 Folwell Avenue, St. Paul, MN 55108, USA. Tel.: +1 651 649 5191; fax: +1 651 649 5140
E-mail address: charleshperry@fs.fed.us (C.H. Perry).

0034-4257/$ – see front matter. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.rse.2012.09.017

[1] Due to limited availability of data, the Canadian portion of the map was considered in a separate, later assessment.

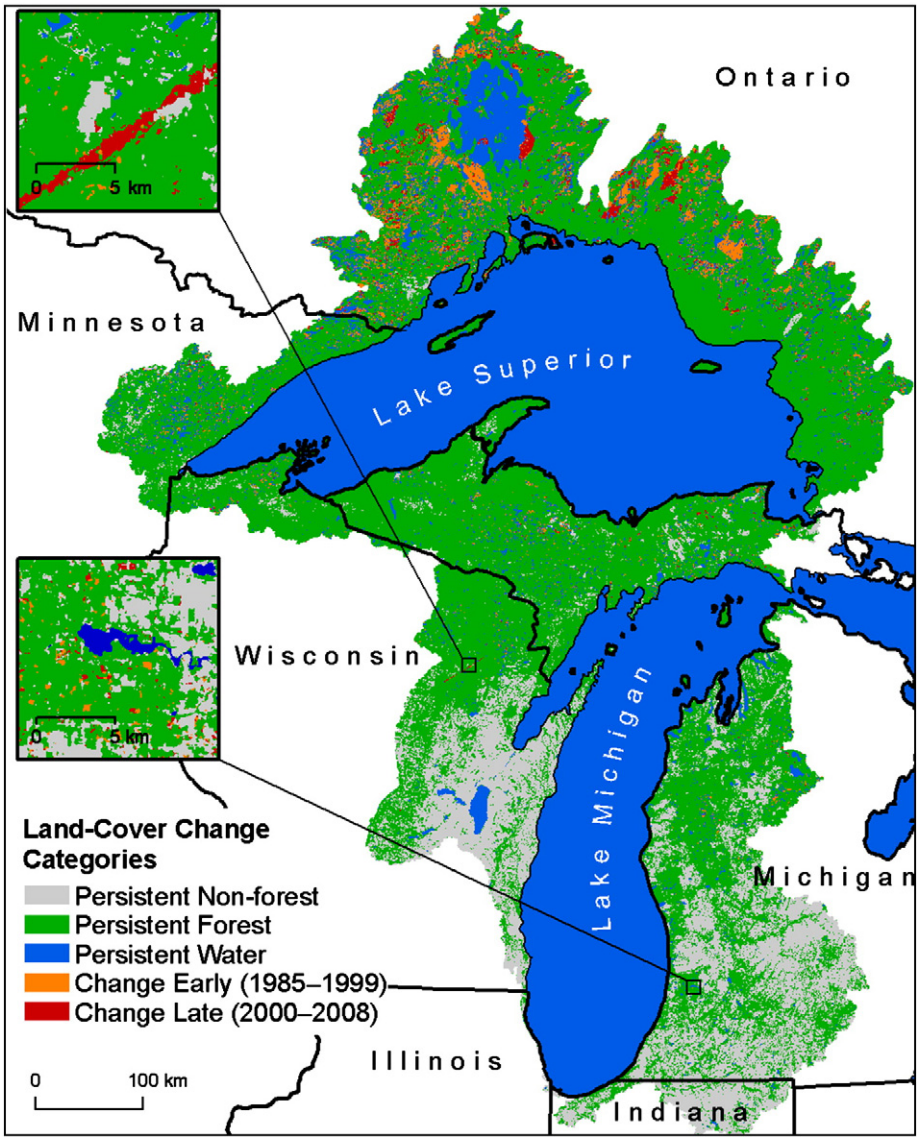

Fig. 1. Forest disturbance map produced by VCTw.

2. Methods

The accuracy assessment was conducted by comparing the forest disturbance map data to "reference data" for a sample of map pixels. By reference data, we mean disturbance classes that were assigned by photointerpreters after inspecting high-resolution imagery (i.e., a proxy for "ground truth"). In this section, we will briefly review the creation of the VCTw data, explain how the sample of map pixels was selected, and how the reference data were obtained for the sample.

2.1. Disturbance mapping

Although VCT has been described in several previous papers (Huang et al., 2009, 2010), it may be helpful to briefly describe the more recently developed VCTw. A detailed description is available in Stueve et al. (2011).

The standard VCT protocol was used to process 36 Landsat path/rows that intersect the Lake Superior and Lake Michigan basins, but the resulting forest disturbance map appeared to contain large areas of incorrectly classified persisting forest and forest disturbance. After reviewing some apparent mapping errors, it was hypothesized that the errors were due to the spectral inseparability between forest and both leafy cultivated crops and emergent herbaceous wetlands. Because cropland and wetlands are highly variable across time, they were also often classified as forest disturbance. This motivated the development of the winter Landsat Time Series Stack (LTSSw) false positive mitigation technique (Stueve et al., 2011). For each path/row, a LTSSw was constructed with roughly quadrennial imagery with the key requirement that the entire image was snow-covered. VCT's cloud masking model was then applied to identify all non-forested, snow-covered areas. Any pixel that was classified as a cloud or cloud edge throughout every mask in the LTSSw, and did not demonstrate a long-term recovery trend was included in a non-forest mask. This mask successfully eliminated a substantial portion of classification errors committed by the standard VCT operating procedure.

A minimum mapping unit (MMU) of 0.356 ha (4 contiguous pixels) was applied to all masked VCT disturbance year outputs. Afterward, the classes initially assigned by VCTw were collapsed into broader classes based on the temporal resolution of available reference data (see Section 2.3.1). Specifically, the typical biennial disturbance classes used by VCT (e.g., "disturbed between 1988 and 1989") were collapsed into an early (1985–1999) (D1) or late (2000–2008) (D2) class. Also, the pre-series disturbance class was collapsed into the persisting forest

(PF) class. The persisting non-forest (PNF) and persisting water (PW) classes remained unchanged.

2.2. Sampling design

A probability sampling design consists of a list of sampling units that form a partition of the mapped area and a randomized process by which each of these units may be selected with known, strictly positive probability. The first decision we made was to use the set of mapped pixels (30 m-by-30 m areas) as our sampling units. A pixel-based approach was chosen over an object-based approach in order to keep the process of assigning classes to the reference data as straightforward as possible. The primary concern with using an object-based approach was that it would often be difficult to assign a single reference class to large objects.

Choosing an appropriate sampling design for an accuracy assessment requires balancing practical considerations (e.g., the cost of collecting of reference data) with statistical considerations (e.g., ensuring that rare classes will be adequately sampled). In our case, there were three main considerations that motivated decisions. First, one of our objectives was to produce results for both the U.S. portion of the map as well as separately for regions, which were considered to be logical landscape units. The US portion of the Lake Superior basin (LSB) constituted one region, and the Lake Michigan basin (LMB) constituted another. The LMB was further split into two separate regions — the northern half of the basin where the typical late-season imagery was used by the VCTw algorithm and the southern half of the basin where an abundance of row-cropped agricultural land required that early-season imagery be used. This split was not motivated by the desire to produce separate sets of results for the two halves of the LMB, but by statistical considerations. If distinct patterns of accuracy were observed in the two half-basins, stratification could result in more precise estimates (Särndal et al., 1992, Section 3.7). Consequently, the map was split into three geographic strata corresponding to these regions, and independent samples were drawn from each (Fig. 2).

Second, there was a two-level cost structure to obtaining the reference data. In order to observe one sampling unit, a raster image had to be processed, and then a photointerpreter had to inspect the imagery at the location of the unit and assign a class. Most of the cost of this observation was incurred during the first step, which suggested that more than one unit should be observed in each processed raster image. Hence, in each geographic stratum, two stages of sampling were conducted. In the first stage, a stratum was tessellated into $1/8 \times 1/8^{\circ}$ (450/450″) quadrangles. These quadrangles constituted the primary

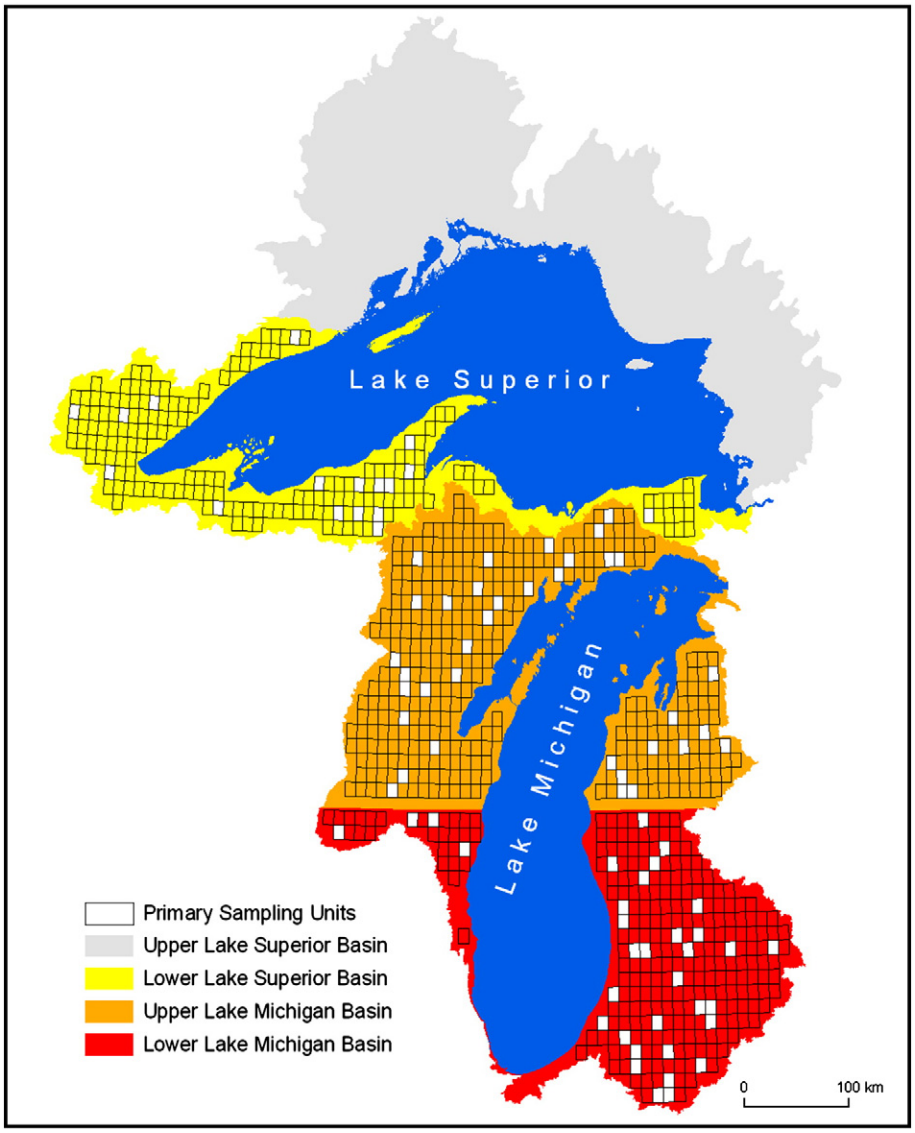

Fig. 2. The first-stage sampling frame for each of three geographic strata with highlighted sample PSUs. The Canadian portion of the map was not included in this assessment.

sampling units (PSUs), and a simple random sample of them was selected. In the second stage, map pixels constituted secondary sampling units (SSUs).

Third, another of our objectives was to produce results specific to each disturbance class. In an effort to provide an adequate amount of information about each class, SSUs were stratified by VCTw disturbance class (PNF, PF, PW, D1, or D2) during the second stage of sampling. A stratified random sample of SSUs was drawn from each PSU sampled during the first stage (Fig. 3).

Once the structure of the design was set, we determined the allocation of sample units. We first ignored the stratification at the second stage of sampling, and used guidelines offered by Cochran (1977, Section 10.10) to help determine the allocation of PSUs across geographic strata and the total number of SSUs sampled per PSU. To this end, we employed estimates of costs and variances based on a small amount of pilot data and on the technical expertise of specialists that had been involved in the creation of the VCTw data. This led to the selection of 17 PSUs from the LSB and 35 PSUs from each half of the LMB. Afterwards, we allocated the SSUs across second-stage strata according to the corresponding disturbance classes' relative importance to the larger project, which emphasized disturbance (Table 1). One exception to the allocations shown in Table 1 was that, if a PSU did not contain any PW pixels, one extra SSU in the D1 and D2 classes were to be sampled. Note that this aspect of the design was planned before observing the sample.

Conditional on our sampling design, the probability of including a given pixel in a sample was known, and is defined in Appendix A.

2.3. Response design

2.3.1. Reference imagery

The reference disturbance classes were obtained through photo-interpreting image sets from the Nation High Altitude Program (NHAP), National Aerial Photography Program (NAPP), and the National Agriculture Imagery Program (NAIP). A total of one NHAP image set (1986–1989), two NAPP image sets (1992–1994, 1998–1999), and

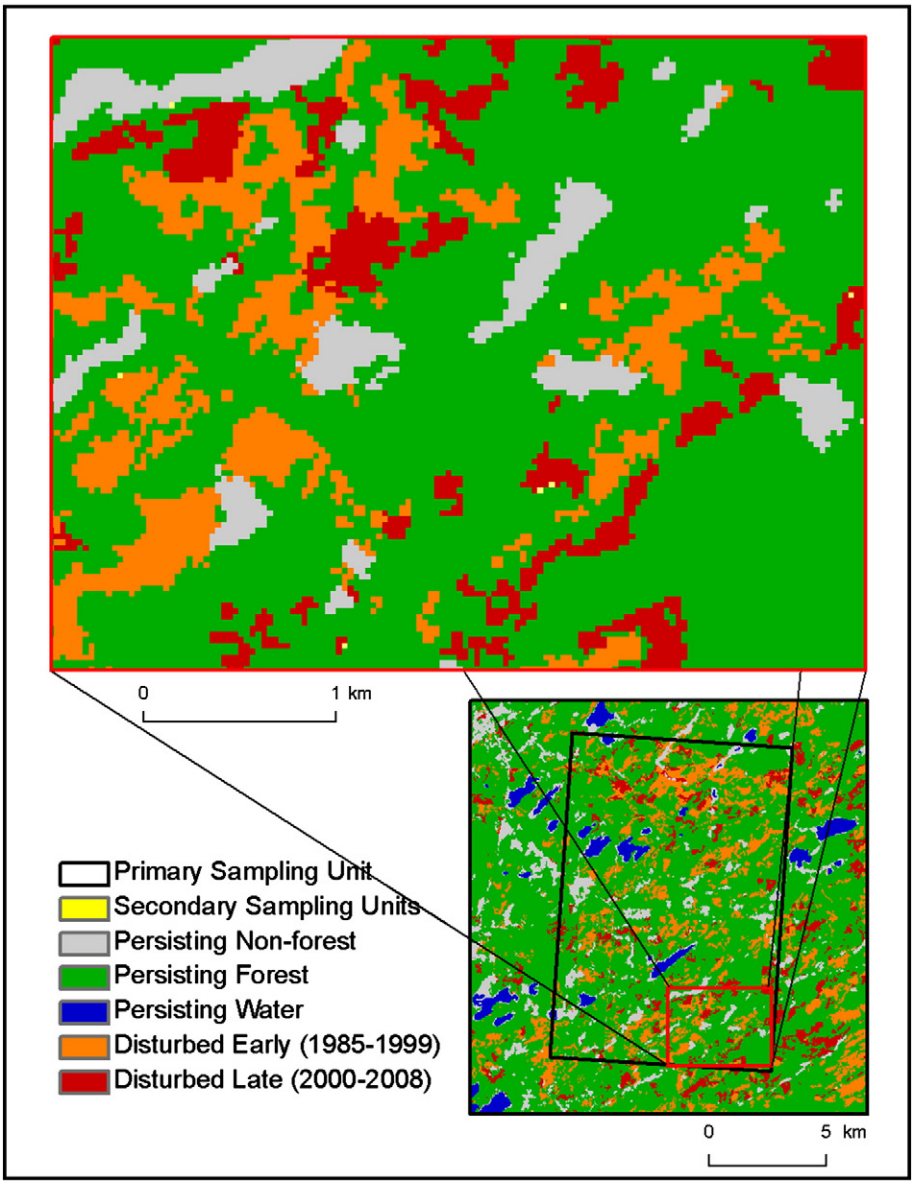

Fig. 3. Stratification used for the second stage of sampling. The lower-right image shows the footprint of a PSU. The prevalence of sample SSUs falling on edges of disturbance classes is discussed in Section 4.1.

Table 1
Allocation of SSUs (pixels) across disturbance classes (strata) per sampled PSU. The northern and southern halves of the LMB are denoted ULMB and LLMB, respectively. PNF = persisting non-forest, PF = persisting forest, PW = persisting water, D1 = disturbed during 1985–1999, and D2 = disturbed during 2000–2008.

	PNF	PF	PW	D1	D2
LSB	9	9	2	10	10
ULMB	5	5	2	6	6
LLMB	10	10	2	5	5

two NAIP image sets (2005, 2008) were available over the entire study area. The NAIP images, which have a spatial resolution of one meter, are required to have 95% of test points fall within 6 m of true ground (Davis, 2011). NHAP and NAPP images were manually georeferenced within the ArcMap environment to match the NAIP images. Due to the time gaps between these sets, Landsat Thematic Mapper (TM) imagery from both winter and summer phenological phases was occasionally used to supplement these data.

2.3.2. Interpretation of reference imagery

Interpreters based classifications of sampled SSUs on inspection of all five imagery sets. Because of the NAIP imagery's high level of spatial resolution and registration accuracy compared to the size of an SSU, interpreters limited inspection of NAIP imagery to a spatial support area defined by the footprint of the SSU (pixel). For NHAP or NAPP imagery however, it was occasionally necessary to expand this spatial support area to a three-by-three grid of Landsat pixels surrounding and including the SSU. This expanded support area was necessitated by the typically coarser spectral resolution of NHAP and NAPP images compared to NAIP images. If classification based on the aerial imagery was problematic (typically due to insufficient spectral resolution of the NHAP or NAPP imagery or to a time gap between the aerial imagery sets), Landsat imagery was also examined. Although the spatial resolution of Landsat is inferior to that of aerial imagery, its superior temporal and spectral resolution provided useful insight.

The disturbance class of each SSU was derived using a rule set that closely resembles that which was used to create the final VCTw disturbance class output. If an SSU's most recent disturbance (as observed in the reference imagery) was between 2000 and 2008, it was assigned to the D2 class. If its most recent disturbance was between 1985 and 1999, it was assigned to the D1 class. If an SSU appeared to be non-forest in the earliest imagery but to transition to forest in later imagery, it was classified as PF. Note that this is consistent with collapsing the VCT pre-series disturbance class into the PF class.

2.4. Statistical analysis

In this section, we explain the methods used to calculate statistical estimates related to map accuracy. The first estimation goal was to estimate accuracy parameters — i.e., confusion matrices, omission and commission error probabilities, and overall accuracies for the entire target region and for each basin. For a summary of the suite of classification accuracy parameters used to assess maps based on remote sensing data, see Stehman and Czaplewski (1998). The second goal was to estimate the percent cover of each disturbance class for the entire target region and for each basin. We first discuss accuracy estimates, then standard error estimates, and finally percent cover estimates.

2.4.1. Accuracy estimates

Every accuracy parameter to be estimated is of the form $p = \frac{N_a}{N_b}$ where N_a is the total number of pixels in some class a (e.g., pixels correctly classified as PF) and N_b is the total in some less restrictive class b (e.g., all pixels classified as PF). For each p to be estimated, we used a ratio estimator, which is of the form $\hat{p}_R = \hat{N}_a/\hat{N}_b$ are the Horvitz–Thompson estimators (also known as π estimators) of N_a and N_b, respectively (Särndal et al., 1992, Section 2.8). To describe the calculation of these estimators, we loosely follow the approach of Stehman et al. (2003). We define $y_a(k)$ and $y_b(k)$ to be the indicator functions that equal one when SSU k is in classes a and b, respectively, and equal zero otherwise. We also define π_k to be the inclusion probability of SSU k, as defined in Appendix A. Finally, we define s to be the set of SSUs in the sample. Then,

$$\hat{p}_R = \frac{\sum_{k\in s} y_a(k)/\pi_k}{\sum_{k\in s} y_b(k)/\pi_k}$$

where $k\in s$ in the subscript of the summation means to sum across all units k in s.

2.4.2. Standard errors

In order to calculate standard errors, so-called "linearized" variances were used. This technique is an approximation, necessitated here by the use of ratio estimators which are non-linear functions of two estimators, and whose variance cannot be easily expressed. The linearization technique substitutes a first-order linear (Taylor series) approximation of $\hat{p}_R$ based at p, and estimates this approximation's variance instead (Särndal et al., 1992, Section 5.5). In two-stage sampling designs, the estimation of variances and standard errors is algebraically complicated, and is therefore described in Appendix B. Computations (for this and all other estimation) were performed using the R statistical package, and the contributed R package `survey`, which are both freely available; see http://cran.r-project.org (Lumley, 2004, 2011; R Development Core Team, 2011). Some details on the `survey` package are given in Appendix C.

2.4.3. Percent cover estimates

In addition to estimating accuracy parameters, a sample of reference data can also be used to produce statistical estimates of the percent cover of each disturbance class over some area (e.g. Czaplewski & Catts, 1992; Magnussen et al., 2003; Stehman, 2009b). To be clear in this discussion, let p be the proportion of pixels that truly (according to our response design) falls into a given disturbance class, C, out of N total pixels. We then define $y_{ref}(k)$ to be the indicator function that equals one when the reference class of the kth pixel is C and zero otherwise. Now,

$$p = \frac{1}{N}\sum_{k=1}^{N} y_{ref}(k).$$

Ratio estimation and linearized standard errors, as described above, could be used to estimate p as the mean of y_{ref}. However, the difference estimator (Särndal et al., 1992, Section 6.3), which has been applied to remote sensing in a similar scenario by McRoberts (2011), is another possible strategy. To see how a difference estimator can be applied here, let $y_{map}(k)$ be the indicator function that equals one when the mapped disturbance class of the kth pixel is C and zero otherwise, and set $d(k) = y_{ref}(k) - y_{map}(k)$. Now, p can be reformulated as follows.

$$p = \frac{1}{N}\sum_{k=1}^{N} y_{map}(k) + \frac{1}{N}\sum_{k=1}^{N} d(k).$$

Since the mean of y_{map} is known, the choice of how to estimate p becomes a decision of whether to estimate the mean of y_{ref} or the mean of d. If we presume that the map data are mostly correct, d will usually equal zero and have little variability. This implies that an estimate of the mean of d will be very precise, and that a difference

Table 2
Estimated confusion matrix (and SEs) for entire target region. Sample size information is denoted by n. Estimates are given in percents, i.e. 100 times proportion. In cases where a SE had been rounded to zero, it is replaced with a "not applicable" (NA).

	VCTw						
		PNF	PF	PW	D1	D2	n
Reference	PNF	34.88 (2.05)	1.83 (0.47)	0 (NA)	0.21 (0.04)	0.28 (0.04)	845
	PF	2.12 (0.41)	50.9 (2)	0 (NA)	0.37 (0.06)	0.17 (0.03)	829
	PW	0.08 (0.03)	0 (NA)	1.72 (0.33)	0 (NA)	0 (NA)	158
	D1	0.35 (0.15)	1.85 (0.53)	0 (NA)	2.3 (0.34)	0.01 (0.01)	409
	D2	0.02 (0.02)	1.54 (0.49)	0 (NA)	0.01 (0.01)	1.36 (0.15)	399
	n	678	678	152	566	566	

Table 3
Estimated commission ($\hat{P}_C$) and omission ($\hat{P}_O$) error probabilities for each disturbance class for the entire target region.

	$\hat{P}_C$	SE($\hat{P}_C$)	$\hat{P}_O$	SE($\hat{P}_O$)
PNF	6.84	1.14	6.24	1.26
PF	9.29	1.31	4.97	0.80
PW	0.26	0.24	4.29	1.71
D1	20.66	3.25	48.93	7.33
D2	25.05	2.75	53.53	7.46

estimator should be used. In estimating the mean of d, we used ratio estimation. The resulting estimator is

$$\hat{p}_d = \frac{1}{N}\sum_{k=1}^{N} y_{map}(k) + \frac{\sum_{k\in s} d(k)/\pi_k}{\sum_{k\in s} 1/\pi_k}.$$

Since the first term in this estimator is known, its variance only depends on the second term (the ratio estimator of the mean of d). We used a linearized standard error estimator.

3. Results

The overall accuracy for the entire target region was estimated to be 91% with an estimated standard error (SE) of 0.8%. Table 2 provides the estimated confusion matrix and the corresponding estimated SEs. Table 3 provides the estimated omission and commission error probabilities along with their SEs. Approximate confidence intervals can be constructed where desired by adding and subtracting the SE multiplied by the relevant critical value from the Normal distribution to the estimate, e.g. use the estimate plus or minus SE times 1.96 for a 95% confidence level. Note that such intervals are based on the assumptions that the estimators are Normally distributed and that SEs are known, not estimated. Simulation studies have suggested that the actual coverage probability of such intervals is typically smaller than the nominal probability (Wolter, 2007). That is, if we repeatedly drew samples and calculated 95% confidence intervals, we could expect that less than 95% of them would contain the true parameter value.

The most striking characteristic of these results is the comparatively high rate of errors associated with the D1 and D2 classes. For example, the estimated commission error probabilities for D1 and D2 are about 21% and 25%, respectively. In other words, we estimate that about one out of five to one out of four pixels classified by the map as disturbed is incorrectly classified. The estimated omission error probabilities for the D1 and D2 classes are about 49% and 54%, respectively, which suggests that VCTw has found somewhere in the neighborhood of one half of the truly disturbed pixels. Note, however, that the standard errors associated with omission errors (here and elsewhere) are large. Studying the confusion matrix, we can see that, when a pixel is inaccurately classified as disturbed, results suggest that it is typically either PF or PNF. Likewise, when a truly disturbed pixel is missed, it is typically misclassified as either PF or PNF.

The overall accuracy for the LSB was estimated to be 87% with an estimated standard error of 2% while the overall accuracy of the LMB was estimated to be 92% with an estimated standard error of 1%. For the LSB, Table 4 provides the estimated confusion matrix and its corresponding SEs, while Table 5 provides the estimated omission and commission error probabilities along with their estimated SEs. For the LMB, Table 6 provides the estimated confusion matrix and its corresponding SEs while Table 7 provides the estimated omission and commission error probabilities along with their estimated SEs. Approximate confidence intervals can be constructed in the same way here as before. Note that the smaller sample sizes for region-specific estimates will negatively affect the coverage probabilities of these intervals.

Comparing the LSB and LMB results, two different patterns of errors emerge. In the LSB, estimated commission error probabilities for all classes besides PF were relatively small (between approximately 0% and 10%), while the estimated probabilities of omission errors for PNF, D1, and D2 were relatively large (greater than 25%). These results, along with an examination of the basin's estimated confusion matrix, suggest that the common errors were misclassifying PNF, D1, and D2 pixels as PF. So, pixels that are classified as disturbed are likely to be disturbed, but we estimate that less than half of truly disturbed pixels are classified correctly. In contrast, in the LMB, only the D1 and D2 classes had estimated omission or commission error probabilities above 10%. Here, both omission and commission errors related to the disturbance classes were common.

Percent cover estimates are presented in Table 8. Comparing the estimates to the map data for the entire target region, the primary difference is the larger estimates of both disturbance classes. Specifically, the percent cover D1 estimate is 4% compared to the 3% derived from the map data, and the percent cover D2 estimate is 3% compared

Table 4
Estimated confusion matrix (and SEs) for the LSB. Sample size information is denoted by n. In cases where a SE had been rounded to zero, it is replaced with a NA.

	VCTw						
		PNF	PF	PW	D1	D2	n
Reference	PNF	10.99 (2.56)	3.93 (1.45)	0 (NA)	0.12 (0.06)	0.16 (0.04)	184
	PF	0.54 (0.23)	67.77 (2.61)	0 (NA)	0.15 (0.06)	0.03 (0.02)	166
	PW	0.13 (0.06)	0 (NA)	2.27 (1.07)	0 (NA)	0 (NA)	34
	D1	0.17 (0.07)	4.45 (1.84)	0 (NA)	4.04 (1.42)	0.03 (0.02)	156
	D2	0.09 (0.09)	3.24 (1.14)	0 (NA)	0.01 (NA)	1.89 (0.53)	140
	n	153	153	30	172	172	

Table 5
Estimated commission ($\hat{P}_C$) and omission ($\hat{P}_O$) error probabilities for each disturbance class for the LSB. In cases where a SE had been rounded to zero, it is replaced with a NA.

	$\hat{P}_C$	SE($\hat{P}_C$)	$\hat{P}_O$	SE($\hat{P}_O$)
PNF	7.83	3.44	27.69	7.39
PF	14.63	2.56	1.06	0.38
PW	0.00	NA	5.29	2.84
D1	6.47	3.31	53.51	13.04
D2	10.11	3.93	63.78	10.13

to 2%. On a regional level, the largest differences between the estimates and the map data appear in the LSB, where the percent cover PF estimate is 67% versus the 78% as derived from the map data. The percent cover D1 and D2 estimates (9% and 6%, respectively), make up much of the difference, and are roughly double what is given by the map data.

As expected, the precision of the percent cover estimates depends on the accuracy of the map. For the PNF, PF, and PW classes, the coefficient of variation (CV) of percent cover estimates is almost always less than 5%. In contrast, for the D1 and D2 classes, which were classified with less accuracy, the CV of percent cover estimates ranges from 10% to 26%.

4. Discussion

4.1. Sampling and response design

The chosen sampling design was successful in that it allowed the estimation of accuracy parameters and SEs with acceptable statistical modeling assumptions and very little approximation (the use of linearized variances constitutes the sole approximation). A key to achieving this simplicity was following the guidelines of invariance and independence (Särndal et al., 1992, Section 4.3). These guidelines place restrictions on what second-stage sampling can depend upon: the second-stage design to be used in a given PSU, if it is selected, cannot depend upon which other PSUs are selected (invariance), and the selection of SSUs in a given selected PSU must be independent from the selection of SSUs in any other selected PSU (independence). These restrictions allow for a relatively simple decomposition of the sampling variability of an estimate into variability due to the first and second stages of sampling.

An important implication of these restrictions is that it would not have been possible to increase the number of disturbed pixels sampled from a given selected PSU because of a shortage of disturbed pixels in other selected PSUs. This means that, unfortunately, the principles of invariance and independence cannot be followed in every accuracy assessment, especially when a two-stage design is used and very rare classes are of interest (e.g. Stehman et al., 2003). An attractive alternative in such a scenario is to conduct a second independent survey designed specifically to sample the rare class and combine the information from both surveys to produce an improved estimator of population characteristics related to the rare class (see Czaplewski, 2010). For this assessment, we were able to allay fears of such a shortage by studying the composition of each PSU in the entire target region before finalizing the design. We discovered that, with the minor exception of the PW class described above, in every PSU the map assigned at least as many pixels to each class as we planned to sample.

Another simplifying characteristic of our accuracy assessment was that sufficient reference data was actually available for every sampling unit in the target region. This was achieved partly through collapsing the biennial disturbance classes and partly through making auxiliary use of Landsat imagery. Consequently, statistical issues related to nonresponse were avoided, and interpretations can legitimately be made on the level of the target region.

As an alternative to using simple random sampling within strata, for either PSUs or SSUs, systematic sampling could have been used.[2] The decision to use simple random sampling was motivated by a desire for statistical simplicity. In order to produce standard errors for estimates based on systematic samples, assumptions must be made concerning the structure of the population. The typical assumption made is that the population is in random order, which implies that SEs should be calculated as if a simple random sample had been selected. If the population actually exhibits positive spatial autocorrelation (things close together are more similar than things far apart), this assumption will likely lead to underestimating the precision of estimates. However, depending on the actual structure of the population, it is possible to overestimate precision (e.g. Moisen et al., 1994). Hence, a gain in precision may have been possible by using systematic sampling, but its realization through the calculation of SEs is not straightforward.

A complication in the sampling procedure for this assessment arose through the use of the Arc tool Create Random Points. The tool had a preference to select pixels on the edges of disturbance class polygons, particularly for polygons with very irregular shapes. This preference is apparent in Fig. 3. We examined the proportion of SSUs that fell on the edge of polygons in our sample and in the population for each combination of first- and second-stage strata. In some of these combinations, the proportion was much higher in the sample than in the population. In other words, our sampling procedure evidently over-sampled SSUs on the edges of disturbance class polygons. This was concerning because it is plausible that SSUs on edges are less likely to be correctly classified than those in interiors. Over-sampling edges could then lead to substantially under-estimating accuracy. We checked whether this was the case by post-stratifying our accuracy estimates (confusion matrix elements and commission and omission error probabilities) by an edge/interior stratification. The post-stratified estimates were negligibly different from our original set of estimates. In particular, all post-stratified estimates were within one percentage point of the original estimate, except in the case of a few omission error probabilities whose large standard errors dwarfed the discrepancy. We chose to present our original set of estimates rather than the post-stratified estimates because the post-stratified sample configuration no longer contained at least two sample SSUs in every stratum within each PSU, necessitating the use of approximations in variance estimation.

A primary difficulty in the response design of this assessment was due to the land cover change aspect of the VCTw algorithm; the long time scale (1985–2008) of the map being assessed necessitated multiple sets of reference imagery. This increased the difficulty of assignments due to temporal gaps and inconsistency in the composition of the reference imagery. In particular, the appearance of disturbances varied among the NHAP, NAPP, and NAIP image sets. Also, even within a NHAP or NAPP image set, the appearance of disturbances varied due to differences in phenology among acquisition years. These obstacles motivated the use of Landsat as an auxiliary source of consistent reference imagery, which was largely successful in resolving problematic classifications.

Another difficulty in assigning reference classes arose when an SSU had apparently heterogeneous land cover. This problem has been solved in other studies by allowing so-called “fuzzy” classification (Gopal & Woodcock, 1994). However, we assigned only a single reference class to all sampled units. In the heterogenous cases, the disturbance class was chosen that reflected the majority of the area within the 30 m-by-30 m sampled unit. We found this method acceptable because, regardless of whether a single classification sufficiently described the true land cover of a unit, the classification's purpose is to assess an algorithm which is solely concerned with land cover characteristics aggregated to this scale.

[2] See Stevens and Olsen (2004) and the associated contributed R package `spatstat` (Baddeley & Turner, 2005; R Development Core Team, 2011) for an interesting alternative to both simple random and systematic sampling.

Table 6
Estimated confusion matrix (and SEs) for the LMB. Sample size information is denoted by n. In cases where a SE had been rounded to zero, it is replaced with a NA.

		VCTw					
		PNF	PF	PW	D1	D2	n
Reference	PNF	41.38 (2.51)	1.25 (0.46)	0.01 (0.01)	0.24 (0.04)	0.31 (0.05)	661
	PF	2.55 (0.52)	46.31 (2.44)	0 (NA)	0.44 (0.08)	0.2 (0.04)	663
	PW	0.06 (0.03)	0 (NA)	1.57 (0.29)	0 (NA)	0 (NA)	124
	D1	0.39 (0.19)	1.14 (0.46)	0 (NA)	1.83 (0.21)	0.01 (NA)	253
	D2	0 (NA)	1.07 (0.54)	0 (NA)	0.02 (0.01)	1.22 (0.13)	259
	n	525	525	122	394	394	

Table 7
Estimated commission ($\hat{P}_C$) and omission ($\hat{P}_O$) error probabilities for each disturbance class for the LMB.

	$\hat{P}_C$	SE($\hat{P}_C$)	$\hat{P}_O$	SE($\hat{P}_O$)
PNF	6.77	1.20	4.19	1.06
PF	6.97	1.49	6.44	1.11
PW	0.36	0.34	3.89	2.13
D1	27.28	3.34	45.73	8.70
D2	29.98	2.83	47.21	11.31

4.2. Statistical analysis

Ratio estimators were used for every accuracy parameter, but these were not the only possible choice. Except for the case of omission error probabilities, the denominator in $p = N_a/N_b$ was known, making $\hat{p}$ another possible estimator where

$$\hat{p} = \hat{N}_a/N_b.$$

Both $\hat{p}$ and $\hat{p}_R$ are design consistent and $\hat{p}$ is design unbiased while $\hat{p}_R$ is approximately design unbiased. However, the decision between the two should really be made based on which has a smaller expected mean square error (which is variance plus squared bias). It seems counterintuitive that estimating the quantity N_b would be preferable to using the known value, but this is often the case. See Särndal et al. (1992, Section 5.7) for a discussion.

The linearized variance technique was used for every SE calculation, but it was also not the only possible choice. Jackknife SEs are based on the idea of repeatedly drawing subsamples from the observed sample and studying the amount of variability among these subsamples. This technique was appropriate for our sampling design and readily available in the `survey` package (Lumley, 2004, 2011). In fact, there is some empirical (i.e., anecdotal) evidence that jackknife SEs are preferable to linearized SEs with respect to the coverage probability of confidence intervals (Wolter, 2007). With this in mind, jackknife SEs were also calculated for every estimate. In almost all cases the jackknife SE was slightly smaller than the linearized SE, but not enough to make any substantial difference in the interpretation of the estimate. Consequently, linearized SEs, which are probably more familiar to most readers, were used. However, for a couple of region-specific estimates (the omission error probabilities for D1 in the LSB and for D2 in the LMB), the confidence interval based on the jackknife was much larger; linearized SEs are suspicious here because they are known to sometimes perform poorly in small sample sizes (Wolter, 2007). The jackknife SEs were not substituted in these cases out of a desire for simplicity and because the confidence intervals based on linearized SEs are so wide already that the difference is probably not practically important. Nevertheless, jackknife SEs should be considered for future accuracy assessments, especially in small sample scenarios.

The use of the accuracy assessment sample to estimate the percent cover of each disturbance class was a substantial improvement over reporting percent cover estimates based only on map data. By using all the available information, not only can these estimates be expected to be more accurate, but their potential magnitude of error can be directly assessed through estimated standard errors. In contrast, the potential magnitude of error of map-only estimates must be interpreted through the raw accuracy estimates. Also, the difference estimator and its standard error can be calculated using the same procedures as those already being used to calculate accuracy estimators. As the use of probability-based accuracy assessments becomes standard practice in the remote sensing field, the use of sample data to improve percent cover (or total area) estimates should also become standard practice.

Table 8
Percent cover estimates for the entire target region, the LSB, and the LMB. p_{map} is the map-only estimate, and $\hat{p}_d$ is the difference estimator described in Section 2.4.3.

	Entire region			LSB			LMB		
	p_{map}	$\hat{p}_d$	SE($\hat{p}_d$)	p_{map}	$\hat{p}_d$	SE($\hat{p}_d$)	p_{map}	$\hat{p}_d$	SE($\hat{p}_d$)
PNF	39.92	39.67	0.69	13.19	16.46	1.60	47.08	45.88	0.75
PF	53.20	50.65	0.89	77.55	66.66	2.21	46.68	46.39	0.95
PW	2.00	2.07	0.03	1.85	1.98	0.06	2.04	2.10	0.04
D1	2.82	4.42	0.57	4.42	8.78	1.84	2.39	3.25	0.53
D2	2.06	3.18	0.49	3.00	6.12	1.14	1.81	2.38	0.54

4.3. Map accuracy

Although this project is currently the only implementation of the VCTw algorithm, it can – with some reservations – be compared to an implementation of the VCT algorithm presented by Huang et al. (2010) and by Thomas et al. (2011). For a collection of six North American Forest Dynamics (NAFD) sites and aggregated across all biennial disturbance classes, Thomas et al. (2011) report estimated commission error probabilities between roughly 15% and 45% and estimated omission error probabilities between roughly 45% to 60% (with the exception of one NAFD site with very few pixels in a disturbed reference class). These results are roughly consistent with the VCTw results given above, but it should be noted that collapsing to only two disturbance classes (D1 and D2) certainly improved the VCTw results compared to what they would have been if the accuracy of disturbance classifications had been assessed within biennial time steps. Thomas et al. (2011) also report very similar results for the PNF class: estimated omission and commission error probabilities are all below 15% except for an estimated probability of omission of 33% which corresponded to a Minnesota site that contained many wetlands. The performance of VCTw for PNF showed the same pattern in that the estimated omission error probability in the LSB (which contains many wetlands) was relatively large. For the PF class, Thomas et al. (2011) report similar estimated omission error probabilities (mostly below 15%) to those seen here, but larger estimated commission error probabilities (between 15% and 43% compared to 5% and 15% for VCTw). This improvement is likely a consequence of the use of winter imagery to mask non-forest areas in the VCTw algorithm (Stueve et al., 2011).

The preponderance of mapping errors related to the D1 and D2 classes may be explained, at least in part, by the fact that VCT – and, thus, VCTw – is best suited for detecting stand-clearing forest

disturbance (Huang et al., 2010). Note that, because a small proportion of land is truly disturbed, VCTw and VCT are still able to achieve relatively high overall accuracies.

Another important theme of the accuracy results presented here, consistent with the results presented by Thomas et al. (2011), is that the accuracy of forest disturbance mapping is highly dependent on the characteristics of the landscape being mapped. The landscapes studied here alternate between abundant and sparse forest cover and have varying degrees of agricultural and developed land. These differences seem to have translated into very distinct patterns of accuracy. For instance, compared to the LMB, the LSB estimates exhibit a reduction of about 8% in the commission error probability for the PF class and about 25% in the omission error probability for the PNF class. These results suggest that efforts to map forest disturbance on large scales (i.e., regional and above) should consider flexible procedures that can take advantage of and avoid the pitfalls associated with features of regional landscapes. In the VCTw algorithm, the use of early-season imagery in the southern half of the LMB to accommodate row-cropped agricultural land and the use of winter imagery to identify snow-covered non-forest land are two examples of this.

5. Conclusion

The usefulness of any land-cover map is dependent on a sound accuracy assessment of its land-cover classifications. Here, we presented an assessment of a map of land cover change in the basins draining into Lake Superior and Lake Michigan produced by Stueve et al. (2011) using the VCTw algorithm. Our use of two-stage cluster sampling improved the efficiency of reference data collection relative to single stage sampling, and our use of stratification allowed us to purposely distribute data collection across geographic regions and mapped disturbance classes. Constructing a reference data set was challenging because of the large spatial and temporal scales of the assessment, but the response design was able to utilize Landsat imagery as an auxiliary source of data when reference classifications based solely on aerial imagery were not reliable. The estimation procedures produced almost-unbiased, design-consistent estimates of accuracy parameters, and all estimates were accompanied by SEs. We also used the observed reference data to improve the map-based percent-cover estimates. As a whole, these decisions produced a useful assessment that was conducted efficiently and is statistically defensible.

Acknowledgments

This work was partially supported by the U.S. EPA Great Lakes Restoration Initiative. Special thanks are extended to Ray Czaplewski and Ron McRoberts for reviewing drafts of this manuscript. Anonymous reviewers' comments were also particularly valuable and their efforts are greatly appreciated. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Appendix A. Inclusion probabilities

In order to define the inclusion probability π_k of SSU k, we will first define the inclusion probability of PSU i, denoted π_{1i}. Let g be the geographic stratum containing i, N_{1g} be the number of PSUs in this stratum, and n_{1g} be the number of sample PSUs in this stratum. Now, $\pi_{1i} = n_{1g}/N_{1g}$.

Next, we define the conditional inclusion probability of SSU k given that the PSU containing it, i, has been selected. Let h be the disturbance class stratum containing k, N_{ih} be the number of SSUs in PSU i and stratum h, and n_{ih} be the number of sample SSUs in PSU i and stratum h. Then, the conditional inclusion probability is $\pi_{k|i} = n_{ih}/N_{ih}$. Finally, the inclusion probability of SSU k, contained in PSU i, can be written as $\pi_k = \pi_i \pi_{k|i}$.

Appendix B. Linearized standard errors

We now present details on estimating the variance of a ratio estimator, $\hat{p}_R$ under the sampling design used in our assessment. In order to remain concise, we will assume that the reader is familiar with some central concepts of design-based sampling statistics, and will refer to results given in Särndal et al. (1992). We must deal with two complicating factors: $\hat{p}_R$ has a non-linear form, and a complex sampling design was used.

First, we deal with the non-linearity of $\hat{p}_R$ by assuming that its variance is equal to that of the linear approximation

$$\tilde{p} = p + \frac{1}{N_b} \sum_{k \in s} \frac{y_a(k) - p y_b(k)}{\pi_k}$$

as in Särndal et al. (1992, Result 5.6.2). So, instead of looking at both y_a and y_b, we only pay attention to the variable $e(k) = y_a(k) - p y_b(k)$, and its sample version $\hat{e}(k) = y_a(k) - \hat{p}_R y_b(k)$. We now can write the linearized variance estimator of $\hat{p}_R$ as $\hat{V}\left(\hat{N}_e\right)/\hat{N}_b^2$ where $\sum_{k \in s} (\hat{e}(k)/\pi_k)$.

Next, we deal with the complex sampling design. To simplify the variance resulting from the two stages of sampling used in our assessment, we decompose the variance of $\hat{N}_e$ into elements that correspond to the first and second stages of sampling, as in Särndal et al. (1992, Result 4.3.1). To recognize the stratification in the second stage of sampling, we use the special case of Eq. (4.3.5) in Särndal et al. (1992) that applies to a stratified random sample. Finally, to recognize the use of stratification in the first stage of sampling, we estimate $\hat{V}\left(\hat{N}_e\right)$ separately for each first-stage stratum, and weight them in the typical fashion.

Appendix C. The `survey` package

We now present details on the R software environment (R Development Core Team, 2011) and the `survey` package (Lumley, 2004, 2011), which were used for all computing. In order to calculate an estimate and corresponding SE, a user needs to first create a survey design object that provides information about the sampling design, and then use an estimator function. We will provide some example code applicable to the estimates calculated for our assessment. Define the following data vectors, all of the same length, where the *kth* element in each vector corresponds to the *kth* sampled SSU:

- `psu` — an index identifying k's PSU
- `ssu` — an index identifying k's SSU
- `region` — the first-stage stratum of k's PSU
- `map` — the second-stage stratum of k's SSU
- `fpc1` — the number of PSUs in the same first-stage stratum as k
- `fpc2` — the number of SSUs in the same second-stage stratum and same PSU as k
- `match` — the correctness of the k's map class
- `ind` — the number 1.

Now, the following code could be used to obtain the ratio estimate and linearized SE (as described in this paper) for the proportion of correct SSUs in a map under the sampling design used in our assessment.

```
des<-svydesign(ids=~psu+ssu,strata=~region+map,fpc=~fpc1+fpc2,vars=data.frame(match,ind))
svyratio(numerator=~match, denominator=~ind, design=des)
```

In order to calculate the SE here, the `svyrecvar` function will be called by `svyratio`. This function uses a recursive algorithm to calculate the estimated variance contributed by both stages of sampling. For more details, see Lumley (2010).

References

Baddeley, A., & Turner, R. (2005). Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*, 1–42.

Cihlar, J. (2000). Land cover mapping of large areas from satellites: Status and research priorities. *International Journal of Remote Sensing*, *21*, 1093–1114.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Congalton, R. G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: Principles and practices*. Boca Raton: Lewis Publishers.

Czaplewski, R. L. (2010). *Complex sample survey estimation in static state-space*. Gen. Tech. Rep. RMRS-239. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Czaplewski, R. L., & Catts, G. P. (1992). Calibration of remotely sensed proportion or area estimates for misclassification error. *Remote Sensing of Environment*, *39*, 29–43.

Davis, D. (2011). *2011 NAIP information sheet*. United States Department of Agriculture, Farm Service Agency, Aerial Photography Field Office, http://www.fsa.usda.gov

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

Gregory, S. V., Swanson, F. J., McKee, W. A., & Cummins, K. W. (1991). An ecosystem perspective of riparian zones. *BioScience*, *41*, 540–551.

Huang, C., Goward, S. N., Masek, J. G., Gao, F., Vermote, E. F., Thomas, N., et al. (2009). Development of time series stacks of Landsat images for reconstructing forest disturbance history. *International Journal of Digital Earth*, *2*, 195–218.

Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., & Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sensing of Environment*, *114*, 183–198.

Karr, J. R., & Schlosser, I. J. (1978). Water resources and the land-water interface. *Science*, *201*, 229–234.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*, 1–19.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. New Jersey: Wiley.

Lumley, T. (2011). *Survey: Analysis of complex survey samples*. R package 3.24.

Magnussen, S., Stehman, S. V., Corona, P., & Wulder, M. A. (2003). A Pólya-urn resampling scheme for estimating precision and confidence intervals under one-stage cluster sampling: Application to map classification accuracy and cover-type frequencies. *Forest Science*, *50*, 810–822.

McRoberts, R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? *Remote Sensing of Environment*, *115*, 715–724.

Miles, P. D., Heinzen, D., Mielke, M. E., Woodall, C. W., Butler, B. J., Piva, R. J., et al. (2011). *Minnesota's Forests 2008*. Resour. Bull. NRS-50. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station.

Moisen, G. G., Edwards, T. C., Jr., & Cutler, D. R. (1994). Spatial sampling to assess classification accuracy of remotely sensed data. In W. K. Michener, J. W. Brunt, & S. G. Stafford (Eds.), *Environmental information management and analysis: Ecosystem to global scales* (pp. 159–176). New York: Taylor & Francis.

Naiman, R. J., & Bilby, R. E. (Eds.). (1998). *River ecology and management: Lessons from the Pacific Coastal Ecoregion*. New York: Springer-Verlag.

Nusser, S. M., & Klaas, E. E. (2003). Survey methods for assessing land cover map accuracy. *Environmental and Ecological Statistics*, *10*, 309–331.

Perry, C. H., Everson, V. A., Brown, I. K., Cummings-Carlson, J., Dahir, S. E., Jepsen, E. A., et al. (2008). *Wisconsin's forest, 2004*. Resour. Bull. NRS-23. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station.

Peterjohn, W. T., & Correll, D. L. (1984). Nutrient dynamics in an agricultural watershed: Observations on the role of a riparian forest. *Ecology*, *65*, 1466–1475.

Pugh, S. A., Hansen, M. H., Pedersen, L. D., Heym, D. C., Butler, B. J., Crocker, S. J., et al. (2009). *Michigan's forests 2004*. Resour. Bull. NRS-34. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model-assisted survey sampling*. New York: Springer-Verlag.

Stehman, S. V. (2009a). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, *30*, 5243–5272.

Stehman, S. V. (2009b). Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sensing of Environment*, *113*, 2455–2462.

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

Stehman, S. V., Wickham, J. D., Smith, J. H., & Yang, L. (2003). Thematic accuracy of the 1992 National Land-Cover Data for the eastern United States: Statistical methodology and regional results. *Remote Sensing of Environment*, *86*, 500–516.

Stevens, D. L., Jr., & Olsen, T. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*, 262–278.

Strahler, A. H., Boschetti, L., Foody, G. M., Friedl, M. A., Hansen, M. C., Herold, M., et al. (2006). *Global land cover validation: Recommendations for evaluation and accuracy assessment of global land cover maps*. GOFC-GOLD Report No. 25. Luxemburg: Office for Official Publications of the European Communities.

Stueve, K. M., Housman, I. W., Zimmerman, P. L., Nelson, M. D., Webb, J. B., Perry, C. H., et al. (2011). Snow-covered Landsat time series stacks improve automated disturbance mapping accuracy in forested landscapes. *Remote Sensing of Environment*, *115*(12), 3203–3219, http://dx.doi.org/10.1016/j.rse.2011.07.005.

Sweeney, B. W. (1992). Streamside forests and the physical, chemical, and trophic characteristics of piedmont streams in Eastern North America. *Water Science and Technology*, *26*, 2653–2673.

Thomas, N. E., Huang, C., Goward, S. N., Powell, S., Rishmawi, K., Schleeweis, K., et al. (2011). Validation of North American forest disturbance dynamics derived from Landsat time series stacks. *Remote Sensing of Environment*, *115*, 19–32.

Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York: Springer-Verlag.