Can. J. Remote Sensing, Vol. 32, No. 3, pp. 268–279, 2006

Use of vector polygons for the accuracy assessment of pixel-based land cover maps

Michael A. Wulder, Joanne C. White, Joan E. Luther, Guy Strickland, Tarmo K. Remmel, and Scott W. Mitchell

Abstract. Identifying appropriate validation sources for large-area land cover products is a challenge, with logistical constraints frequently necessitating the use of preexisting data sources. Several issues exist when comparing polygon (vector-based) datasets to raster imagery: geolocational mismatches, differences in features or classes mapped, disparity between the scale of polygon delineation and the spatial resolution of the image, and temporal discrepancies. To evaluate the potential impact of using vector coverages to assess the accuracy of pixel-based land cover maps, five evaluation protocols are applied to test sites located in British Columbia and Newfoundland and Labrador, Canada. One protocol directly compared the land cover of the sample unit to the land cover of the forest inventory polygon within which the sample unit fell, two protocols used different regions around the sample unit to define the land cover class, and two protocols were based on homogeneity criteria that restricted the selection of sample units. For the protocols tested, the overall accuracy values ranged from 34% to 58%. Given the broad range of accuracies achieved, the results suggest that caution is needed when making spatially explicit comparisons between raster and vector datasets. When possible, the use of purpose-collected validation data is recommended for the accuracy assessment of maps derived from remotely sensed data; if preexisting vector-based data are the only option for the validation, approaches accounting for the heterogeneity of classes within a given polygon are recommended.

Résumé. L'identification des sources appropriées de validation pour les produits du couvert végétal à grande surface pose un défi, les contraintes logistiques requérant fréquemment l'utilisation de sources de données préexistantes. Plusieurs problèmes se manifestent lorsque l'on compare des ensembles polygonaux de données (basés sur les vecteurs) par rapport aux images matricielles : des décalages de géo-localisation; des différences dans les caractéristiques ou classes cartographiées; des disparités entre l'échelle de la délimitation polygonale et la résolution spatiale de l'image; et des divergences temporelles. Afin d'évaluer l'impact potentiel de l'utilisation de couvertures vectorielles pour l'estimation de la précision des cartes du couvert à base de pixels, cinq protocoles d'évaluation sont appliqués à des sites tests situés en Colombie-britannique et à Terre-Neuve et au Labrador, au Canada. Un des protocoles compare directement le couvert de l'unité d'échantillonnage et celui du polygone d'inventaire forestier à l'intérieur duquel l'unité se situait; deux des protocoles utilisaient différentes régions autour de l'unité d'échantillonnage pour définir la classe de couvert, alors que les deux autres protocoles étaient basés sur un critère d'homogénéité qui a limité le choix des unités d'échantillonnage. Pour les protocoles testés, la précision globale variait de 34 % à 58 %. Étant donné la grande variété de taux de précision atteints, les résultats indiquent qu'une certaine prudence est de mise lorsque l'on fait des comparaisons spatialement explicites entre des ensembles de données matricielles et vectorielles. Lorsque possible, il est recommandé d'utiliser des données de validation acquises spécifiquement pour les besoins de l'évaluation de la précision des cartes dérivées des données de télédétection. Si les données vectorielles préexistantes constituent la seule option possible pour la validation, les approches tenant compte de l'hétérogénéité des classes à l'intérieur d'un polygone donné sont recommandées.
[Traduit par la Rédaction]

Received 10 November 2005. Accepted 13 June 2006.

M.A. Wulder[1] and J.C. White. Pacific Forestry Centre, Canadian Forest Service, Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada.

J.E. Luther and G. Strickland. Atlantic Forestry Centre, Canadian Forest Service, Natural Resources Canada, P.O. Box 960, Corner Brook, NF A2H 6J3, Canada.

T.K. Remmel. Department of Geography, University of Toronto at Mississauga, 3359 Mississauga Road North, Mississauga, ON L5L 1C6, Canada.

S.W. Mitchell. Department of Geography and Environmental Studies, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada.

[1]Corresponding author (e-mail: mike.wulder@nrcan.gc.ca).

Introduction

Remotely sensed data enable the routine production of land cover maps, representing small areas (Ehlers et al., 2003), regions (Homer et al., 1997), nations (Wulder et al., 2003), continents (Stone et al., 1994), and the globe (Loveland and Belward, 1997). An accuracy assessment is a prerequisite to the subsequent use of these maps for various applications (Stehman and Czaplewski, 2003). Accuracy-assessment protocols require validation data independent from information used in map development; however, it is expensive and logistically challenging to collect data specifically for the validation of large-area land cover products (Cihlar, 2000). Preexisting reference data (vector based) are often available as validation data sources. Application of these data for accuracy assessment may be limited, however, because of issues such as locational disparities between the available reference data and the remotely sensed land cover map product, differences in classification systems or legends, discrepancies between the spatial scale at which the polygons in the vector-based data were delineated and the resolution of the image from which the land cover map was produced (Remmel and Perera, 2002), and temporal discrepancies in data collection.

Errors in location can be calculated and corrected, or compensated for, by the use of a spatial support region (SSR) (Verbyla and Hammond, 1995), which is defined as “the size, geometry and orientation of the space on which an observation is defined” (Atkinson and Curran, 1995, p. 768). Categorical differences between the maps may be addressed through the harmonization of class legends (McConnell, 2002). It is more difficult to account for disparities in spatial and temporal scales in data collection. Polygons are delineated as a generalization of specific characteristics, whereas remotely sensed data capture discrete attributes for a regularly sized area of the earth’s surface. As a result, the remotely sensed data often provide greater detail than the generalized polygonal counterpart. Temporally, map and reference data rarely correspond (Liu and Zhou, 2004); for land cover mapping, this disparity can be significant, particularly in dynamic landscapes. Although it is not expected that pixel-based maps will perfectly emulate independently generated vector-based maps, it may be expected, given ideal conditions of compatible map categories over homogeneous areas with temporal convergence, that the vector- and raster-based classes will correspond more frequently than not. Unfortunately, such ideal conditions rarely exist. Discrepancies between products can be anticipated for heterogeneous areas, difficult to map classes (e.g., mixed classes), and rare classes.

The goal of this research is to explore the ramifications of using preexisting vector-based forest inventory data to validate a raster-based large-area land cover product and to make recommendations based on the results. Insights are sought on how to optimize a comparison between vector and raster representation of land cover, beyond simple overlap statistics, in support of accuracy assessment. Towards these goals, five different protocols for comparing a raster image classification with vector forest inventory data are tested. Recognizing that no single measure of accuracy will be optimal in every situation (Foody, 2002), the objective of this work is to investigate the implications of these various protocols to the estimation and reporting of accuracy assessment results. Our approach is transparent and provides the rationale for the selections made, such that users of the land cover maps are empowered to assess the utility of the protocols for their own applications. The analysis is undertaken on two study areas in Canada representing different ecological conditions. The data selection is intended to be representative of commonly used datasets to enable inferences beyond the datasets used in this study.

Study areas

The study areas for this analysis are located in Prince George, British Columbia, and Deer Lake, Newfoundland and Labrador, Canada (**Figure 1**). The Prince George study area covers approximately 14 500 km^2, with a mean elevation of 889 m, minimum elevation of 247 m, and maximum elevation of 1992 m. Located in the Sub-Boreal Spruce biogeoclimatic zone (Meidinger and Pojar, 1991), the forests in the study area are dominated by conifer tree species; however, areas of urban and agricultural land use are also found within the study area. The primary land cover disturbances in the study area are forest harvesting and insect damage (and related salvage activities), with some isolated fire events. The Deer Lake study area covers approximately 16 400 km^2, with a mean elevation of 306 m, minimum elevation of 0 m, and maximum elevation of 836 m. This study area is located in the Boreal Shield ecozone, with forests dominated by conifer tree species and areas of wetlands and barrens; agricultural and urban land use is also present. Forest harvesting and insect damage are the primary land cover disturbances in this study area.

Data

Forest inventory data

Each Canadian province produces and maintains some form of forest inventory (Leckie and Gillis, 1995). The attributes recorded for individual mapping units in the forest inventory (polygons) typically include stand species composition, density, height, and age. Much of the information collected in the forest inventory is generated by interpretation of aerial photographs, with field visits taken to ensure the quality and consistency of the interpretation. The Canadian provincial forest inventory programs are developed to address both common and regionally specific needs, often resulting in jurisdiction-specific timelines and standards. The National Forest Inventory (NFI) program encompasses the entire forest area of Canada, applying a regular network of samples within which a standard set of attributes will be collected, maintained, and reported (Gillis et al., 2005). Provincial and territorial agencies are working cooperatively with the Canadian Forest Service to generate the national inventory database (Gillis, 2001). The NFI classification hierarchy includes vegetation cover, tree cover, landscape position, vegetation type, forest density, and species diversity (**Figure 2**) (Gillis, 2001). The minimum mapping unit for the forest inventory data is 2.00 ha.

Classified satellite image data

The Canadian Forest Service and Canadian Space Agency joint project, Earth Observation for Sustainable Development of Forests (EOSD), is using remotely sensed data (Landsat Thematic Mapper (TM) and Enhanced Thematic Mapper (ETM+)) to produce a land cover map of the forested area of Canada representing land cover conditions in the year 2000. Produced through a partnership of federal, provincial, and territorial governments, universities, and industry (Wulder et al., 2003), the land cover product is planned for completion in 2006. The final product, created from over 450 Landsat scenes (Wulder and Seemann, 2001), will require accuracy assessment to ensure the integrity and quality of the output and facilitate the use of the output in a broad range of applications. Planning is currently underway for a sample-based accuracy assessment of the EOSD final product (Wulder et al., 2006).

Landsat-7 ETM+ images were classified following unsupervised hyperclustering and labelling methods summarized in Wulder et al. (2003) and detailed in Wulder et al. (2004). The EOSD products for these areas were developed following the locational constraints and nomenclature of the Canadian National Topographic System (NTS) 1 : 250 000 scale maps (Wulder et al., 2003). NTS map sheet 93 F (Nechako River) represents the Prince George test site and

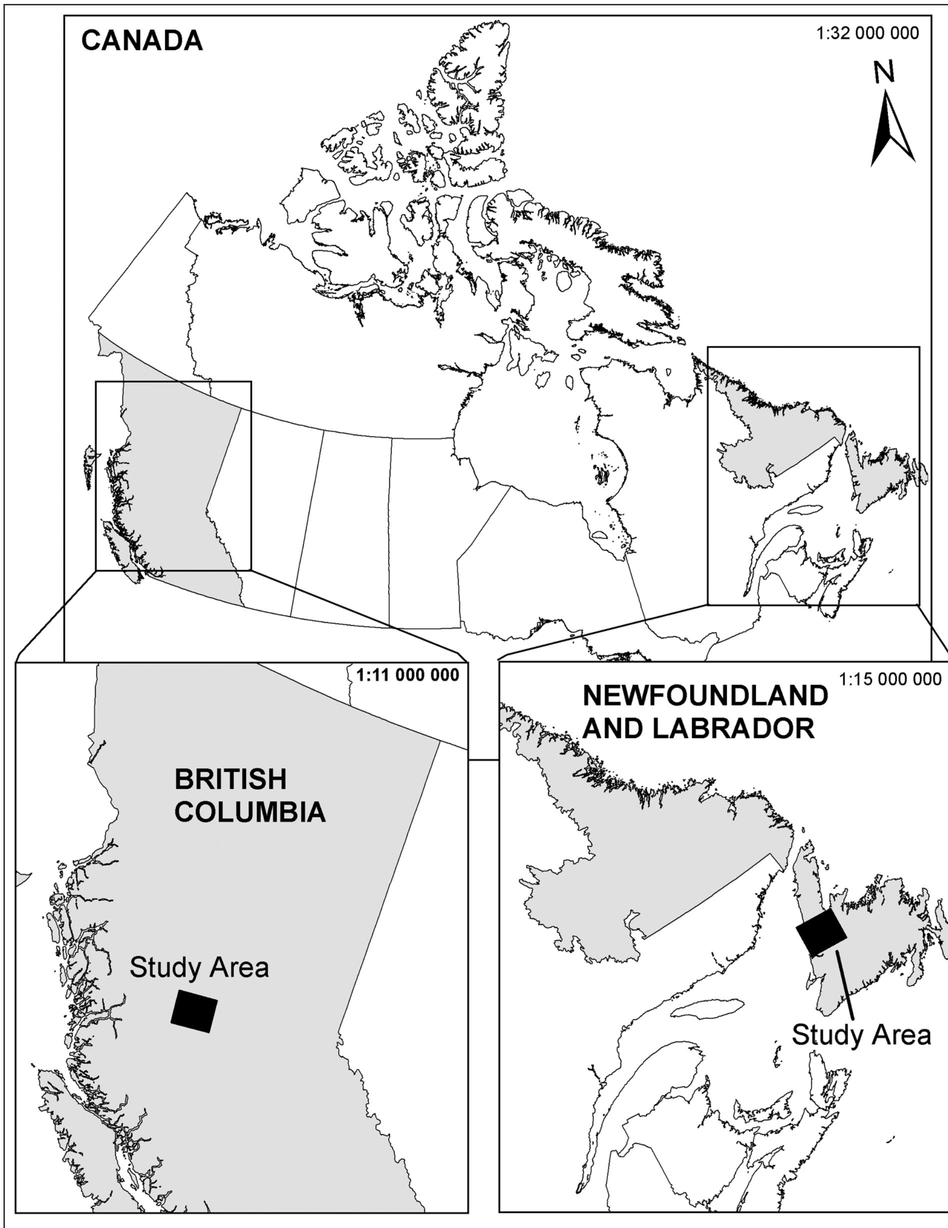

Figure 1. Location of the two study areas in British Columbia and Newfoundland and Labrador, Canada.

contains data from four Landsat-7 ETM+ images: path 50, row 22, 4 August 2000; path 50, row 23, 2 August 1999; path 49, row 22, 12 September 1999; and path 49, row 23, 12 September 1999. NTS map sheet 12 H (Sandy Lake) represents the Deer Lake test site and contains data from four Landsat-7 ETM+ images: path 03, row 26, 2 September 2002; path 04, row 26, 30 May 2000; path 05, row 25, 13 September 2001; and path 05, row 26, 13 September 2001.

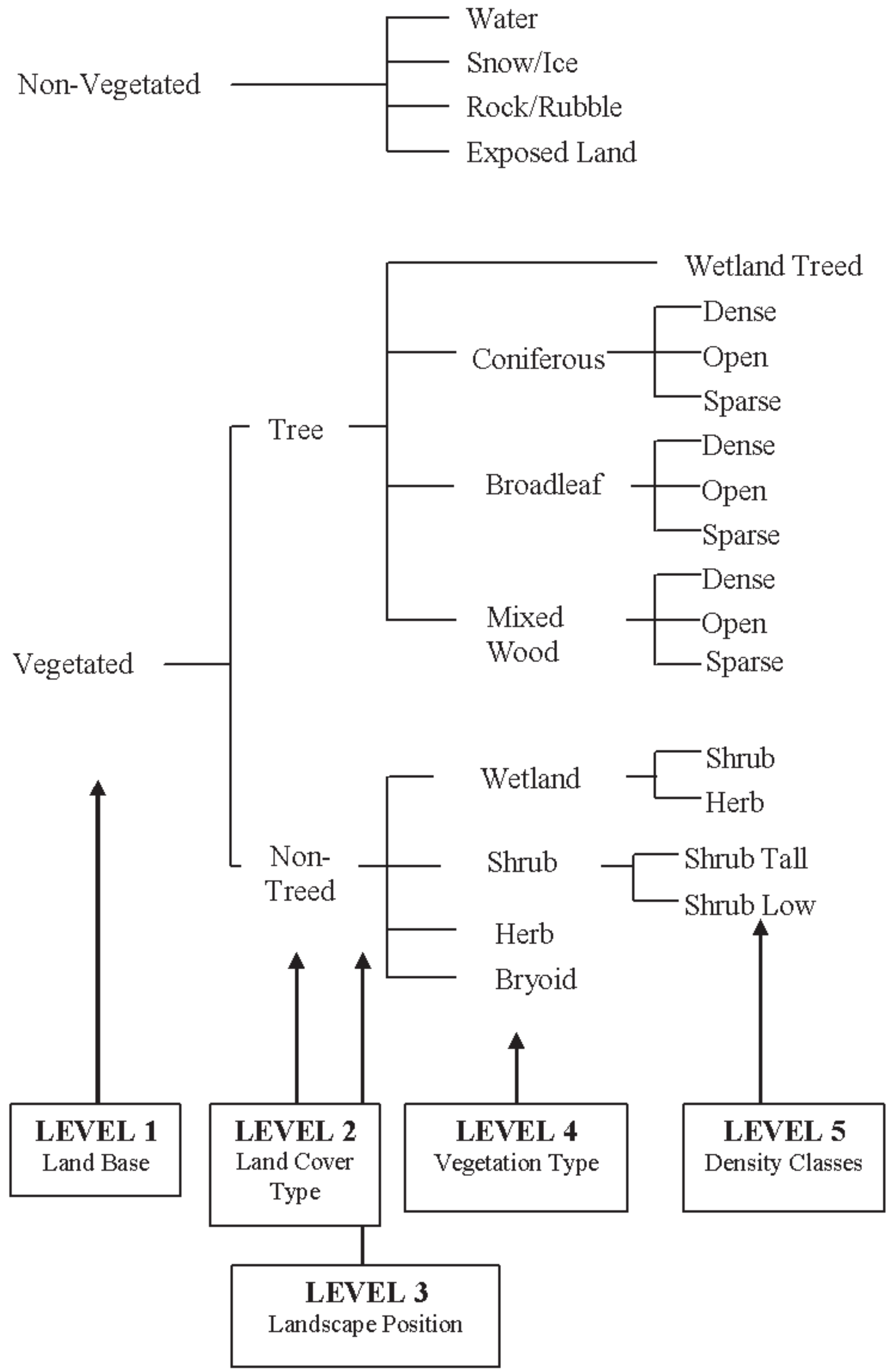

Figure 2. Five of the six levels of the NFI classification hierarchy. The classes in the EOSD legend are identical to level 5 of the NFI scheme, enabling generalization following the nested NFI classification structure. Level 3 is not illustrated here because it does not exist in the same manner in the EOSD and NFI classification systems. In the NFI land cover classification scheme, level 3 refers to landscape position (e.g., alpine, upland, or wetland).

The EOSD classification has 22 classes, which include vegetation density classes of dense, open, and sparse (**Figure 2**). The EOSD classification hierarchy is compatible with the NFI hierarchy but does not include species diversity (Wulder and Nelson, 2003) and was developed to represent a level of detail discernable using Landsat imagery (Wulder et al., 2003; Wulder and Nelson, 2003). The minimum mapping unit for the EOSD product is the 25 m pixel. The integrated and hierarchical nature of the EOSD and NFI legends provides opportunities to generalize land cover patterns to successively coarser levels of the hierarchy, and the analysis of these generalizations can provide useful insights on the nature of the EOSD product (Remmel et al., 2005). For this analysis, vegetation density classes were considered when sample selection was conducted, with 100 samples selected for each class; however, the analysis and reporting of results were restricted to level 4 of the classification hierarchy (vegetation type). Since it is difficult to distinguish between tall and low shrub with Landsat data, these two classes were combined into a single shrub class. Similar difficulties exist when attempting to separate the various wetland classes (treed, shrub, herb) in the classification hierarchy, and these were combined into a single wetland class. Therefore, potential land cover classes for comparison included vegetated classes (coniferous, broadleaf, mixed wood, wetland, shrub, herb, bryoid) and nonvegetated classes (water, snow–ice, rock–rubble, and exposed land).

Methods

Sample design

Sampling units are usually either point or area, and their various forms and relative advantages and disadvantages are detailed in Stehman and Overton (1996) and Stehman and Czaplewski (1998). In this study, a point sampling unit was used (the centroid of a 25 m × 25 m unit, which corresponds to the size of an image pixel and the minimum mapping unit in the EOSD land cover product). The common practice, when using polygonal data for accuracy assessment, is to use the label of the polygon within which the selected sample unit falls, and this approach was followed in this study. However, it must be acknowledged that since a polygon is a generalization of land cover characteristics (the degree of generalization depends on the minimum mapping unit used when the polygons were delineated), this approach will result in conservative bias in the accuracy estimates (Verbyla and Hammond, 1995). In other words, the generalized nature of the polygon data will result in estimates of accuracy that are systematically lower than the actual estimate. In this study, conservative bias was an issue for the protocols that compared the sample unit directly to the polygon label, or which used a 3 × 3 pixel spatial support region. The minimum mapping unit of the NFI was 2.00 ha, the area of the sample unit (25 m pixel) is 0.06 ha, and the area of a 3 × 3 pixel neighbourhood is 0.50 ha.

The direct comparison of the sample unit to the polygon label is further confounded by the complexity of land cover, as captured by remotely sensed data. For example, previous investigations have shown that there is a large range in spectral variability within each forest inventory polygon, with spectral values for different land cover classes occupying the same spectral feature space (Wulder et al., 1999). Furthermore, this spectral variability often increases as proximity to the boundary of an inventory polygon increases (Boudewyn et al., 2000). Due to the limitations associated with using the polygon label, additional insights and approaches for comparing vector-based forest inventory to raster-based image classifications are required. One approach has been to restrict sample selection to areas where land cover is less complex; however, optimistic bias in accuracy estimates can occur when sample selection is restricted to homogeneous areas (Hammond and Verbyla, 1996), resulting in estimates that are systematically greater than the actual estimate.

In this study, samples were selected using a stratified random sampling approach, with the strata being defined by the EOSD land cover classes. In the Prince George study area, eight of the 11 possible NFI level 4 classes were present, including exposed land, water, shrub, herb, wetland, coniferous, broadleaf, and mixed wood. In the Deer Lake study area, nine of the NFI level 4 classes were present, including all the classes found at the Prince George site, with the addition of the rock–rubble class. Although analysis was conducted and results presented at level 4, the sample selection was conducted at level 5. For each land cover class, 100 samples were selected. The total number of samples selected for each protocol varied according to the presence of the individual land cover classes.

Evaluation protocols

Five distinct evaluation protocols were tested to compare the EOSD land cover product and the polygonal forest inventory data (**Figures 3–5**). One protocol directly compared the sample unit with the reference data, two of the protocols used different spatial support regions (SSR) around the sample unit, and the other two protocols restricted the samples that were selected based on homogeneity criteria. For all these protocols, the comparison was between two land cover labels, namely the label of the sampling unit (which was the EOSD 25 m × 25 m pixel) and the label of the NFI polygon within which the sampling unit fell. The first protocol compared the land cover class of the sample unit directly to the label of the forest inventory polygon within which the sample unit was located (**Figure 3**). The second method compared the land cover class mode of a 3 × 3 pixel SSR surrounding the sample unit directly to the label of the inventory polygon within which the sample unit was located (**Figure 4**). The third method compared the land cover class mode of all the EOSD pixels in the forest inventory polygon within which the sample unit was located directly to the label of the inventory polygon (**Figure 4**).

The remaining two evaluation protocols compared the EOSD sample directly to the forest inventory polygon label as per the first method described previously; however, homogeneity criteria restricted sample selection to relatively homogeneous

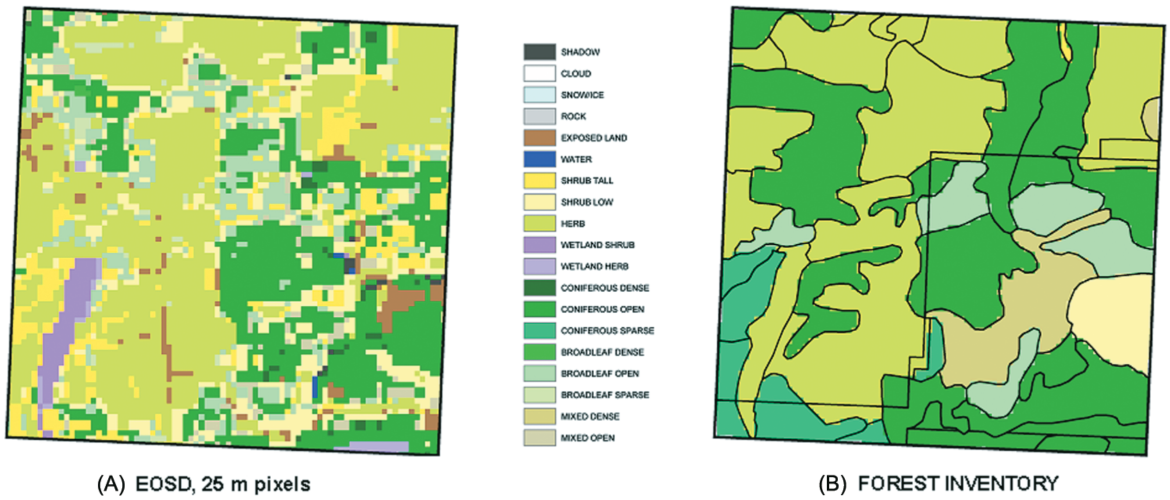

Figure 3. Illustration of the protocol used to compare the land cover classification of the EOSD 25 m sample pixel (A) directly to the land cover classification of the forest inventory polygon (B). Note the heterogeneity of the land cover product generated from the remotely sensed data (A), in comparison to the generalized nature of the forest inventory polygons (B).

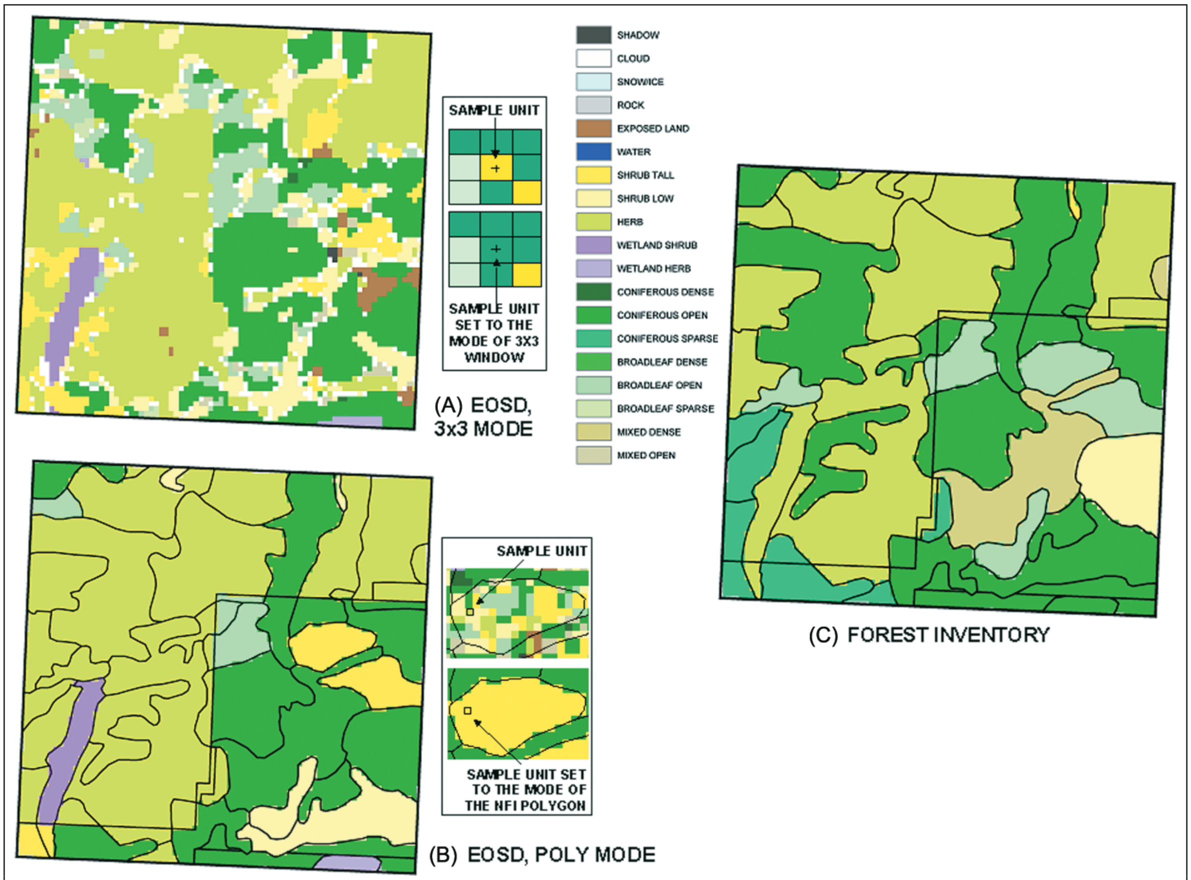

Figure 4. Illustration of the protocol used to compare the land cover classification mode of a 3 × 3 pixel neighbourhood around the EOSD sample (A) to the land cover classification of the forest inventory polygon (C). Also shown is the protocol used to compare the land cover classification mode of EOSD classes found within the forest inventory polygon (B) to the land cover classification of the original forest inventory polygon (C).

areas, an approach previously reported for accuracy assessments of other large-area land cover products (e.g., Stehman et al., 2003; Wickham et al., 2004). For the fourth method, therefore, only those pixel samples found within a completely homogeneous 3 × 3 pixel neighbourhood (**Figure 5**) were used for analysis. Similarly, for the fifth method, only those pixel samples found within polygons having less than a predetermined number of classes were selected (**Figure 5**). Homogeneity criteria for this fifth method were determined by examining the distribution of samples and the number (variety) of classes in the polygon. This subjective determination of homogeneity for the polygons was necessary, since the likelihood of homogeneous polygons (e.g., polygons containing only one EOSD class) was low (given the differences in the representation of land cover; see **Figure 3**). A trade-off was made between restricting the number of different classes found within the polygon and retaining approximately half of the available samples. For the Prince George study area, samples were restricted to homogeneous polygons with eight or fewer classes (which represented 52% of the total number of samples) (**Table 1**). For the Deer Lake study area, the homogeneity criterion was set to 13 or fewer classes (which represented 52% of the total number of samples) (**Table 2**).

Analysis and estimation

To generate unbiased estimates of accuracy, the stratified sampling approach used in this study necessitated that the raw counts from the error matrix be converted to estimated joint probabilities prior to estimation (Czaplewski, 2003). Joint probabilities account for the proportion of samples in the strata that are mapped correctly and the proportion of the total study area occupied by the strata. The appropriate estimators for stratified random sampling are provided in Czaplewski (2003). To assist in the interpretation of the results, confidence intervals were produced for the estimates of overall accuracy (at the 95% confidence level).

Results and discussion

The total number of samples used for each of the five evaluation protocols and the overall accuracy results for the

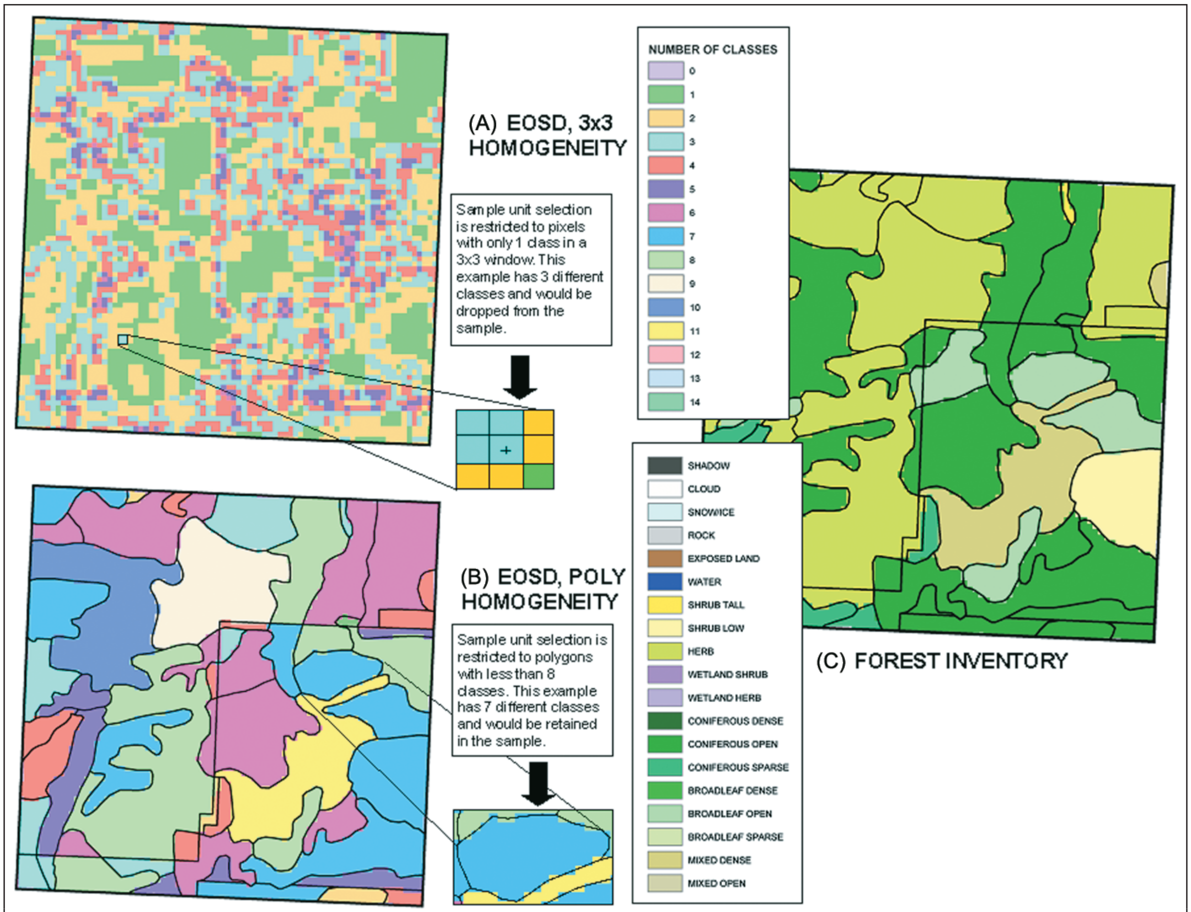

Figure 5. Illustration of the remaining two protocols used to compare the sample land cover classification to the forest inventory. In (A), sample selection is restricted to pixels that fall in homogeneous 3 × 3 neighbourhoods. The classifications for these "homogeneous" or "pure" samples are compared to those for the forest inventory (C), as in **Figure 3**. Similarly in (B), sample selection is restricted to those forest inventory polygons that have fewer than seven different land cover classes. These homogeneous or pure polygons are then compared to the land cover attribute for the original forest inventory (C).

Table 1. Determination of homogeneity criterion in the Prince George study area.

No. of classes within the inventory polygon	No. of samples affected	Percentage of samples affected	Cumulative percentage of samples affected
1	9	0.60	0.60
2	16	1.10	1.80
3	49	3.50	5.30
4	72	5.10	10.40
5	105	7.50	17.90
6	132	9.40	27.40
7	175	12.50	39.90
8	175	12.50	52.40
9	210	15.00	67.40
10	223	15.90	83.30
11	123	8.80	92.10
12	56	4.00	96.10
13	42	3.00	99.10
14	6	0.40	99.50
15	7	0.50	100.00
Total	1400	100.00	

Table 2. Determination of homogeneity criterion in the Deer Lake study area.

No. of classes within the inventory polygon	No. of samples affected	Percentage of samples affected	Cumulative percentage of samples affected
1	1	0.07	0.07
2	8	0.57	0.64
3	8	0.57	1.21
4	11	0.79	2.00
5	26	1.86	3.86
6	47	3.36	7.21
7	66	4.71	11.93
8	67	4.79	16.71
9	80	5.71	22.43
10	98	7.00	29.43
11	96	6.86	36.29
12	103	7.36	43.64
13	114	8.14	51.79
14	97	6.93	58.71
15	135	9.64	68.36
16	89	6.36	74.71
17	354	25.29	100.00
Total	1500	100.00	

Prince George, British Columbia, study area are presented in **Table 3**. A separate set of samples was selected for each of the accuracy assessment protocols, with the exception of the homogeneity measures, whose samples were subset from samples collected for the mode of the 3 × 3 pixel SSR and the polygon SSR. For the 3 × 3 pixel homogeneity measure, 311 sample units (out of 1476) were found within a homogeneous 3 × 3 pixel SSR (e.g., all nine surrounding pixels are of the same land cover class). For the polygon homogeneity measure, 733 samples (out of 1400) were found within polygons that have eight or fewer different land cover classes present.

Overall accuracy in the Prince George study site ranged from 34% (when directly comparing the EOSD sample to the inventory polygon label) to 58% (when comparing a 3 × 3 homogeneous EOSD sample to the inventory polygon label). This range of accuracies reflects the conservative bias associated with the protocols used to compare the EOSD and NFI products (Verbyla and Hammond, 1995). The use of the mode of a 3 × 3 pixel neighbourhood emulates the application of a 3 × 3 pixel filter to the EOSD product (**Figure 4**). In this study area, the generalization associated with using the modal class of a 3 × 3 pixel neighbourhood increased the estimate of overall accuracy by 3% (to 37%). A further 12% increase (to 49%) in overall accuracy resulted from using the mode of all the EOSD pixels found within the inventory polygon in which the sample unit was found, effectively matching the level of generality associated with the NFI to the EOSD product (**Figure 4**).

Restricting sample selection to polygons with eight or fewer classes at the Prince George site results in an increase in overall accuracy from 34% to 51%. Similarly, restricting sample selection to homogeneous 3 × 3 neighbourhoods resulted in an increase in overall accuracy to 58%, reflecting some of the optimistic bias described by Hammond and Verbyla (1996). The homogeneity criteria substantially decreased the sample size available for estimation (and thereby decreased the width of the confidence intervals associated with these estimates); only 21% of 1476 samples were found in homogeneous 3 × 3 neighbourhoods, and 52% of 1400 samples were found in polygons with eight or fewer different land cover classes.

Measures of producer and user accuracy were estimated using methods for stratified random samples provided in Czaplewski (2003). Selecting samples based on the strata of mapped classes (e.g., from the EOSD product) favours precision in the estimation of user accuracy (Stehman et al., 2003). At the Prince George study site, the user accuracy for coniferous forest (which represents 58% of the total study area) was consistently greater than or equal to 88% for all of the protocols (**Table 4**). Water also had consistently high user accuracies, resulting from its distinct spectral response relative to that of the other classes. Estimates of user accuracy for other classes are highly variable.

The overall accuracy results for the Deer Lake site are presented in **Table 5**. Overall accuracy ranged from 34% (when directly comparing the EOSD sample to the inventory polygon label) to 55% (when comparing a 3 × 3 homogeneous EOSD sample to the inventory polygon label). In the Deer Lake study area, the use of the mode for a 3 × 3 pixel neighbourhood surrounding the sample unit resulted in a 5% increase (to 38%) in overall accuracy over the direct comparison of sample unit to polygon label. A 7% increase (to 41%) in overall accuracy

Table 3. Number of samples used for each of the five evaluation protocols, the resulting overall accuracies, and associated confidence intervals for the Prince George study area.

EOSD class	Pixel	3 × 3 mode	Polygon mode	3 × 3 homogeneity	Polygon homogeneity
Water	100	100	100	49	25
Exposed land	100	100	100	88	93
Shrub	200	200	200	14	55
Herb	100	100	100	43	46
Wetland	200	200	200	38	91
Coniferous	300	300	300	67	209
Broadleaf	300	300	300	12	122
Mixed wood	200	176	100	0	92
Total	1500	1476	1400	311	733
Overall accuracy (%)	34.31	37.37	48.99	58.08	51.20
95% confidence interval (%)	±2.04	±2.10	±2.23	±4.76	±3.11

Table 4. Individual class accuracies (%) for the Prince George study area.

	Pixel		3 × 3 mode		Polygon mode		3 × 3 homogeneity		Polygon homogeneity	
EOSD class	P	U	P	U	P	U	P	U	P	U
Water	92.23	91.67	95.05	88.21	98.00	97.61	98.90	84.41	98.91	93.81
Exposed land	19.23	36.07	25.00	39.72	14.29	33.14	40.00	39.74	—	—
Shrub	13.33	12.37	22.07	11.79	28.94	16.39	—	—	15.56	12.98
Herb	7.97	13.25	12.68	20.84	8.02	7.93	35.30	35.05	13.85	12.50
Wetland	—	—	—	—	—	—	—	—	—	—
Coniferous	38.81	88.58	49.05	92.89	61.35	98.61	68.58	94.48	63.51	98.72
Broadleaf	41.87	17.48	88.56	16.23	47.85	14.34	66.67	46.52	33.64	13.54
Mixed wood	54.41	9.28	50.97	4.21	81.82	7.90	—	—	66.67	8.90

Note: P, producer accuracy for the class; U, user accuracy for the class.

Table 5. Number of samples used for each of the five evaluation protocols, the resulting overall accuracies, and associated confidence intervals for the Deer Lake study area.

EOSD class	Pixel	3 × 3 mode	Polygon mode	3 × 3 homogeneity	Polygon homogeneity
Water	100	100	100	94	25
Exposed land	100	100	100	3	5
Rock–rubble	100	100	100	13	23
Shrub	200	200	200	14	86
Herb	100	100	100	12	58
Wetland	300	300	300	2	202
Coniferous	300	300	300	66	159
Broadleaf	300	200	200	10	130
Mixed wood	200	200	100	4	59
Total	1700	1600	1500	218	725
Overall accuracy (%)	33.89	38.19	40.82	54.63	35.29
95% confidence interval (%)	±1.92	±2.03	±2.12	±5.78	±2.93

resulted from using the mode of the EOSD pixels found within the inventory polygon.

At the Deer Lake site, restriction of sample selection to homogeneous polygons with 13 or fewer EOSD classes resulted in an increase in overall accuracy of 1% (to 35%), which contrasts with the 17% increase in overall accuracy for the Prince George study site. The overall accuracy in homogeneous 3 × 3 neighbourhoods was 55%. The homogeneity criteria considerably reduced the sample size: the 3 × 3 homogeneity criteria retained 14% of the original 1600 samples, and the polygon homogeneity criteria retained 52% of the original samples. As with the Prince George data, individual class

Table 6. Individual class accuracies (%) for the Deer Lake study area.

EOSD class	Pixel		3 × 3 mode		Polygon mode		3 × 3 homogeneity		Polygon homogeneity	
	P	U	P	U	P	U	P	U	P	U
Water	82.20	99.93	87.50	99.72	85.96	99.67	97.90	100.00	78.13	99.92
Exposed land	24.31	13.21	15.09	1.33	59.12	54.07	—	—	8.70	7.60
Rock–rubble	35.82	14.17	5.71	6.72	32.43	31.58	—	—	18.52	12.14
Shrub	18.92	35.43	28.57	7.10	42.86	2.46	—	—	30.77	47.75
Herb	30.77	13.92	25.00	8.96	37.08	6.23	25.00	6.23	—	—
Wetland	43.89	14.66	50.80	32.90	48.58	18.23	33.33	82.28	—	—
Coniferous	26.67	77.84	32.51	84.54	38.52	89.09	76.02	94.16	31.56	84.52
Broadleaf	67.67	4.62	52.36	4.29	73.60	28.72	28.79	5.35	20.93	97.84
Mixed wood	30.13	32.84	48.80	18.50	39.74	26.31	9.21	23.26	45.09	19.12

Note: P, producer accuracy for the class; U, user accuracy for the class.

accuracies for Deer Lake were variable. The user accuracy for coniferous forest (which represents 44% of the total study area) was consistently greater than 77% (**Table 6**). As in the other study area, water also has consistently high user accuracies.

In both study sites, the use of a modal class for a 3 × 3 pixel neighbourhood increased overall accuracy by an average of 4%, and the restriction of estimation to homogeneous 3 × 3 pixel neighbourhoods increased overall accuracy by an average of 22%. Stehman et al. (2003) employed similar protocols when estimating accuracy for the 1992 National Land Cover Data for the eastern United States. On average, the use of a modal class for a 3 × 3 pixel neighbourhood surrounding the sample pixel increased overall accuracy by 4%, and the restriction of sample selection to homogeneous 3 × 3 pixel neighbourhoods increased overall accuracy by an average of 18%. This increase in accuracy reflects the optimistic bias described by Hammond and Verbyla (1996).

The use of the modal class within a forest inventory polygon increased accuracy by an average of 11% (15% in Prince George and 7% in Deer Lake). Restricting the accuracy analysis to "relatively" homogeneous polygons resulted in an average increase in overall accuracy of 9% (17% in Prince George and 1% in Deer Lake). The differences in the gain in overall accuracy between these two sites may be partly attributable to differences in polygon sizes. Although the mean polygon size (for non-water polygons) in both study sites is similar (10.28 ha in Prince George and 11.14 ha in Deer Lake), the standard deviations differ by 130.00 ha (19.00 ha in Prince George and 149.00 ha in Deer Lake).

The impacts of both conservative and optimistic biases in accuracy assessment were found in this analysis (Verbyla and Hammond, 1995; Hammond and Verbyla, 1996). The fundamental difference in both the structure and design of the EOSD and forest inventory products is evidenced by the overall accuracies reported. In an effort to make the EOSD and inventory products more compatible for validation, the detail inherent in the EOSD product must be compromised. Specifically, the protocol whereby the land cover class mode of all the EOSD pixels found within the forest inventory polygon is used for comparison to the land cover class of the inventory polygons produced the greatest level of overall accuracy (without restriction of sample selection by homogeneity criteria). This method addresses conservative bias by matching the level of detail in the sample unit to that in the minimum mapping unit of the reference data. It must be noted, however, that end-users who subscribe to the accuracy estimates achieved with this protocol should be using a product that reflects the level of generalization associated with this method (Czaplewski, 2003). The spatial support region used to generate accuracy estimates must be transparent and explicitly communicated in conjunction with the results of the accuracy assessment. For instance, if accuracy assessment of the classified imagery is undertaken with a 3 × 3 modal filter, the accuracy definition should state that the accuracies reported pertain to a product that was generalized with a 3 × 3 modal filter.

Other data sources could have been used for the validation protocols applied in this study; however, the NFI was selected for its concordance with the legend of the land cover product, in terms of both its spatial coverage and land cover classification scheme (Wulder and Nelson, 2003; Remmel et al., 2005). In addition, the standardized nature of the inventory design across Canada simplified the acquisition of validation data in both of the study sites. It must be emphasized that the NFI is used in this study as an example of a preexisting polygonal data source, and that this work was not intended to be prescriptive with regards to how the NFI may be optimized for use as validation data for the EOSD land cover product. Rather, the focus of this work was more general and concentrated on issues that arise when a preexisting polygonal data source is used as a source of validation data for a raster-based land cover product.

The results of this study suggest that the use of a preexisting polygonal data source to validate a raster-based land cover product requires that the land cover product be generalized to a level comparable with that of the validation data to achieve estimates that are within suggested acceptable accuracy targets (Wulder et al., 2006). This generalization of the land cover product may or may not be desirable depending on the intended application. Without generalization, the estimated accuracies may not truly represent the actual accuracy of the product.

As noted by Remmel et al. (2005), vector and raster datasets can be integrated fruitfully for data audit, bias detection, update needs, and quality assurance. Cross-validation between the datasets can emphasize the relative comparison between the datasets rather than the absolute differences common to accuracy assessment. To validate the accuracy of large-area land cover products generated from remotely sensed data, the use of purpose-collected validation data is recommended. Although the cost of collecting purpose-acquired data for validation may be perceived as high, the costs of compromising the level of detail or integrity of the land cover product must also be considered (particularly in reference to the expense associated with producing a large-area land cover product). Cost-effective purpose-acquired data may be collected using digital camera or video equipment combined with a differentially corrected global positioning system (DGPS). The digital photograph or captured video frame can be consistently attributed according to the classification scheme of the land cover product. Depending on the timing of the collection of the aerial survey data, calibration data may also be collected that could be used in the classification process directly. The use of such technology, combined with a well-planned sampling protocol (Stehman and Czaplewski, 1998), can help to overcome many of the limitations associated with the use of preexisting, vector-based data for accuracy assessment purposes.

Conclusion

The use of preexisting polygon data for the validation of large-area land cover products is often perceived as a cost-effective means of associating some level of confidence with the land cover product. The results of this analysis illustrate the difficulties associated with such a validation approach; point-based approaches for relating pixels to forest inventory data are impacted by both spatial and categorical mismatches. Accounting for class heterogeneity present in the imagery (constrained by neighbourhood windows or polygons) and differences in spatial support resulted in improved relationships between the forest inventory and the satellite map data. Given the wide range of accuracies reported in this study, the results suggest that caution is needed when making spatially explicit comparisons between raster and vector datasets. Moreover, when necessary, validation approaches accounting for the heterogeneity of classes within a given polygon are recommended. The fundamental difference between these data types, however, may result in an accuracy assessment that compromises the level of detail and integrity of one or both of the datasets. It is for this reason that the use of purpose-acquired validation data, where plausible, is preferable to the use of preexisting data sources.

Although not directly comparable for validation purposes, the forest inventory data and land cover product described in this paper should be considered complementary. Synergies between the classified satellite image data and forest inventory datasets must continue to be explored, such as in the case of cross-validation for statistically based estimation (area-based estimation as opposed to spatially explicit estimation), audit, and bias detection. The research reported in this communication encourages consideration of purpose-acquired validation data such as video data, coupled with well-designed sampling approaches, to characterize the accuracy of large-area land cover products.

Acknowledgements

This research is aided through funding of the Canadian Space Plan and the Government Related Initiatives Program (GRIP) of the Canadian Space Agency and Natural Resources Canada, Canadian Forest Service. The British Columbia Ministry of Forests and Range is acknowledged for provision of the Prince George study site data. The Newfoundland Forest Service and Gros Morne National Park are thanked for provision of forest inventory databases that were translated to NFI classes for the purpose of this study. Two anonymous reviewers are thanked for offering suggestions that improved the final manuscript.

References

Atkinson, P., and Curran, P.J. 1995. Defining an optimal size of support for remote sensing investigations. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 33, pp. 768–776.

Boudewyn, P., Seemann, D., Wulder, M., and Magnussen, S. 2000. The effects of polygon boundary pixels on image classification accuracy. In *Proceedings of the 22nd Symposium of the Canadian Remote Sensing Society*, 20–25 August 2000, Victoria, B.C. Canadian Aeronautics and Space Institute (CASI), Ottawa, Ont. pp. 637–643.

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

Czaplewski, R.L. 2003. Statistical design and methodological considerations for the accuracy assessment of maps of forest condition. In *Remote sensing of forest environments: concepts and case studies*. Edited by M.A. Wulder and S.E. Franklin. Kluwer Academic Publishers, Boston Mass. pp. 115–141.

Ehlers, M., Gähler, M., and Janowsky, R. 2003. Automated analysis of ultra high resolution remote sensing data for biotope mapping: new possibilities and challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 57, pp. 315–326.

Foody, G.M. 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, Vol. 90, pp. 185–201.

Gillis, M.D. 2001. Canada's National Forest Inventory (responding to current information needs). *Environmental Monitoring and Assessment*, Vol. 67, pp. 121–129.

Gillis, M.D., Omule, A.Y., and Brierley, T. 2005. Monitoring Canada's forests: the National Forest Inventory. *Forestry Chronicle*, Vol. 81, pp. 214–221.

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

Homer, C., Ramsey, R., Edwards, T., Jr., and Falconer, A. 1997. Landscape cover-type modeling using a multi-scene thematic mapper mosaic. *Photogrammetric Engineering & Remote Sensing*, Vol. 63, pp. 59–67.

Leckie, D., and Gillis, M. 1995. Forest inventory in Canada with an emphasis on map production. *Forestry Chronicle*, Vol. 71, No. 1, pp. 74–88.

Liu, H., and Zhou, Q. 2004. Accuracy analysis of remote sensing change detection by rule-based rationality evaluation with post-classification comparison. *International Journal of Remote Sensing*, Vol. 25, pp. 1037–1050.

Loveland, T.R., and Belward, A.S. 1997. The International Geosphere Biosphere Programme Data and Information System global land cover data set (DISCover). *Acta Astronautica*, Vol. 41, pp. 681–689.

McConnell, W. 2002. Meeting in the middle: the challenge of meso-level integration. *Land Use Policy*, Vol. 19, pp. 99–101.

Meidinger, D., and Pojar, J. (Editors). 1991. *Ecosystems of British Columbia*. British Columbia Ministry of Forests, Special Report Series 6. 330 pp.

Remmel, T.K., and Perera, A. 2002. Accuracy of discontinuous binary surfaces: a case study using boreal forest fires. *International Journal of Geographical Information Science*, Vol. 16, pp. 287–298.

Remmel, T.K., Csillag, F., Mitchell, S., and Wulder, M. 2005. Integration of forest inventory and satellite imagery: a Canadian status assessment and research issues. *Forest Ecology and Management*, Vol. 207, pp. 405–428.

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

Stehman, S.V., and Czaplewski, R.L. 2003. Introduction to special issue on map accuracy. *Environmental and Ecological Statistics*, Vol. 10, pp. 301–308.

Stehman, S.V., and Overton, W.S. 1996. Spatial sampling. In *Practical handbook of spatial statistics*. Edited by S.L Arlinghaus and D.A. Griffith. CRC Press, Boca Raton, Fla. pp. 31–63.

Stehman, S.V., Wickham, J.D., Smith, J.H., and Yang, L. 2003. Thematic accuracy of the 1992 land cover data for the eastern United States: statistical methodology and regional results. *Remote Sensing of Environment*, Vol. 86, pp. 500–516.

Stone, T., Schlesinger, P., Houghton, T., and Woodwell, G. 1994. A map of the vegetation of South America based on satellite imagery. *Photogrammetric Engineering & Remote Sensing*, Vol. 60, pp. 541–551.

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

Wickham, J.D., Stehman, S.V., Smith, J.H., and Yang, L. 2004. Thematic accuracy of the 1992 national land-cover data for the western United States. *Remote Sensing of Environment*, Vol. 91, pp. 442–468.

Wulder, M.A., and Nelson, T. 2003. *EOSD land cover classification legend report, version 2*. Pacific Forestry Centre, Canadian Forest Service, Victoria, B.C. 83 pp. Available from http://www.eosd.cfs.nrcan.gc.ca/cover/EOSD_legend_report-v2.pdf [accessed on 25 October 2005].

Wulder, M.A., and Seemann, D. 2001. Spatially partitioning Canada with the LANDSAT worldwide referencing system. *Canadian Journal of Remote Sensing*, Vol. 27, No. 3, pp. 225–231.

Wulder, M.A., Magnussen, S., Boudewyn, P., and Seemann, D. 1999. Spectral variability related to forest inventory polygon stored within a GIS. In *Proceedings of the IUFRO Conference on Remote Sensing and Forest Monitoring*, 1–3 June 1999, Rogow, Poland. Edited by T. Zawila-Niedzwiecki and M. Brach. European Commission and Warsaw Agricultural University, Warsaw, Poland. pp. 141–153.

Wulder, M.A., Dechka, J.A., Gillis, M.A., Luther, J.E., Hall, R.J., Beaudoin, A., and Franklin, S.E. 2003. Operational mapping of the land cover of the forested area of Canada with Landsat data: EOSD land cover program. *Forestry Chronicle*, Vol. 79, pp. 1075–1083.

Wulder, M.A., Cranny, M., Dechka, J., and White, J. 2004. *An illustrated methodology for land cover mapping of forests with Landsat-7 ETM+ data: methods in support of EOSD land cover, version 3*. Pacific Forestry Centre, Canadian Forest Service, Victoria, B.C. 35 pp. Available from http://www.eosd.cfs.nrcan.gc.ca/cover/methods_vers3_e.pdf [accessed 25 October 2005].

Wulder, M.A., Franklin, S., White, J., Linke, J., and Magnussen, S. 2006. An accuracy assessment framework for large area land cover classification products derived from medium resolution satellite data. *International Journal of Remote Sensing*, Vol. 27, No. 4, pp. 663–683.