Article

Comparison of Simple Averaging and Latent Class Modeling to Estimate the Area of Land Cover in the Presence of Reference Data Variability

Dingfan Xing [1], Stephen V. Stehman [1,*], Giles M. Foody [2] and Bruce W. Pengra [3]

1 Department of Sustainable Resources Management, SUNY ESF, 1 Forestry Drive, Syracuse, NY 13210, USA; dixing@esf.edu
2 School of Geography, University of Nottingham, Room C7 Sir Clive Granger, University Park, Nottingham NG7 2RD, UK; giles.foody@nottingham.ac.uk
3 KBR, Contractor to the U.S. Geological Survey, Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD 57198, USA; bpengra@contractor.usgs.gov
* Correspondence: svstehma@esf.edu; Tel.: +1-315-470-6692

Abstract: Estimates of the area or percent area of the land cover classes within a study region are often based on the reference land cover class labels assigned by analysts interpreting satellite imagery and other ancillary spatial data. Different analysts interpreting the same spatial unit will not always agree on the land cover class label that should be assigned. Two approaches for accommodating interpreter variability when estimating the area are simple averaging (SA) and latent class modeling (LCM). This study compares agreement between area estimates obtained from SA and LCM using reference data obtained by seven trained, professional interpreters who independently interpreted an annual time series of land cover reference class labels for 300 sampled Landsat pixels. We also compare the variability of the LCM and SA area estimates over different numbers of interpreters and different subsets of interpreters within each interpreter group size, and examine area estimates of three land cover classes (forest, developed, and wetland) and three change types (forest gain, forest loss, and developed gain). Differences between the area estimates obtained from SA and LCM are most pronounced for the estimates of wetland and the three change types. The percent area estimates of these rare classes were usually greater for LCM compared to SA, with the differences between LCM and SA increasing as the number of interpreters providing the reference data increased. The LCM area estimates generally had larger standard deviations and greater ranges over different subsets of interpreters, indicating greater sensitivity to the selection of the individual interpreters who carried out the reference class labeling.

Keywords: land cover monitoring; sampling; Landsat; LCMAP; remote sensing

1. Introduction

Land cover and land cover change are fundamental variables with great importance to natural environmental science, as well as critical factors that impact global and regional climate [1]. Therefore, estimating the area of land cover and land cover change plays a prominent role in studies of climate change, carbon cycling, and biodiversity conservation. Area estimation is also a key element of land cover monitoring. For example, in a national land cover monitoring program for the time period 2000 to 2020, it would be of interest to know the area of forest cover (km^2) and the percent area of forest cover in 2000, 2010, and 2020, as well as the area and the percent area of change in forest cover from 2000 to 2010 and 2010 to 2020. Area estimates for other years and time intervals would, of course, be of interest for all land cover classes included in the monitoring objectives.

The good practice recommendations for area estimation [2] specify: (1) Selecting a probability sample of pixels or other spatial units; (2) obtaining the reference class of each

sample unit; and (3) estimating the area of land cover or land cover change based on these reference sample data. Using the reference classification as the basis for area estimation is recommended to avoid the bias of pixel counting (i.e., summing the area of all pixels mapped as the target land cover class). Olofsson et al. [2,3] provide formulas for estimating the area and associated standard errors for the common special case of stratified random sampling, while Stehman [4] and Gallego [5] provide comprehensive overviews of the general topic of remote sensing-based methodology for estimating the area of land cover and land cover change.

The reference class land cover and land cover change labels assigned to the spatial units of the assessment (e.g., Landsat pixels) are often obtained by human interpretation of satellite imagery and other ancillary data [2]. In our study, we limit attention to the case where the reference class label is a single, nominal attribute (i.e., a hard classification). Ideally, this reference class label would represent the true ground condition, and as such, would be observed without error or variability among interpreters. In reality, interpreters may disagree when assigning the reference class label for a given sample unit. For example, Powell et al. [6] examined agreement among five interpreters providing reference labels for 790 sample pixels in Brazil and reported about 30% interpreter disagreement for a legend of five land cover classes. Pengra et al. [7] selected a simple random subsample from a sample of 12,000 reference pixels, and the pixels in the subsample were independently interpreted by a second analyst. For an eight-class land cover legend, the overall agreement between the duplicated interpretations was 88%, with the smallest agreement of 46% occurring for the disturbed class and the highest agreement of 94% occurring for water. Foody and Boyd [8] reported agreement among pairs of volunteer interpreters ranged from 63% to 80% when assigning forest/non-forest reference labels interpreted from ground-based photographs in West Africa.

Given that reference data may be imperfect, due to variability among interpreters, we investigated two options for estimating area in the presence of this variability, simple averaging, and latent class modeling. In simple averaging (SA), all interpreters obtain a reference land cover class label for every unit in the sample, and the percent area of a land cover class can be estimated from the data of each individual interpreter. The final percent area estimate from SA combines the sample data from all interpreters by taking the mean (average) of the individual interpreter estimates. For example, suppose three interpreters assign reference land cover class labels to each pixel of a simple random sample of 100 pixels, and the percent areas of forest cover estimated from the sample were 43%, 40%, and 49% for the three interpreters. Then the SA estimated percent area of forest cover would be 44%.

Latent class modeling (LCM) combines the results of multiple interpreters to obtain an estimate of the prevalence of a land cover class, such as forest or developed in situations where there is no "gold standard" reference classification [9,10]. The fundamental premise of LCM is that disagreement among interpreters represents imperfect recognition of the true land cover class. However, from the associations among the land cover labels assigned by multiple interpreters, the true land cover, in the form of a latent class variable, can be identified. In other words, the true land cover cannot be directly observed, but it can be determined from the observed land cover class labels provided by the interpreters (see Section 2.3).

An early application of LCM to land cover studies demonstrated that LCM could potentially reduce the impacts of reference data error on estimates of producer's and user's accuracies of change, as well as estimates of the area of change [9]. Foody and Boyd [8] applied LCM to estimate the area of forest cover using class labels provided by four volunteers and the class labels from the GlobCover map [11]. They found that the percent area estimate of 44.4% for the forest cover class produced from LCM showed strong agreement with the estimate of 42.7% produced from SA of the four volunteer interpreter estimates and with the estimate of 40.4% produced using the reference class most frequently assigned to each pixel by the set of interpreters. These results for a single case study [8]

demonstrated the viability of LCM for estimating area and set the stage for our study in which we greatly extend the comparison of SA with LCM to include multiple land cover and change classes.

In addition to SA and LCM, other approaches to combining reference data from multiple interpreters include: (1) having the interpreters reach a consensus decision; (2) using the dominant reference class label; (3) using the majority (>50% of the interpreters) class label; or (4) having a "super interpreter" resolve inconsistent cases. A unique dominant or majority reference class may not always exist (e.g., two classes tie), and consensus and super interpreter approaches cannot readily be compared objectively to other methods because they depend on processes that are not easily replicated (e.g., the expertise of the super interpreter). SA is a commonly used approach, and LCM offers a novel modeling approach to accommodate reference data variability. Therefore, in this article, our focus is limited to SA and LCM and addresses the impact of reference data variability on the area estimates obtained using these two approaches. Given the absence of gold standard reference data, it is not possible to determine whether either SA or LCM provides an unbiased estimator of area. However, we can quantify the degree to which the area estimates from the two approaches differ, and we do so for several land cover classes and land cover change types. In addition, we compare the variability of the area estimates over different numbers of interpreters (i.e., interpreter group size) and over different subsets (combinations) of interpreters. The latter variability over different combinations of interpreters for each group size is particularly important because an area estimation technique that is less sensitive to the subset of interpreters selected is preferable.

2. Materials and Methods

2.1. Reference Data

The reference data used for the comparison of LCM and SA area estimates were collected by the U.S. Geological Survey's (USGS) Land Change Monitoring, Assessment, and Projection (LCMAP) initiative [12]. The objectives of LCMAP are to quantify and map land cover and land cover change within the conterminous United States (US) on an annual basis starting from 1985. Two key elements of LCMAP are estimating the area of land cover for each year (e.g., area of water, tree cover, cropland, wetland, etc.) and estimating the annual change in the area of the land cover classes. The LCMAP area estimates are produced from annual reference land cover labels determined by professional interpreters for a simple random sample of nearly 25,000 pixels (30 m × 30 m). Prior to initiating reference data labeling Supplementary Materials of the sample pixels, LCMAP conducted a pilot study in which seven trained interpreters independently assessed a simple random sample of 300 pixels from the Puget Sound region of Washington State within the US (Figure 1). These pilot study data were used for our comparative analyses of SA and LCM. From the seven land cover classes of the pilot study data, we selected forest and developed to represent common classes and wetland to represent a rare class (see Appendix A for definitions of the land cover classes). We also compared SA and LCM area estimates for three land cover change types, forest loss, forest gain, and developed gain.

Reference class interpretation protocols were defined in a Joint Response Design (JRD) document [7]), and the TimeSync reference data collection tool [13] was adapted to collect the variables defined in the JRD. TimeSync includes Landsat images for each year and provides interpreters the flexibility to change the TimeSync display image to any one of the Landsat images for that year. TimeSync also provides a graphic time series display of Landsat values for all usable Landsat observations. Interpreters also were expected to examine fine resolution aerial imagery in Google Earth™ and older aerial imagery available through EarthExplorer (USGS, 2018) from the USGS National Aerial Photography Program (NAPP) and the National High-Altitude Photography (NHAP) program. The seven interpreters had a range of educational backgrounds but were all experienced in the visual interpretation of Landsat images and aerial photos. The interpreters all received the same training in the use of TimeSync and the application of the JRD protocols. Senior U.S.

Forest Service and U.S. Geological Survey staff reviewed a set of practice interpretations prior to the analysts beginning interpretation of the 300 sample pixels included in the final dataset.
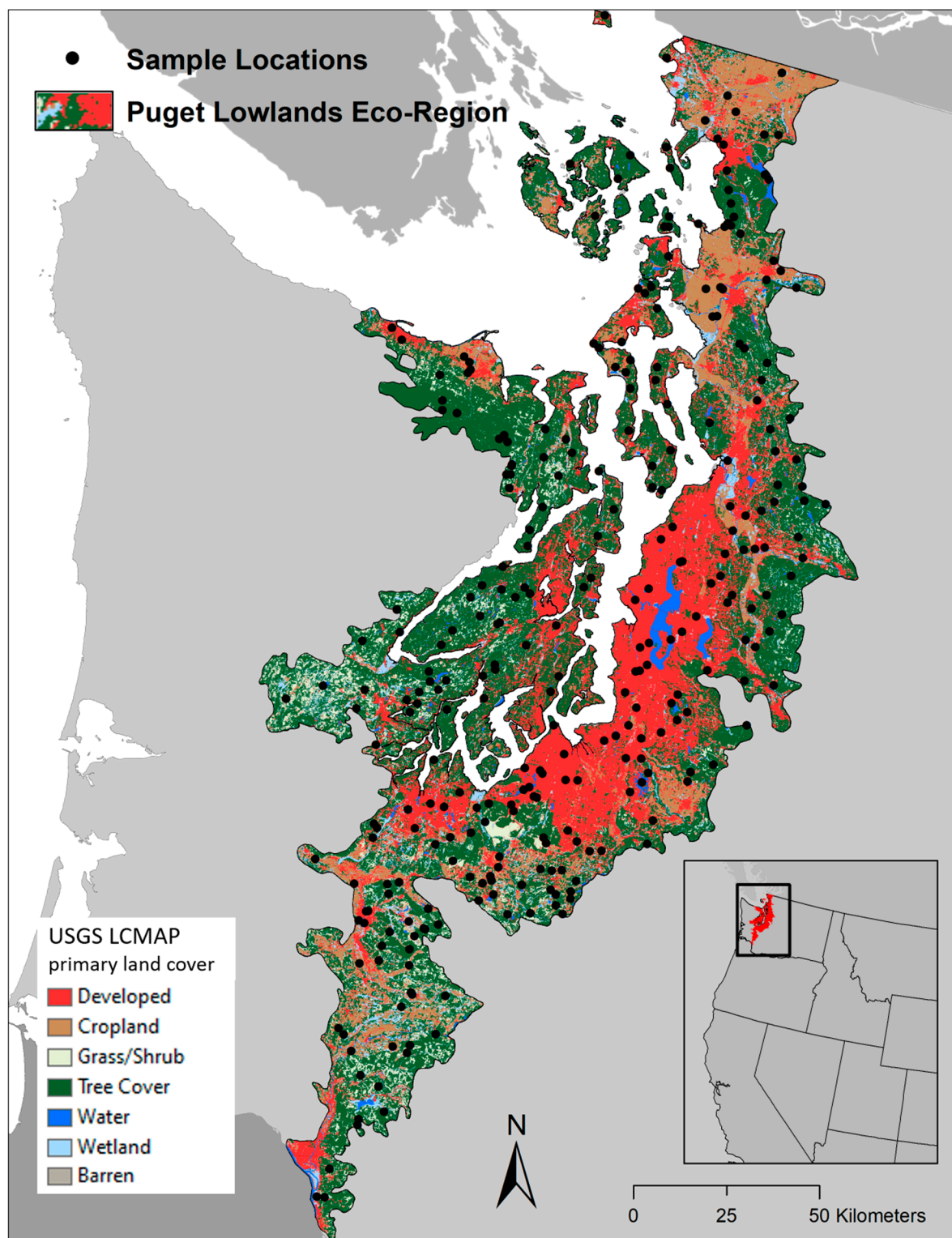

Figure 1. Spatial distribution of the 300 sample pixels in the Puget Sound region of Washington State in the northwest United States.

2.2. Simple Averaging (SA)

Simple averaging estimates the proportion of area from multiple interpreters by averaging the proportion estimates obtained by the individual interpreters. For example, suppose the sampling design is simple random, and the class of interest is wetland. For a specific interpreter, a single reference class label is assigned to each pixel in the sample, and the estimated proportion of the area of wetland would be the proportion of pixels in the sample labeled as wetland. If the estimated proportion of wetland area from interpreter i is denoted as p_i and there are k interpreters, the estimated proportion of the area of wetland combining the information from all k interpreters is simply the average of the individual interpreter estimated proportions,

$$p = \sum_{i=1}^{k} p_i / k \tag{1}$$

An advantage of SA is that it is readily applied to any sampling design. For example, if a stratified sampling design were used to collect the reference data, the proportion of the area of the class of interest would be estimated for each interpreter by

$$p_i = \sum_{h=1}^{H} W_h p_{hi} \tag{2}$$

where p_{hi} = sample proportion of pixels in stratum h that had the reference label of the class of interest, W_h = proportion of the area in the study region in stratum h, and H = number of strata. Once p_i for each interpreter i is estimated, simple averaging would be used to combine the estimates for all interpreters using Equation (1).

2.3. Latent Class Modeling (LCM)

The use of LCM to estimate area requires a set of reference class labels obtained by multiple interpreters, and these labels are the observed or manifest variables of the latent class model. That is, each interpreter assigns a single reference class label to each sample pixel, and the set of labels for the sample provided by each interpreter constitutes a manifest variable. LCM then uses the associations or response patterns among those manifest variables to derive the underlying latent classes and to produce an estimate of the proportion of cases that belong to each latent class, where these latent classes are assumed to represent the true land cover classes.

To estimate the area of each land cover class or land cover change class using LCM, we converted the reference labels of each of the 300 sample pixels to a binary classification consisting of the target class (e.g., developed) and everything else (e.g., not developed). The manifest variables used in the LCM will impact the outcome of the modeling. For example, using a binary classification of land cover as the manifest variables would yield a different outcome than including all land cover classes as the manifest variables. The simplification to a binary classification allows a direct interpretation of the two latent classes generated from the model, as one latent class is the target class, and the other latent class is the complement of the target class. Conversely, the interpretation of the seven latent classes that would be produced when the manifest variables included all seven land cover classes would be more challenging. It is beyond the scope of this work to evaluate fully the differences in area estimates that arise depending on the classes used as the manifest variables. However, we do include a limited exploratory comparison of the area estimates resulting from using just the binary target / non-target classification versus all land cover classes as the manifest variables (Section 3.6).

We also limited the scope of our analysis to the simplest LCM model, which assumes conditional independence. This allows us to compare SA to estimates from the LCM approach that is most readily accessible to practitioners. However, alternative latent class models may be specified when the assumption of conditional independence is un-

tenable [14]. For details of the specific statistical formulation of LCM when conditional independence is assumed, Foody et al. [8] provide a concise summary extracted from the more comprehensive treatments of LCM reported in Vermunt [14] and Vermunt and Magidson [15]. Foody [16] presents data and an accompanying analysis to provide a numerical example of the results and interpretation of LCM.

We implemented LCM using the poLCA library [17] in the R environment (R Core Team 2020, version 4.0.0) [18]. In LCM, parameter estimates for the same sample data can vary because of the random initialization of the Expectation Maximization (EM) algorithm used to produce the estimates. Therefore, for each implementation of LCM to estimate area, we ran 100 iterations of the model in R to counter the risk of identifying a local minimum instead of the global minimum [14]. The Bayesian information criterion (BIC) was used to evaluate the fit of each of the 100 iterations of the LCM model, with smaller BIC preferred. The estimated proportion of the area was obtained from the iteration with the smallest BIC.

Although it is recommended to apply LCM for three or more manifest variables (i.e., three or more interpreters), our initial analyses indicated that the LCM estimates using just two interpreters were generally consistent with SA estimates. In projects with limited resources, it may be that only two interpreters can be afforded, so we included the two-interpreter case in our evaluation. However, an anomaly in the LCM results was introduced when using only two interpreters in that different proportion estimates could occur for iterations that resulted in the same smallest BIC. In such cases, the proportion estimates from the 100 iterations of the LCM algorithm were averaged. When the data for three to seven interpreters were used, the minimum BIC was associated with a unique estimate of the proportion of area.

2.4. Perspectives of Analyst Variability Underlying SA and LCM

SA and LCM represent different approaches to accommodating interpreter variability when estimating area. For SA, inconsistency among interpreters is assumed to be the result of random mis-labeling errors that compensate over the full sample. For example, suppose in the case of three interpreters, if one interpreter mis-labeled a true forest sample pixel as non-forest, there would be a compensating error for which a true non-forest pixel was mis-labeled as forest by that or another interpreter. Consequently, on average, over the full sample, the proportion of the area is estimated without bias because labeling errors are assumed to be random. LCM assumes that the true state of land cover is a latent variable unobservable to the interpreters. Although the interpreters assign what they believe to be the true land cover label, the interpreters are providing the manifest variables from which LCM can extract the latent (true) land cover status. If gold standard reference data existed, then it would be possible to evaluate if the latent classes corresponded to the gold standard classes.

The models underlying SA and LCM would not necessarily result in similar area estimates, and without gold standard reference data, it would not be possible to determine which approach yields the better estimates. However, given the data from the pilot study, we can compare the degree to which the SA and LCM estimates differ from each other and evaluate the sensitivity of each approach to inconsistency among the interpreters who determine the reference classification. In this article, we take the first step of identifying circumstances (e.g., the prevalence of the land cover or change class, interpreter agreement, and interpreter group size) in which SA and LCM differ in terms of the area estimates produced, as well as the variability of the area estimates over different combinations of interpreters within each interpreter group size.

2.5. Analysis

We assessed variability of the area estimates over different numbers of interpreters (one to seven interpreters) and different sets of interpreters (e.g., different combinations of two interpreters selected from the seven interpreters) for each land cover class. The number of possible combinations of interpreters (i.e., interpreter subsets) varies depending

on the group size. We examined all seven possible subsets of six interpreters, all 21 possible subsets of two interpreters, and all 21 possible subsets of five interpreters. For three and four interpreter group sizes, we randomly selected 21 of the 35 possible subsets so that the number of subsets of three and four interpreters matched the number of subsets of two and five interpreters. We used the land cover data from 1990, 2000, and 2010 for our analyses, roughly the mid-range of the 1985−2016 monitoring timeframe of the LCMAP Puget Sound pilot study.

For each interpreter group size, we obtained SA and LCM area estimates for different combinations (subsets) selected from the seven interpreters. We summarized agreement between SA and LCM estimates using the mean difference (MD) between the SA and LCM percent area estimates, the mean absolute difference (MAD) (i.e., the mean of the absolute values of the differences), and r^2 = square of the correlation (*r*) between the SA and LCM estimates. We quantified variability of the area estimates over interpreter subsets within each group size using the range of the estimates (maximum estimated area minus minimum estimated area), the standard deviation of the area estimates, and the coefficient of variation, defined as

$$CV = 100\% * (\text{standard deviation})/\text{mean}. \qquad (3)$$

CV provides a more interpretable metric for comparing variability among the different land cover classes because it scales variability relative to the mean estimated area of each class.

3. Results

We first describe the pairwise agreement among interpreters (Section 3.1) to characterize the degree to which interpreters varied in their reference class labeling. We then evaluate the variability of the area estimates obtained by the individual interpreters (Section 3.2). The initial comparison of the SA and LCM area estimates focuses on the case in which all seven interpreters were used (Section 3.3), followed by an assessment of agreement between SA and LCM estimates over subsets of different interpreters for each group size from two through six interpreters (Section 3.4). The comparison of the variability of the LCM and SA area estimates over interpreter subsets is reported in Section 3.5. Lastly, we explore differences in LCM area estimates when including all land cover classes as the manifest variables compared to LCM area estimates when two classes (target and non-target) are used as the manifest variables (Section 3.6).

3.1. Interpreter Agreement

The pairwise agreement was consistent over time, thus we report agreement results for 2010 land cover and 2000 to 2010 change as representative of the agreement for the 1985−2016 time series. For the 2010 data and all land cover classes, the pairwise agreement between interpreters ranged from 85% to 91% (Table 1). All seven interpreters unanimously agreed on 73% of the 300 sample pixels. For 97% of the sample, four or more interpreters provided the same reference label (Table 2). In terms of the class-specific agreement among interpreters, of the 173 pixels for which at least one interpreter labeled the pixel as forest, 124 (72%) were labeled as forest by all seven interpreters, and 147 (85%) were labeled as forest by a majority (>50%) of interpreters (Table 2). Of the 120 pixels labeled as developed by at least one interpreter, 71 (59%) were labeled as developed by all seven interpreters, and 92 (77%) were labeled as developed by a majority of interpreters. Of the 18 pixels labeled as wetland by at least one interpreter, 1 (6%) was labeled as wetland by all seven interpreters, and 11 (61%) were labeled as wetland by a majority of interpreters.

Lower agreement between interpreters was observed for land cover change (Table 2). Forest gain had the lowest agreement, with only one pixel labeled as forest gain by a majority of interpreters. Forest loss had only two sample pixels (10%), for which all interpreters agreed, and a majority of the seven interpreters assigned forest loss to only 40% of all pixels for which at least one interpreter labeled the pixel as forest loss. Interpreters more consistently distinguished developed gain and forest loss compared to forest gain.

Still, only 40% of the forest loss pixels and 24% of the developed gain pixels were labeled by a majority of the interpreters.

Table 1. The pairwise agreement (expressed as a percent) between the seven interpreters (A to G) based on all land cover classes for the 2010 data.

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	—	87	86	87	88	85	85
B		—	86	89	86	85	86
C			—	90	89	89	90
D				—	86	87	91
E					—	86	87
F						—	87
G							—

Table 2. Interpreter agreement expressed as a cumulative percent of cases for which k or more interpreters agreed on the class label (e.g., for the forest class and $k \geq 5$, of the 173 (Denom) sample pixels labeled as forest by at least one interpreter, 80% had five or more interpreters agreeing on the forest label).

Agreement (%)	**k = Number of Interpreters in Agreement**						
	≥ 7	**≥ 6**	**≥ 5**	**≥ 4**	**≥ 3**	**≥ 2**	**Denom**
All classes	73	83	91	97	100	100	300
Forest	72	76	80	85	88	92	173
Developed	59	68	73	77	83	91	120
Wetland	6	17	39	61	78	83	18
Forest loss	10	25	30	40	55	70	20
Forest gain	0	0	4	4	17	44	23
Developed gain	18	18	24	24	35	59	100

3.2. Area Estimates from the Individual Interpreters

The sample data from each individual interpreter can be used to produce area estimates for the six classes evaluated. The percent area estimates from an individual interpreter represent a situation in which the study uses only a single interpreter. From the sample data for the pilot study, we can evaluate how much area estimates vary over the set of individual interpreters. These percent area estimates varied considerably (Table 3). The range in the individual interpreter estimated percent area over the set of all seven interpreters was smallest for developed gain (1.7%) and largest for forest (10.0%). In the case of forest and the year 1990, depending on which interpreter was selected for the study, the percent area of the forest could be as low as 45.3% or as high as 55.3%. The coefficient of variation (CV, see Equation (3)) of the individual interpreter area estimates was smaller for forest and developed (the two common classes) compared to the CV of the wetland and three change classes (the rare classes). CV exceeded 25% for all three change classes and reached a maximum of 97% for forest gain in 2000−2010. Forest gain was the class with the least consistent agreement among interpreters (Table 2), so this high CV for forest gain can be attributed to the lack of agreement among interpreters for this class.

For forest, developed and wetland, the individual interpreter estimates were generally consistent over time (Figure 2). For example, interpreter B provided the highest area estimate for the forest class in both 1990 and 2000, and the second-highest estimate in 2010, whereas interpreter F tended to yield lower estimates than other interpreters in all three years. A similar pattern of relative consistency occurred for the developed class as interpreters C and E produced the smallest estimates, and interpreters A and F produced the highest estimates for all three years. For wetland, interpreters D and G provided the highest area estimates, and interpreter A obtained the lowest estimates for all three years. The estimated percent areas of change in land cover did not show as consistent of a pattern

for the two change periods assessed (Figure 3), perhaps attributable to higher interpreter disagreement in assigning change class labels.

Table 3. Mean and variability of the land cover and land cover change percent area estimates produced from single interpreters.

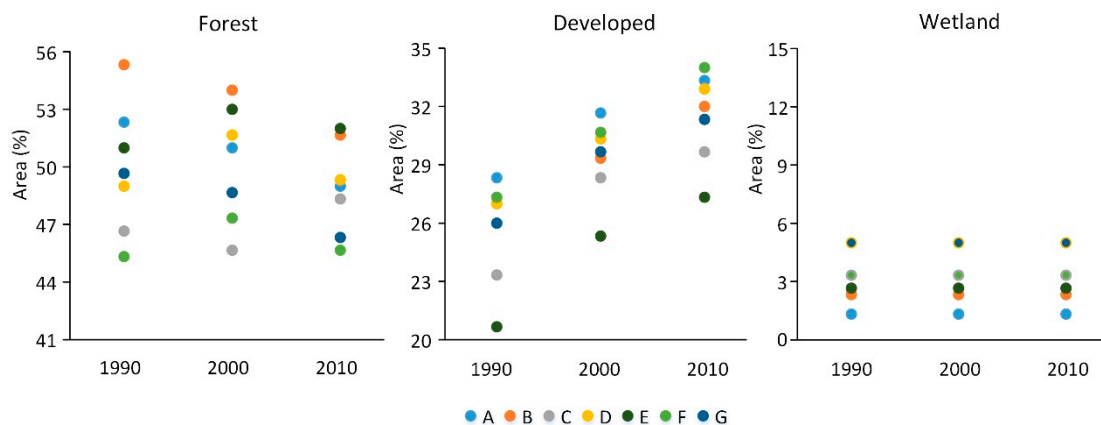| Parameter	Year	Mean	StDev	CV (%)	Min	Max	Range
Forest	1990	49.9	3.39	6.8	45.3	55.3	10.0
	2000	50.2	3.06	6.1	45.7	54.0	8.3
	2010	48.9	2.41	4.9	45.7	52.0	6.3
Developed	1990	25.7	2.71	10.5	20.7	28.3	7.6
	2000	29.3	2.06	7.0	25.3	31.7	6.4
	2010	31.5	2.28	7.2	27.3	33.7	6.4
Wetland	1990	3.3	1.35	41.2	1.3	5.0	3.7
	2000	3.3	1.35	41.2	1.3	5.0	3.7
	2010	3.3	1.35	41.2	1.3	5.0	3.7
Forest loss	1990−2000	4.0	1.11	27.8	3.0	5.7	2.7
	2000−2010	3.1	0.94	30.3	2.0	5.0	3.0
Forest gain	1990−2000	4.3	2.29	53.3	1.7	7.0	5.3
	2000−2010	1.9	1.85	97.4	0.0	5.3	5.3
Developed gain	1990−2000	3.9	0.86	22.1	2.3	5.0	2.7
	2000−2010	2.2	0.66	30.0	1.3	3.0	1.7

Figure 2. Estimated percent area of land cover from the individual interpreters (A to G).
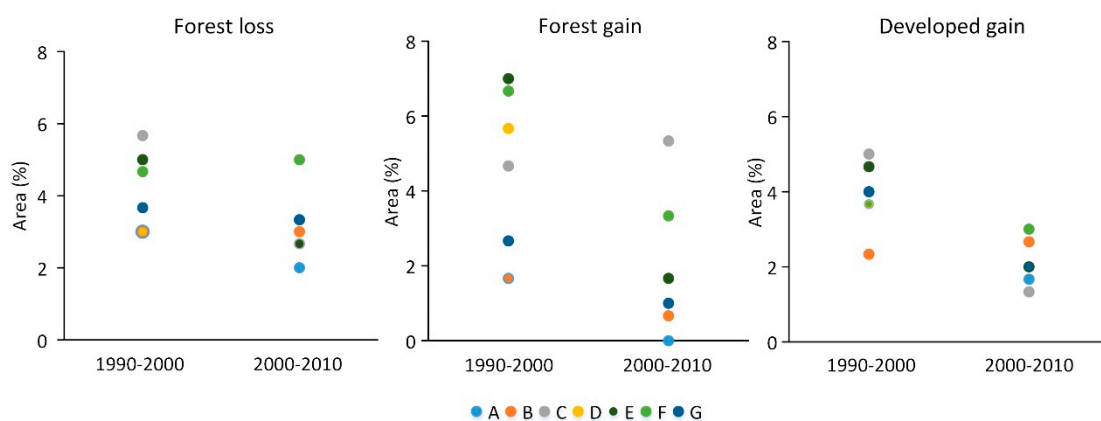

Figure 3. Estimated percent area of land cover change classes for each interpreter (A to G).

3.3. Comparison of SA and LCM Area Estimates Using Data from all Interpreters

When combining the data from all seven interpreters to produce the area estimates, the SA and LCM estimates differed by less than 1.5% for the forest and developed classes (Table 4). For the wetland and all three change classes, the differences between the SA and LCM estimates ranged from 0.6% to 2.3%, with LCM producing larger area estimates in most cases (Table 4). Given that the estimated areas of wetland and change were less than 5% of the total area of the pilot study region, these differences between SA and LCM are large relative to the extent of the areas of wetland and change. For example, LCM estimated 50% greater area of forest loss for 1990–2000 relative to SA even though the difference in the estimates was only 2% (6% for LCM compared to 4% for SA). Forest gain for 2000–2010 had the greatest difference as LCM estimated over twice as much area of forest gain as estimated by SA.

Table 4. Estimated percent area of land cover and land cover change obtained using SA compared to LCM for reference data from all seven interpreters.

Class	Year	SA	LCM	Difference
	1990	49.9	49.0	−0.9
Forest	2000	50.2	51.2	1.0
	2010	48.9	49.1	0.2
	1990	25.7	27.2	1.5
Developed	2000	29.3	28.9	−0.4
	2010	31.5	31.2	−0.3
	1990	3.3	5.0	1.7
Wetland	2000	3.3	5.0	1.7
	2010	3.3	5.0	1.7
Forest loss	1990−2000	4.0	6.0	2.0
	2000−2010	3.1	4.2	1.1
Forest gain	1990−2000	4.3	5.1	0.8
	2000−2010	1.9	4.2	2.3
Developed gain	1990−2000	3.9	3.3	−0.6
	2000−2010	2.2	3.4	1.2

3.4. Agreement between SA and LCM Area Estimates over Interpreter Subsets

Comparisons of SA and LCM area estimates for different subsets of interpreters were based on the land cover data from 2000 and 2010. For two to six interpreters, the MD and MAD values, which reflect the agreement between the SA and LCM estimates, were less than or equal to 2% for all cases except for the group size of two interpreters for developed and the group sizes of two and three interpreters for forest gain (Table 5). Although a difference in percent area of 2% is relatively small for the common classes forest and developed, a difference of 2% is relatively large when estimating the area of a rare class, such as wetland or land cover change. LCM had higher area estimates than SA (i.e., MD > 0) for all interpreter group sizes for the rare classes wetland, forest loss, and forest gain, but MD was both negative and positive for the forest, developed, and developed gain classes. There was not a consistent relationship between r^2 and the number of interpreters for the different land cover area estimates. For forest, developed and wetland, r^2 generally decreased as the interpreter group size increased, but for the three change types, no association between variation in r^2 and interpreter group size was evident.

Table 5. Mean estimated percent area of SA and LCM and agreement between SA and LCM over different interpreter subsets within each interpreter group size for 2010 land cover and 2000–2010 land cover change (MD = mean difference and MAD = mean absolute difference).

Interpreter Group Size	Mean		Agreement		
Forest	SA	LCM	MD	MAD	r^2
2	48.90	49.02	0.11	0.20	0.98
3	48.81	48.60	−0.21	0.51	0.77
4	48.81	47.87	−0.94	1.43	0.25
5	48.90	48.74	−0.17	0.66	0.35
6	48.90	48.49	−0.41	1.14	0.06
Developed	SA	LCM	MD	MAD	r^2
2	31.48	33.62	2.15	2.15	0.92
3	31.60	31.30	–0.30	0.70	0.58
4	31.45	32.59	1.14	1.32	0.35
5	31.48	31.33	−0.14	0.63	0.21
6	31.48	32.05	0.57	0.75	0.34
Wetland	SA	LCM	MD	MAD	r^2
2	3.29	3.30	0.01	0.63	0.52
3	3.24	3.81	0.57	0.60	0.65
4	3.33	4.57	1.24	1.24	0.15
5	3.29	4.74	1.46	1.46	0.19
6	3.29	4.92	1.63	1.63	0.03
Forest Loss	SA	LCM	MD	MAD	r^2
2	3.14	4.71	1.56	1.59	0.41
3	3.19	3.27	0.08	0.43	0.31
4	3.17	3.67	0.50	0.53	0.50
5	3.14	3.96	0.82	0.82	0.41
6	3.14	4.12	0.98	0.98	0.24
Forest Gain	SA	LCM	MD	MAD	r^2
2	1.86	4.52	2.66	0.81	0.49
3	1.85	3.52	1.67	2.14	0.02
4	1.90	2.45	0.55	1.09	0.09
5	1.87	2.89	1.02	1.09	0.51
6	1.86	3.65	1.79	1.79	0.71
Developed Gain	SA	LCM	MD	MAD	r^2
2	2.24	3.02	0.78	0.87	0.21
3	2.26	1.75	−0.51	0.57	0.58
4	2.24	2.08	−0.16	0.43	0.54
5	2.24	2.52	0.28	0.61	0.41
6	2.24	2.81	0.57	0.84	0.50

3.5. Variability of SA and LCM Area Estimates over Interpreter Subsets

SA had smaller variability than LCM over different interpreter subsets within each interpreter group size, indicating that SA would be less sensitive to the interpreters selected to determine the reference land cover data. That is, for each interpreter group size, the range and standard deviation of the SA area estimates were smaller than the range and standard deviation of the LCM area estimates (Table 6). The decrease in the variability of SA relative to LCM also generally was magnified as the number of interpreters increased. For example, for the area estimates of developed, the ratio of the standard deviation of SA to the standard deviation of LCM for the group size of two interpreters was 1.40/1.47 = 0.95, whereas for the group size of six interpreters, the ratio was 0.38/0.78 = 0.49.

Table 6. Comparison of the variability of SA and LCM percent area estimates by interpreter group size (Min = minimum, Max = maximum, StDev = standard deviation, and CV % = coefficient of variation).

Group Size	Simple Averaging (SA)					Latent Class Modeling (LCM)				
Forest	Min	Max	Range	StDev	CV (%)	Min	Max	Range	StDev	CV (%)
1	45.7	52.0	6.3	2.41	4.9	—	—	—	—	—
2	46.0	51.8	5.8	1.48	3.0	46.1	51.9	5.8	1.37	2.8
3	46.8	50.9	4.1	0.99	2.0	46.6	51.0	4.4	1.15	2.4
4	47.3	50.3	3.0	0.82	1.7	45.8	51.6	5.8	1.53	3.2
5	47.7	50.1	2.3	0.59	1.2	46.9	50.6	3.7	1.02	2.1
6	48.4	49.4	1.1	0.40	0.8	46.9	50.5	3.6	1.31	2.7
Developed	Min	Max	Range	StDev	CV (%)	Min	Max	Range	StDev	CV (%)
1	27.3	33.7	6.4	2.28	7.2	—	—	—	—	—
2	28.5	33.5	5.0	1.40	4.4	30.6	36.1	5.6	1.47	4.4
3	30.0	33.3	3.3	0.96	3.0	29.7	33.8	4.1	1.20	3.8
4	30.3	33.0	2.7	0.77	2.4	30.6	34.7	4.1	1.24	3.8
5	30.7	32.7	2.0	0.56	1.8	29.7	33.0	3.3	0.87	2.8
6	31.1	32.2	1.1	0.38	1.2	31.0	33.0	2.0	0.78	2.4
Wetland	Min	Max	Range	StDev	CV (%)	Min	Max	Range	StDev	CV (%)
1	1.3	5.0	3.7	1.35	40.9	—	—	—	—	—
2	1.8	5.0	3.2	0.83	25.2	1.1	5.4	4.3	1.17	35.5
3	2.1	4.4	2.3	0.62	19.1	2.0	5.0	3.0	0.93	24.4
4	2.4	4.2	1.8	0.47	14.1	3.3	5.0	1.7	0.51	11.2
5	2.6	3.9	1.3	0.33	10.0	3.3	5.0	1.7	0.41	8.6
6	3.0	3.6	0.6	0.23	7.0	4.7	5.0	0.3	0.14	2.8
Forest Loss	Min	Max	Range	StDev	CV (%)	Min	Max	Range	StDev	CV (%)
1	2.0	5.0	3.0	0.94	30.3	—	—	—	—	—
2	2.3	4.2	1.8	0.58	18.5	2.9	7.7	4.8	1.39	29.5
3	2.6	3.9	1.3	0.43	13.5	2.1	5.2	3.2	0.76	23.2
4	2.6	3.7	1.1	0.34	10.7	2.8	5.0	2.2	0.61	16.6
5	2.7	3.5	0.7	0.23	7.3	3.0	5.1	2.1	0.59	14.9
6	2.8	3.3	0.5	0.16	5.1	3.3	4.7	1.4	0.59	14.3
Forest Gain	Min	Max	Range	StDev	CV (%)	Min	Max	Range	StDev	CV (%)
1	0.0	5.3	5.3	1.85	97.4	—	—	—	—	—
2	0.3	4.3	4.0	1.14	61.3	1.0	14.4	13.4	4.14	91.6
3	0.7	3.4	2.8	0.85	45.9	0.7	9.4	8.7	2.42	68.8
4	0.8	2.8	2.0	0.59	31.1	0.6	5.3	4.7	1.53	62.4
5	0.9	2.8	1.9	0.48	25.7	1.0	5.3	4.3	1.32	45.7
6	1.3	2.2	0.9	0.31	16.7	1.8	5.0	3.2	1.13	31.0
Devel Gain	Min	Max	Range	StDev	CV (%)	Min	Max	Range	StDev	CV (%)
1	1.3	3.0	1.7	0.66	30.0	—	—	—	—	—
2	1.5	3.0	1.5	0.40	17.9	1.7	4.6	2.9	0.95	31.5
3	1.9	2.9	1.0	0.28	12.4	1.0	3.1	2.1	0.54	30.9
4	1.8	2.7	0.9	0.26	11.6	1.3	3.3	2.0	0.61	29.3
5	1.9	2.5	0.6	0.16	7.1	1.3	3.5	2.1	0.69	27.4
6	2.1	2.4	0.3	0.11	4.9	1.7	3.4	1.8	0.80	28.5

　　　　Variability of the SA and LCM area estimates was associated with interpreter agreement. That is, developed and forest had the smallest CVs among interpreter subsets (Table 6), and the developed and forest classes generally had a good agreement among interpreters (Table 2). The larger CV values occurred for the wetland and three change classes that had a lower agreement among interpreters than developed and forest. Based on CV, the variability of the area estimates within each interpreter group size was much larger for the rare classes (wetland, forest loss, forest gain, and developed gain) relative to the common classes. Thus, the impact of reference data variability would be much stronger on rare classes.

3.6. Comparison of LCM Estimates Using All Classes versus Two Classes as Manifest Variables

The LCM area estimates discussed in Sections 3.3–3.5 were based on LCM applied to a binary classification of each target class (i.e., a binary classification as the manifest variables). Here we compare the LCM area estimates when all seven land cover classes are retained for the manifest variables (referred to as LCM7) to the area estimates based on a binary classification of land cover used for the manifest variables (referred to as LCM2). We report results for subsets of interpreters within group sizes of three and six interpreters.

The LCM7 area estimates (averaged over all interpreter subsets within three and six interpreter group sizes) were less than or equal to the LCM2 estimates for all three land cover classes for both group sizes (Table 7). The differences between the LCM7 and LCM2 estimates were larger for the six interpreter group size relative to three interpreters. The LCM2 percent area estimates exceeded the LCM7 estimates for forest and developed by 3% for the group size of six interpreters. The LCM7 estimates had a wider range and larger standard deviation compared to the LCM7 estimates. The variability of LCM7 estimates relative to that of the LCM2 estimates increased for the six interpreter group size relative to the three interpreter group size. That is, the ratio of the standard deviation of the LCM7 area estimates divided by the standard deviation of the LCM2 area estimates was larger when there were six interpreters. Given that one of the advantages of SA relative to LCM was less sensitivity to the set of interpreters selected (Section 3.5), this advantage of SA would be more pronounced relative to LCM7 than it was relative to LCM2.

Table 7. Comparison of LCM area estimates produced from models with all seven land cover classes included in the manifest variables (LCM7) and estimates from models with a binary classification (LCM2) for three and six interpreter group sizes (2010 land cover data).

Interpreter Group Size	Mean Area (%)		Range		Standard Dev	
Forest	LCM2	LCM7	LCM2	LCM7	LCM2	LCM7
Three	48.6	48.0	4.4	4.8	1.15	1.19
Six	48.5	45.2	3.6	5.4	1.31	1.60
Developed	LCM2	LCM7	LCM2	LCM7	LCM2	LCM7
Three	31.3	30.8	4.1	4.4	1.20	1.35
Six	32.0	28.5	2.0	3.8	0.78	1.45
Wetland	LCM2	LCM7	LCM2	LCM7	LCM2	LCM7
Three	3.8	3.8	3.0	3.4	0.93	1.06
Six	4.9	4.8	0.3	0.7	0.14	0.25

4. Discussion

In this study, we compared area estimates produced from LCM to estimates computed by SA for three land cover classes and three land cover change types. In addition to examining agreement between the LCM and SA area estimates, our results extend previous work (e.g., Foody and Boyd [8]) by addressing the question of the variability of the area estimates over different combinations of interpreters within each interpreter group size of two to seven interpreters. Interpreter variability can greatly impact area estimates produced from reference data even when the interpreters are experienced and undergo common training. Although the interpreters in the pilot study received common training prior to initiating data collection, they did not receive further feedback regarding their interpretations once data collection was in progress. Conversely, in the operational implementation of LCMAP reference data collection, interpreters received ongoing feedback and participated in protocols designed to improve interpreter consistency over the time span of data collection [7]. Therefore, interpreter consistency may be greater for the LCMAP sample, and consequently, variability of the area estimates less, relative to what we observed for the Puget Sound pilot study data used in our analyses of SA and LCM.

Area estimates obtained by SA had less sensitivity to the subset of interpreters selected for each group size than area estimates obtained by LCM. For example, for the seven different subsets of interpreters for the interpreter group size of six, the variability of

the LCM area estimates was greater than the variability of the SA estimates. In practice, a method of area estimation that is robust to the set of interpreters selected is preferable, so SA was superior to LCM in that regard. Variability of the estimates over different interpreter subsets decreased as the number of interpreters increased (Tables 5 and 6). This result was expected for the SA estimates because the variance of an estimated mean decreases as the sample size (i.e., number of interpreters) increases. The reduction in variability for a greater number of interpreters was more pronounced for the SA area estimates than for the LCM area estimates.

In addition to the advantage of simplicity, SA also has the desirable feature that the estimated areas are additive (i.e., the sum of the estimated percent areas of all classes will be 100%). Additivity would not be assured for the LCM estimates when the manifest variables were defined using the binary classification of land cover (i.e., target class, not the target class), but would be achieved if the manifest variables included all land cover classes (Table 8). The lack of additivity of the binary class (LCM2) manifest variables is demonstrated by two example subsets of four interpreters (Table 8) in which the total percent area summed over all classes exceeds 100% (in the case of interpreters A, B, C, and E the sum is 105%). By construction, the area estimates using all classes as the manifest variables (LCM7) and SA will sum to 100%.

Table 8. Evaluation of additivity of percent area estimates (2010 land cover and two example cases of four interpreters) for LCM using two (LCM2) or seven (LCM7) classes as the manifest variables and SA (additivity defined as the percent area of all land cover classes summing to 100%).

Class	Interpreters C, D, E, G			Interpreters A, B, C, E		
	LCM7	LCM2	SA	LCM7	LCM2	SA
Forest	47.00	47.01	49.00	47.13	50.35	50.25
Developed	29.01	31.48	30.33	28.44	31.09	30.58
Agriculture	8.34	8.34	8.75	10.71	10.80	9.33
Grass/Shrub	6.00	5.67	5.42	4.95	5.63	4.58
Wetland	5.00	5.00	4.00	3.57	4.28	2.58
Water	2.33	2.33	2.33	2.67	2.67	2.42
Bare ground	2.32	0.33	0.17	2.53	0.33	0.25
Total	100.00	100.16	100.00	100.00	105.15	100.00

4.1. Alternatives to Simple Averaging

Several alternatives to SA could be considered when multiple analysts have interpreted the same sample units. For example, if outliers are a concern (e.g., if one interpreter deviated considerably from others), the median of the individual interpreter proportion estimates could be used to reduce the impact of outliers. The median proportion does not have the desirable additivity property possessed by SA. Outliers should be unusual in applications where trained and coordinated interpreters obtain the reference data, as was the case for our pilot study data (Table 2).

Using the majority reference class label is another alternative to SA for producing area estimates when multiple analysts have interpreted each sample pixel. For our data and the set of all seven interpreters, 97% of the sample pixels had a majority reference class label, so a protocol for determining the reference class label would be needed for the remaining pixels that did not have a majority label. Yet another option would be to re-evaluate all sample pixels that did not have agreement among all interpreters and resolve the reference class labels by consensus or by adjudication by an expert interpreter.

4.2. Additional Considerations for LCM

In our study, the sample pixels for estimating the area were acquired via simple random sampling. However, stratified and cluster sampling are also often implemented [19], so it would be important to be able to produce area estimates for these sampling designs as well. For LCM, model fitting and parameter estimation would need to accommodate strata

or clusters if present in the sampling design. LCM can be implemented for these more complex sampling designs [20,21], but additional research is needed to explore properties of the area estimates generated from LCM for these sampling designs.

We restricted our analyses to the most basic latent class model, targeting an analysis approach that could be readily implemented by practitioners. Specifically, we implemented a model assuming conditional independence applied to a binary land cover classification. Practitioners more strongly versed in the application of LCM could explore more complex models (e.g., models that do not assume conditional independence). Our brief exploration of LCM area estimates based on different manifest variables (Section 3.6) demonstrated that the area estimates and the variability of these estimates over interpreter subsets depended on how the manifest variables were defined (Section 3.6). Additional research is needed to better understand the interpretation of the LCM outputs under different manifest variables.

LCM can be applied to map accuracy studies to provide a model-based estimate of producer's accuracy. This feature is particularly relevant for analyzing volunteer geographic information (VGI) used as reference data in land cover studies. The application of LCM to VGI data introduces considerations different from those present in our study. For the LCMAP pilot study data, we had a fixed set of similarly trained, experienced interpreters who all examined the complete sample of 300 pixels. Typically, VGI studies have a much larger set of interpreters (for example, 65 volunteers participated in the study reported by Foody et al. [22]), volunteers may not interpret the same set of sample locations, and the volunteers may be expected to have greater variation in skill relative to the seven interpreters in the LCMAP study. In general, the results we observed for LCM would not necessarily translate to applications using VGI unless the VGI data collection protocol was similar to that of our study (i.e., two to seven interpreters who each obtained the reference data for every sample unit).

5. Conclusions

The finding that area estimates of land cover and land cover change can be highly variable over different sets of interpreters re-affirms the importance of implementing protocols to improve and maintain interpreter consistency when multiple interpreters are used to collect reference data [2,7]. Increasing the number of interpreters per sample pixel will reduce the impact of interpreter variability on area estimates, but increasing the number of interpreters comes at the cost of reduced sample size (assuming a fixed total cost for collecting reference data). Area estimates produced from SA and LCM were different, but in the absence of gold standard reference data, it was not possible to conclude that one approach was better in terms of accurately reflecting the true area. The fact that the SA and LCM area estimates differed suggests the need to develop diagnostic tools to assess the assumptions underlying SA and LCM (Section 2.5). SA generally had smaller variability than LCM over different combinations of interpreters for each interpreter group size. Smaller variability over interpreter subsets is preferable because it conveys less sensitivity to the selection of the interpreters providing the reference data.

trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A.

Appendix A.1. Definition of Land Cover Classes Used in the Analysis

Condensed versions of the class definitions are presented in this Appendix, with more detailed definitions provided in Pengra et al. [7]. The developed class was defined as land covered by man-made structures (e.g., high-density residential, commercial, industrial, mining or transportation), or a mixture of both vegetation (including trees) and structures (e.g., low-density residential, lawns, recreational facilities, cemeteries, transportation, and utility corridors, etc.), including any land functionally altered by human activity. Forest was defined as land that is planted or naturally vegetated and which contains (or is likely to contain) 10% or greater tree cover at some time during a near-term successional sequence. Forest may also include deciduous, evergreen, and/or mixed categories of natural forest, forest plantations, and woody wetlands. Wetland was defined as land adjacent to or within a visible water table (either permanently or seasonally saturated) dominated by shrubs or persistent emergents, and in the case of forest use, where constant or recurrent shallow inundation or saturation of water is a determining factor in shaping the physical characteristics of the underlying vegetation and soils.

References

1. Pielke, R.A. Land Use and Climate Change. *Science* **2005**, *310*, 1625. [CrossRef] [PubMed]
2. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
3. Olofsson, P.; Foody, G.M.; Stehman, S.V.; Woodcock, C.E. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sens. Environ.* **2013**, *129*, 122–131. [CrossRef]
4. Stehman, S.V. Model-assisted estimation as a unifying framework for estimating the area of land cover and land-cover change from remote sensing. *Remote Sens. Environ.* **2009**, *113*, 2455–2462. [CrossRef]
5. Gallego, F.J. Remote sensing and land cover area estimation. *Int. J. Remote Sens.* **2004**, *25*, 3019–3047. [CrossRef]
6. Powell, R.L.; Matzke, N.; de Souza, C.; Clark, M.; Numata, I.; Hess, L.L.; Roberts, D.A. Sources of error in accuracy assessment of thematic land-cover maps in the Brazilian Amazon. *Remote Sens. Environ.* **2004**, *90*, 221–234. [CrossRef]
7. Pengra, B.W.; Stehman, S.V.; Horton, J.A.; Dockter, D.J.; Schroeder, T.A.; Yang, Z.; Cohen, W.B.; Healey, S.P.; Loveland, T.R. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. *Remote Sens. Environ.* **2020**, *238*, 111261. [CrossRef]
8. Foody, G.M.; Boyd, D.S. Using volunteered data in land cover map validation: Mapping West African forests. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1305–1312. [CrossRef]
9. Foody, G.M. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sens. Environ.* **2010**, *114*, 2271–2285. [CrossRef]
10. Goodman, L.A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **1974**, *61*, 215–231. [CrossRef]
11. Bicheron, P.; Defourny, P.; Brockmann, C.; Schouten, L.; Vancutsem, C.; Huc, M.; Bontemps, S.; Leroy, M.; Achard, F.; Herold, M.; et al. *GlobCover-Products Description and Validation Report*; MEDIAS-France: Levallois-Perret, France, 2008.
12. Brown, J.F.; Tollerud, H.J.; Barber, C.P.; Zhou, Q.; Dwyer, J.L.; Vogelmann, J.E.; Loveland, T.R.; Woodcock, C.E.; Stehman, S.V.; Zhu, Z.; et al. Lessons learned implementing an operational continuous United States national land change monitoring capability: The Land Change Monitoring, Assessment, and Projection (LCMAP) approach. *Remote Sens. Environ.* **2020**, *238*, 111356. [CrossRef]
13. Cohen, W.B.; Yang, Z.; Kennedy, R. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync—Tools for calibration and validation. *Remote Sens. Environ.* **2010**, *114*, 2911–2924. [CrossRef]
14. Vermunt, J.K. *LEM: A General Program for the Analysis of Categorical Data*; Department of Methodology, Tilburg University: Tilburg, The Netherlands, 1997.
15. Vermunt, J.K.; Magidson, J. Latent class models for classification. *Comput. Stat. Data Anal.* **2003**, *41*, 531–537. [CrossRef]
16. Foody, G.M. Latent class modeling for site- and non-site-specific classification accuracy assessment without ground data. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 2827–2838. [CrossRef]
17. Linzer, D.A.; Lewis, J.B. poLCA: An R package for polytomous variable latent class analysis. *J. Stat. Softw.* **2011**, *42*, 1–29. [CrossRef]
18. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.

19. Stehman, S.V. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **2009**, *30*, 5243–5272. [CrossRef]
20. Vermunt, J.K.; Magidson, J. Latent class analysis with sampling weights: A maximum-likelihood approach. *Sociol. Methods Res.* **2007**, *36*, 87–111. [CrossRef]
21. Patterson, B.H.; Dayton, C.M.; Graubard, B.I. Latent class analysis of complex sample survey data. *J. Am. Stat. Assoc.* **2002**, *97*, 721–741. [CrossRef]
22. Foody, G.M.; See, L.; Fritz, S.; Van der Velde, M.; Perger, C.; Schill, C.; Boyd, D.S. Assessing the accuracy of Volunteered Geographic Information arising from multiple contributors to an internet based collaborative project. *Trans. GIS* **2013**, *17*, 847–860. [CrossRef]