Chapter 5

ACCURACY ASSESSMENT OF MAPS OF FOREST CONDITION

Statistical Design and Methodological Considerations

Raymond L. Czaplewski
USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado, 80526-1891, USA

1. INTRODUCTION

No thematic map is perfect. Some pixels or polygons are not accurately classified, no matter how well the map is crafted. Therefore, thematic maps need metadata that sufficiently characterize the nature and degree of these imperfections. To decision-makers, an accuracy assessment helps judge the risks of using imperfect geospatial data. To analysts, an accuracy assessment helps describe the reliability of the map for geospatial analyses and modeling, and the distribution of different types of "true" land cover within each mapped category. To producers of thematic maps, an accuracy assessment measures the degree of technical success for alternative algorithms or techniques. To project managers, an accuracy assessment helps determine contract compliance or measure performance of technical staff.

1.1 Random sampling

There are two general methods to obtain reference data for an accuracy assessment: *ad hoc* sampling and probability sampling. Both methods commonly appear in remote sensing projects. However, for the following reasons, only probability sampling is considered in this Chapter.

Ad hoc methods often rely on a sampling plan that selects convenient sites in order to minimize cost of reference data. Experts purposively select sites believed to be representative of the mapped area. This method can

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

produce an error matrix (e.g., Tables 5-1 to 5-3) at relatively low cost. This matrix accurately describes the results for the sampled sites. However, does that same error matrix provide a useful assessment of classification accuracy for the entire thematic map? The producer of the accuracy assessment would argue that this approach is good enough for practical purposes. In some benign cases, this can be true. But what if a user of the thematic map is skeptical of its accuracy, or what if there are disagreements over performance or contract compliance? Furthermore, convenient sampling sites are often near roads, which are frequently associated with unique landforms, land uses, management histories, and landscape patterns. Are such sites truly representative of the entire map? Are the conditions that cause misclassification errors similar among convenient and inconvenient sampling sites? Other than expert opinion, there is no good way to answer these questions. And what if there is a disagreement among the experts? In pathological situations, some "experts" might intentionally seek atypical sites to discredit the map's accuracy. It is far easier to discredit the accuracy of a map than prove otherwise with *ad hoc* methods.

The cost savings offered by *ad hoc* methods can also be achieved with probability sampling. The random sample can be constrained to a sub-population that is accessed relatively inexpensively (e.g., all portions of the thematic map that are less than 500 m from a road). Valid inference is limited to this sampled population, but at least the inference is scientifically defensible. The remainder of this Chapter considers only probability sampling methods.

1.2 Objectives

The objective of this Chapter to present simple, statistically valid, and cost-effective statistical methods to estimate a contingency table during an accuracy assessment. Each row of the contingency table represents a category from the thematic map, and each column represents a category from the reference data. In the remote sensing literature, the most familiar contingency table (e.g., Table 5-1) is the "error matrix" or "confusion matrix" (Story and Congalton 1986). However, the contingency table could use different classification systems for the reference data and the map. For example, the reference data could include categories that can only be reliably applied by a field crew, and these data used to characterize the map categories in greater thematic detail. Another type of contingency table uses "fuzzy-set" categories (Gopal and Woodcock 1994), which cross-classifies points based on their thematic categories on the map and categories such as "Correct", "Acceptable", "Not right" or "Very wrong" (e.g., Table 5-2). These contingency tables share most of the same statistical issues, and are considered simultaneously in this Chapter. In addition, this Chapter covers only simple random sampling of points on the thematic map, and one

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

specialized type of stratified random sampling. An informed non-statistician can apply these simple and robust designs with little risk of procedural error.

Table 5-1. Error matrix based on a simple random sample of 100 points from a 1,000,000-ha sample population, which are cross-classified by both the Map Classifier and the Reference Classifier.

		Count of sample points				Mapped area (ha)	
		Reference Classifier					
		Forest	Old-growth forest	Non-forest	Total	Estimated (100 sampled points)	Exact (all Map Objects in the GIS)
Map Classifier	Forest	43	1	4	48	480,000	409,346
	Old-growth forest	2	6	0	8	80,000	41,634
	Non-forest	14	3	27	44	440,000	549,020
	Total	59	10	31	100		
Estimated true area (ha)		590,000	100,000	310,000	1,000,000	1,000,000	1,000,000

Estimated overall accuracy: (43+6+27)/100=76%

Kappa: $[100\cdot(43+6+27)-(59\cdot48+10\cdot8+31\cdot44)]/[100^2-(59\cdot48+10\cdot8+31\cdot44)]= 0.58$

	Estimated producer's accuracy	Estimated user's accuracy
Forest	43/59= 73%	43/48= 90%
Old-growth forest	6/10= 60%	6/8= 75%
Non-forest	27/31= 87%	27/44= 61%

1.3 Definitions

. "Map Objects" are either pixels or polygons. Map Objects are modeled as internally homogeneous, even though in reality they often include some internal heterogeneity, at least at a fine spatial scale. Therefore, problems associated with "mixed pixels", or atypical inclusions within a polygon, are not covered in this Chapter. A "Classifier" is defined as a process that assigns a Map Object into one, and only one, thematic category, such as water, forest or barren classes, within a user-defined classification system. The Map Classifier might be a supervised or unsupervised algorithm operating on Landsat data, a photo-interpreter using aerial photography for a "wall-to-wall" stand map, or a geospatial model (e.g., wildlife habitat suitability). The Map Classifier is applied to every Map Object in the map. The Reference Classifier is the process that assigns an infinitesimally small point on the map into its "true" or "correct" category. This category could exactly correspond to the classification system used for the map (e.g., Table 5-1), a user-defined classification system that differs from the system used

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

for the map, or it could be a fuzzy-set category (e.g., Table 5-2). This point on the map is surrounded by a larger pixel or polygon (i.e., a Map Object), which serves as a "support region" for application of a classification protocol. The Reference Classifier might be a ground crew or an interpreter using high-resolution aerial photography. The Reference Classifier is considerably more expensive to apply than the Map Classifier. Therefore, the Reference Classifier can only be applied to a relatively small sample of points (and the Map Objects that form their support regions).

Table 5-2. Simple contingency table in which the Reference Classifier uses fuzzy-set categories to assign different degrees of correctness or error to each sample point.

		Reference Classifier				
		Correct	Acceptable	Not right	Very wrong	Row Total
Map Classifier	Forest	43	2	2	1	48
	Old-growth forest	6	0	1	1	8
	Non-forest	27	10	4	3	44
	Column Total	76	11	8	5	100

Table 5-1 is an example of a simple error matrix, in which a sample of n=100 points is cross-classified by both the Map Classifier and the Reference Classifier. In this example, both the Map and Reference Classifiers classified 43 points as forest, while only one point is incorrectly classified as old-growth forest on the map but classified as forest with the reference data. On the other hand, Table 5-2 is a simple example of a fuzzy-set contingency table, where the Reference Classifier uses linguistic values to describe the relative degree of correctness or error to each point. Assume this sample of 100 points came from a statistically valid, simple random sample of points on the map. Then the counts in Tables 5-1 and 5-2 can be directly used to estimate probabilities that any point in the map would be so classified by the Map and Reference Classifiers.

1.4 Accuracy assessment statistics

The following statistics quantitatively describe classification accuracy and other types of metadata for a thematic map.

- Overall accuracy: What is the probability that any point on the map is assigned to exactly the same category by the Map Classifier and the Reference Classifier? For example, the overall accuracy in Table 5-1 is estimated to be 76 %.
- Fuzzy-set accuracy: What is the probability that any point on the map would be assigned to one or several linguistic categories by the Reference Classifier? For example, the probability that the classification

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

of any point on the map is "Correct" or "Acceptable" is estimated to be (76 %+11 %)=87 % from Table 5-2.

- Marginal proportions: What is proportion of the population is classified as Category *X*? Using Table 5-1 as an example, 10 % of the thematic map (100,000 ha) is estimated to be truly old-growth forest based on the Reference Classifier. Using that same sample of 100 points, the estimated proportion of the map classified as old-growth forest is 8 % (80,000 ha).
- Kappa coefficient of agreement: Kappa is a scalar statistic that quantifies the agreement between the Reference and Map Classifiers in a multivariate error matrix (Campbell 1996). Values of kappa exceeding 0.6 are considered good (Czaplewski 1994). However, the analytical value of the kappa statistic is questionable (Stehman, 2002), and this Chapter treats kappa as merely a conventional descriptive statistic.
- Conditional probabilities given the Reference Classifier: For any point assigned to Category *Y* by the Reference Classifier, what is the probability that this same point would be assigned to Category *X* by the Map Classifier? For example, assume you are standing over a point on the ground that is surrounded by old-growth forest (based on the protocol used by the Reference Classifier for the support region). Table 5-1 estimates that there is a 3/10=30 % chance of that same point being classified as non-forest on the map (i.e., an "omission error"). "Producer's accuracy" is a special case (Story and Congalton 1986). Given that the Reference Classifier assigns the point to category X, what is the probability that the point is correctly assigned to category *X* by the Map Classifier? Producer's accuracy is 60 % in this old-growth forest example.
- Conditional probabilities given the Map Classifier: For any point assigned to Category *X* by the Map Classifier, what is the probability that this same point would be assigned to Category *Y* by the Reference Classifier? For example, pick any point on the map that is classified as old-growth forest, and then find that same point in the field. Table 5-1 estimates that there is a (8-6)/8=25 % chance of that point being mislabeled as some other map class (i.e., "commission error"). User's accuracy is a special case of this statistic (Story and Congalton 1986), which is 75 % in this example.

2. FIVE STEPS IN AN ACCURACY ASSESSMENT

Five steps are necessary to produce a successful, yet simple and cost effective, accuracy assessment.

1. Select a probability sample of points for which expensive reference data will be collected to compare with classifications of corresponding points on the map.

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

2. Define the response design, which is the protocol used to apply the Reference Classifier to the support region for each sample point selected in Step 1. The response design must produce a "true" classification that is acceptable to users of the accuracy assessment.
3. Use correct statistical methods to estimate accuracy assessment statistics with the sample data from Steps 1 and 2.
4. Use diagnostic statistics to detect potential problems in the execution of Steps 1, 2, and 3.
5. Present the results to both the users and producers of the map in an informative and intuitive format.

The value of the accuracy assessment depends on how well each and every step is conducted.

The general subject of sample surveys is well covered in general references, such as Cochran (1977), Särndal et al. (1992), Schreuder et al. (1993), Salant and Dillman (1994), Lloyd (1999), and Lohr (1999). In the context of accuracy assessments in remote sensing studies, Stehman and Czaplewski (1998) and Stehman (2001) discuss the first three steps.

2.1 Selection of the reference sample

The sample design specifies how to select the points at which the reference data are gathered. The "target population" is the area or region represented by the thematic map, while the "sampled population" is the portion of the target population that is chosen for sampling. Ideally, the target and sampled populations are identical, but practical constraints often require selecting a sampled population that is a smaller segment of the target population.

For example, assume the target population is a 4,000,000 ha region covered by the thematic map. To reduce the cost of the accuracy assessment, the probability sample of reference points will be drawn from only those areas within a 500 m "buffer" from any road. A GIS is used to identify all portions of the thematic map within the 500 m buffer. In this example, only 1,000,000 ha of the entire map is within the 500 m buffer. The estimated error matrix will validly estimate classification accuracy for that 1,000,000 ha. Unfortunately, inferences for the remaining 3,000,000 ha of target population cannot be scientifically supported with data from this accuracy assessment.

2.1.1 Simple random sampling

The most straightforward and robust design is a simple random sample of n points on the map. Tables 5-1 and 5-2 are examples. Every sample point has exactly the same probability of being included in the sample, regardless

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

of the map classification. Therefore, a simple random sample can be drawn before the final thematic map is available.

2.1.2 Stratified random sampling

There are often rare map categories that are especially important to the user. Unfortunately, a simple random sample is expected to allocate only a small number of reference samples to any rare category. However, "stratified random sampling" can be used to allocate a different sampling intensity to each map category based on the importance of that category to the user. Usually, stratification improves statistical precision, even when the sample size is in each stratum is proportional to the prevalence of each stratum (i.e., proportional allocation).

This Chapter defines a "stratum" as all portions of the thematic map that are classified into the same map category on the final thematic map. Rather than select a single random sample from the entire sampled population, a simple random sample is independently drawn within each category after the thematic map is completed. Compared to random sampling, stratified random sampling is slightly more complex; however, it remains a feasible choice for non-statisticians. Other approaches to stratification can be valid, but estimation is more complex, thus requiring the aid of a consulting statistician (Czaplewski 2000).

Table 5-3A is an example of a stratified random sample, in which an equal number of sample points is allocated to each stratum, regardless of the stratum's prevalence. However, unlike a simple random sample (e.g., Table 5-1), raw counts in a stratified sample cannot be directly used to make unbiased estimates for statistics that are computed from multiple strata (i.e., multiple rows in the error matrix). With stratified random sampling, each cell in Table 5-3A must be converted into an estimated joint probability in order to consider the full suite of assessment statistics. This conversion has been done to produce Table 5-3B. Table 5-8 gives all estimators needed with stratified sampling.

Stratified random sampling has several disadvantages compared to simple random sampling. In order to stratify on the mapped categories, the final thematic map must be available before reference data are collected. This can be several years after the remotely sensed data were originally acquired. Therefore, some misclassifications will actually be caused by changes in land cover. Furthermore, if changes are made to the thematic map after drawing the stratified random sample, then more complex estimators are needed (e.g., Czaplewski 2000), which are not covered in this Chapter. Alternatively, a simple random sample of reference data can be implemented during the earliest phases of a remote sensing study. The next section recommends use of an expected error matrix during the planning phase to evaluate the expected precision of the accuracy assessment statistics, and

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

this same planning tool can help evaluate the advantages of simple random sampling compared to stratified random sampling.

2.1.3 Sample size

One of the most fundamental questions in any accuracy assessment is "How many points should be sampled?" For a simple random sample, Czaplewski and Catts (1992) found relatively little gain in statistical precision beyond 500 to 1,000 sample points. Congalton (1991) recommends a sample size of 50 for each map category if a stratified random sample is used, and map categories are used to define the strata. Some stratified random samples allocate half of the total sampling intensity to emulate proportional allocation, and the remaining half to improve estimates for rare categories. For example, if $n_{i\bullet}$ is the sample size allocated to mapped stratum i, n is the total sample size, $p_{i\bullet}$ is the proportion of the sampled population mapped as category i, and there are a total of k mapped categories, then $n_{i\bullet}=[p_{i\bullet}(n/2)+(1/k)(n/2)]$.

Figure 5-1 and Tables 5-4 to 5-7 help the practitioner chose sample sizes in more specific applications. Consider a simple random sample of 100 points. If each point

Table 5-3. Error matrix based on a stratified sample of 100 from the population in Table 5-1

		Count of points from stratified sample				Mapped area (ha)	
	A	Reference Classifier					
		Forest	Old-growth forest	Non-forest	Total	"Estimated"	Exact
Map Classifier (strata)	Forest	30	1	3	34		409,346
	Old-growth forest	9	22	2	33		41,634
	Non-forest	10	2	21	33		549,020
	Total	49	25	26	100		1,000,000
	B	Estimates from stratified sample					
Map Classifier	Forest	36.1%	1.2%	3.6%	40.9%	409,346	409,346
	Old-growth forest	1.1%	2.8%	0.3%	4.2%	41,634	41,634
	Non-forest	16.7%	3.3%	35.0%	54.9%	549,020	549,020
	Total	53.9%	7.3%	38.9%	100.0%		
Estimated true area (ha)		538,912	73,070	388,018	1,000,000	1,000,000	1,000,000
Estimated overall accuracy: (36.1+2.8+35.0)/100=73.9%							
		Estimated producer's accuracy			Estimated user's accuracy		

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

	Count of points from stratified sample			
Forest	36.1/53.9=	67.0%	36.1/40.9=	30/34=88.2%
Old-growth forest	2.8/7.3=	38.4%	2.8/4.2=	22/33=66.7%
Non-forest	35.0/38.9=	90.0%	35.0/54.9=	21/33=63.6%

is classified into "Correct" or "Incorrect", then the proportion of correct classifications is equivalent to Overall Accuracy. Assume the expected Overall Accuracy is 75 %; the 90 % confidence interval from Table 5-5 is 67 % to 82 %, meaning there is a 5 % chance that the true Overall Accuracy is less than 67 %, and another 5 % chance that the true Overall Accuracy is greater than 82 %. If the simple random sample were increased to 250 points, then the confidence interval would shrink to approximately 70 % to 80 % (Table 5-5). Likewise, assume old-growth forest is expected to cover 10 % of the sampled population. The 90 % confidence interval from a simple random sample of 100 points would be 6 % to 16 % (Table 5-5), while the confidence interval from 250 points would be 7 % to 14 % (Table 5-5). Only the user and producer of an accuracy assessment can judge whether the increase in statistical precision is worth the extra cost for reference data.

The prudent practitioner will use the literature and expert judgment to construct an expected error matrix during the planning phase of the accuracy assessment. Then, the expected confidence intervals for alternate sample sizes and sampling designs can be computed with methods that follow in this Chapter. This hypothetical exercise will help anticipate the precision of the accuracy assessment statistics for various funding levels and sample designs.

2.1.4 Independence of reference sample

The sample of reference sites should not include any sites used for training or labeling the Map Classifier. In some sense, the Map Classifier is optimized to agree with a sample of training or labeling sites; therefore, the classification accuracy for that particular sample of sites will likely be greater than the accuracy expected at other sites. If training sites are used for accuracy assessment, then the metadata should indicate that the estimated accuracy is likely an overestimate of the true accuracy. If the training or labeling sites will be a subset of the probability sample of reference sites, then a simple random sub-sampling procedure should be used to separate the sample into two independent groups: reference sites and training sites. This assures that the reference sites remain a reliable probability sample. More complex sub-sampling procedures could produce reliable results, but a consulting statistician should become involved to assure that correct statistical estimators are used.

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

Table 5-4. Confidence belts for estimated proportions (*p*) in an error matrix for 95 % confidence coefficient (see Figure 5-1).

		95% Confidence limits					
p	Limit	n=10	20	50	100	250	1000
0%	Upper	30.8	16.8	7.1	3.6	1.5	0.4%
	Lower	0.0	0.0	0.0	0.0	0.0	0.0%
5%	Upper		24.9	15.1	11.3	8.5	6.5%
	Lower		0.1	0.9	1.6	2.7	3.7%
10%	Upper	44.5	31.7	21.8	17.6	14.4	12.0%
	Lower	0.3	1.2	3.3	4.9	6.6	8.2%
15%	Upper		37.9	27.9	23.5	20.0	17.4%
	Lower		3.2	6.5	8.6	10.8	12.8%
20%	Upper	55.6	43.7	33.7	29.2	25.5	22.6%
	Lower	2.5	5.7	10.0	12.7	15.2	17.6%
25%	Upper		49.1	39.3	34.7	30.8	27.8%
	Lower		8.7	13.8	16.9	19.8	22.3%
30%	Upper	65.2	54.3	44.6	40.0	36.1	32.9%
	Lower	6.7	11.9	17.9	21.2	24.4	27.2%
35%	Upper		59.2	49.8	45.2	41.3	38.0%
	Lower		15.4	22.1	25.7	29.1	32.0%
40%	Upper	73.8	63.9	54.8	50.3	46.4	43.1%
	Lower	12.2	19.1	26.4	30.3	33.9	36.9%
45%	Upper		68.5	60.7	55.3	51.4	48.1%
	Lower		23.1	31.8	35.0	38.7	41.9%
50%	Upper	81.3	72.8	64.5	60.2	56.4	53.1%
	Lower	18.7	27.2	35.5	39.8	43.6	46.9%
55%	Upper		76.9	69.1	65.0	61.3	58.1%
	Lower		31.5	40.3	44.7	48.6	51.9%
60%	Upper	87.8	80.9	73.6	69.7	66.1	63.1%
	Lower	26.2	36.1	45.2	49.7	53.6	56.9%
65%	Upper		84.6	77.9	74.3	70.9	68.0%
	Lower		40.8	50.2	54.8	58.7	62.0%
70%	Upper	93.3	88.1	82.1	78.8	75.6	72.8%
	Lower	34.8	45.7	55.4	60.0	63.9	67.1%
75%	Upper		91.3	86.2	83.1	80.2	77.7%
	Lower		50.9	60.7	65.3	69.2	72.2%
80%	Upper	97.5	94.3	90.0	87.3	84.8	82.4%
	Lower	44.4	56.3	66.3	70.8	74.5	77.4%
85%	Upper		96.8	93.5	91.4	89.2	87.2%
	Lower		62.1	72.1	76.5	80.0	82.6%
90%	Upper	99.7	98.8	96.7	95.1	93.4	91.8%
	Lower	55.5	68.3	78.2	82.4	85.6	88.0%
95%	Upper		99.9	99.1	98.4	97.3	96.3%
	Lower		75.1	84.9	88.7	91.5	93.5%
100%	Upper	100	100	100	100	100	100%
	Lower	69.2	83.2	92.9	96.4	98.5	99.6%

Table 5-5. Confidence belts for estimated proportions (*p*) in an error matrix for 90 % confidence coefficient (see Figure 5-1)..

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

		90% Confidence limits					
p	Limit	n=10	20	50	100	250	1000
0%	Upper	25.9	13.9	5.8	3.0	1.2	0.3%
	Lower	0.0	0.0	0.0	0.0	0.0	0.0%
5%	Upper		21.6	13.4	10.2	7.9	6.3%
	Lower		0.3	1.2	2.0	3.0	3.9%
10%	Upper	39.4	28.3	19.9	16.4	13.7	11.7%
	Lower	0.5	1.8	4.0	5.5	7.0	8.5%
15%	Upper		34.4	25.9	22.2	19.2	17.0%
	Lower		4.2	7.5	9.5	11.4	13.2%
20%	Upper	50.7	40.1	31.6	27.7	24.6	22.2%
	Lower	3.7	7.1	11.3	13.7	15.9	17.9%
25%	Upper		45.6	37.0	33.1	29.9	27.4%
	Lower		10.4	15.3	18.0	20.5	22.8%
30%	Upper	60.7	50.8	42.4	38.4	35.1	32.5%
	Lower	8.7	14.0	19.5	22.5	25.2	27.6%
35%	Upper		55.8	47.6	43.6	40.3	37.6%
	Lower		17.7	23.8	27.1	30.0	32.5%
40%	Upper	69.6	60.6	52.6	48.7	45.4	42.6%
	Lower	15.0	21.7	28.3	31.8	34.8	37.4%
45%	Upper		65.3	57.5	53.7	50.4	47.6%
	Lower		25.9	32.9	36.5	39.7	42.4%
50%	Upper	77.8	69.8	62.4	58.6	55.4	52.6%
	Lower	22.2	30.2	37.6	41.4	44.6	47.4%
55%	Upper		74.1	67.1	63.5	60.3	57.6%
	Lower		34.7	42.5	46.3	49.6	52.4%
60%	Upper	85.0	78.3	71.7	68.2	65.2	62.6%
	Lower	30.4	39.4	47.4	51.3	54.6	57.4%
65%	Upper		82.3	76.2	72.9	70.0	67.5%
	Lower		44.2	52.4	56.4	59.7	62.4%
70%	Upper	91.3	86.0	80.5	77.5	74.8	72.4%
	Lower	39.3	49.2	57.6	61.6	64.9	67.5%
75%	Upper		89.6	84.7	82.0	79.5	77.2%
	Lower		54.4	63.0	66.9	70.1	72.6%
80%	Upper	96.3	92.9	88.7	86.3	84.1	82.1%
	Lower	49.3	59.9	68.4	72.3	75.4	77.8%
85%	Upper		95.8	92.5	90.5	88.6	86.8%
	Lower		65.6	74.1	77.8	80.8	83.0%
90%	Upper	99.5	98.2	96.0	94.5	93.0	91.5%
	Lower	60.6	71.7	80.1	83.6	86.3	88.3%
95%	Upper		99.7	98.8	98.0	97.0	96.1%
	Lower		78.4	86.6	89.8	92.1	93.7%
100%	Upper	100	100	100	100	100	100%
	Lower	74.1	86.1	94.2	97.0	98.8	99.7%

Table 5-6. Confidence belts for estimated proportions (*p*) in an error matrix for 80 % confidence coefficient (see Figure 5-1)..

		80% Confidence limits					
p	Limit	*n*=10	20	50	100	250	1000

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

		80% Confidence limits					
0%	Upper	20.6	10.9	4.5	2.3	0.9	0.2%
	Lower	0.0	0.0	0.0	0.0	0.0	0.0%
5%	Upper		18.1	11.6	9.1	7.3	6.0%
	Lower		0.5	1.6	2.5	3.3	4.1%
10%	Upper	33.7	24.5	17.8	15.0	12.9	11.3%
	Lower	1.0	2.7	4.9	6.3	7.6	8.8%
15%	Upper		30.4	23.6	20.6	18.3	16.5%
	Lower		5.6	8.8	10.5	12.1	13.6%
20%	Upper	45.0	36.1	29.1	26.1	23.6	21.7%
	Lower	5.5	9.0	12.8	14.9	16.7	18.4%
25%	Upper		41.5	34.5	31.4	28.9	26.8%
	Lower		12.7	17.1	19.4	21.5	23.2%
30%	Upper	55.2	46.7	39.8	36.6	34.0	31.9%
	Lower	11.6	16.6	21.5	24.0	26.2	28.1%
35%	Upper		51.8	45.0	41.8	39.1	37.0%
	Lower		20.7	26.0	28.7	31.0	33.0%
40%	Upper	64.6	56.7	50.1	46.9	44.2	42.0%
	Lower	18.8	24.9	30.6	33.4	35.9	38.0%
45%	Upper		61.5	55.0	51.9	49.2	47.1%
	Lower		29.3	35.3	38.3	40.8	42.9%
50%	Upper	73.3	66.2	59.9	56.9	54.2	52.1%
	Lower	26.7	33.8	40.1	43.1	45.8	47.9%
55%	Upper		70.7	64.7	61.7	59.2	57.1%
	Lower		38.5	45.0	48.1	50.8	52.9%
60%	Upper	81.2	75.1	69.4	66.6	64.1	62.0%
	Lower	35.4	43.3	49.9	53.1	55.8	58.0%
65%	Upper		79.3	74.0	71.3	69.0	67.0%
	Lower		48.2	55.0	58.2	60.9	63.0%
70%	Upper	88.4	83.4	78.5	76.0	73.8	71.9%
	Lower	44.8	53.3	60.2	63.4	66.0	68.1%
75%	Upper		87.3	82.9	80.6	78.5	76.8%
	Lower		58.5	65.5	68.6	71.1	73.2%
80%	Upper	94.5	91.0	87.2	85.1	83.3	81.6%
	Lower	55.0	63.9	70.9	73.9	76.4	78.3%
85%	Upper		94.4	91.2	89.5	87.9	86.4%
	Lower		69.6	76.4	79.4	81.7	83.5%
90%	Upper	99.0	97.3	95.1	93.7	92.4	91.2%
	Lower	66.3	75.5	82.2	85.0	87.1	88.7%
95%	Upper		99.5	98.4	97.5	96.7	95.9%
	Lower		81.9	88.4	90.9	92.7	94.0%
100%	Upper	100	100	100	100	100	100%
	Lower	79.4	89.1	95.5	97.7	99.1	99.8%

Table 5-7. Confidence belts for estimated proportions (*p*) in an error matrix for 50 % confidence coefficient (see Figure 5-1)..

			50% Confidence limits				
p	Limit	*n*=10	20	50	100	250	1000
0%	Upper	12.9	6.7	2.7	1.4	0.6	0.1%

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

				50% Confidence limits			
	Lower	0.0	0.0	0.0	0.0	0.0	0.0%
5%	Upper		12.9	8.8	7.3	6.3	5.5%
	Lower		1.4	2.7	3.4	4.0	4.5%
10%	Upper	24.7	18.7	14.5	12.8	11.6	10.7%
	Lower	2.8	4.8	6.8	7.8	8.6	9.3%
15%	Upper		24.2	19.9	18.2	16.8	15.8%
	Lower		8.7	11.2	12.4	13.4	14.2%
20%	Upper	35.5	29.6	25.2	23.4	22.0	20.9%
	Lower	9.6	12.8	15.7	17.0	18.2	19.1%
25%	Upper		34.8	30.5	28.6	27.1	26.0%
	Lower		17.1	20.3	21.8	23.0	24.0%
30%	Upper	45.8	40.0	35.6	33.7	32.2	31.0%
	Lower	17.6	21.6	25.0	26.6	27.9	29.0%
35%	Upper		45.1	40.7	38.8	37.3	36.1%
	Lower		26.1	29.7	31.4	32.8	33.9%
40%	Upper	55.5	50.1	45.8	43.9	42.3	41.1%
	Lower	26.1	30.7	34.5	36.3	37.7	38.9%
45%	Upper		55.0	50.8	48.9	47.3	46.1%
	Lower		35.4	39.4	41.2	42.7	43.9%
50%	Upper	64.9	59.8	55.7	53.9	52.3	51.1%
	Lower	35.1	40.2	44.3	46.1	47.7	48.9%
55%	Upper		64.6	60.6	58.8	57.3	56.1%
	Lower		45.0	49.2	51.1	52.7	53.9%
60%	Upper	73.9	69.3	65.5	63.7	62.3	61.1%
	Lower	44.5	49.9	54.2	56.1	57.7	58.9%
65%	Upper		73.9	70.3	68.6	67.2	66.1%
	Lower		54.9	59.3	61.2	62.7	63.9%
70%	Upper	82.4	78.4	75.0	73.4	72.1	71.0%
	Lower	54.2	60.0	64.4	66.3	67.8	69.0%
75%	Upper		82.9	79.7	78.2	77.0	76.0%
	Lower		65.2	69.5	71.4	72.9	74.0%
80%	Upper	90.4	87.2	84.3	83.0	81.8	80.9%
	Lower	64.5	70.4	74.8	76.6	78.0	79.1%
85%	Upper		91.3	88.8	87.6	86.6	85.8%
	Lower		75.8	80.1	81.8	83.2	84.2%
90%	Upper	97.2	95.2	93.2	92.2	91.4	90.7%
	Lower	75.3	81.3	85.5	87.2	88.4	89.3%
95%	Upper		98.6	97.3	96.6	96.0	95.5%
	Lower		87.1	91.2	92.7	93.7	94.5%
100%	Upper	100	100	100	100	100	100%
	Lower	87.1	93.3	97.3	98.6	99.4	99.9%

2.2 Response design

The response design specifies how to determine the reference classification at a given sample point. The first step is to select the "spatial support region" for which the reference classification is made (Stehman and

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

Czaplewski 1998). The support region for a pixel map is discussed first, followed by consideration of a polygon map.

2.2.1 Pixel map

Assume the Map Object is a 30 m pixel. The support region might be a 30 m square area that surrounds the sample point, even though the true "boundaries" of the target pixel on the map are not well known in the field or on an aerial photograph. Or the support area could be a circle with 50 m radius centered on the sample point. Regardless, the objective is to determine, as accurately as possible, the correct reference classification of the target pixel. This is accomplished within a support region that is well defined on the ground or an aerial photograph.

Misregistration of sample points to the thematic map can bias the accuracy assessment statistics. The classification on the map at the intended sample point might be correct, but misregistration causes the reference classification to be compared to a different, nearby pixel, which might not have the same classification on the map. A larger support region can partially compensate for misregistration between the map and the sample point. However, if the support region becomes too large, then the reference classifier may not well represent the true classification of the single map pixel that surrounds the sample point. The support region should be similar to the spatial scale of the Map Object.

Some studies lower the spatial resolution of the map near a sample point so that the sampled pixel will likely have the same map classification as the intended pixel. For example, assume the map consists of unique classifications of individual pixels (i.e., a moving window is not used to spatially filter classifications of adjacent pixels). However, during the accuracy assessment, a 3x3 block of pixels is formed around each sample point used for reference data, and some rule is used to assign a single map category to this block of nine potentially heterogeneous map pixels. This can mitigate affects of registration error, and it is possible to compare a reference classification for the sample point to the classification of this surrounding block of pixels. Regrettably, this process assesses the accuracy of an imaginary thematic map to which the same filtering rule was applied to all pixels, not the individual-pixel map that was actually produced. The accuracy of the unfiltered map might be similar to the accuracy of the filtered map, but this is speculative. If registration error is significant, then perhaps the entire map should be subjected to the same filtering rule. At least the map being assessed is the same map being delivered.

Boundaries between adjacent forest stands and other types of land cover are often more heterogeneous than interiors of stands. Therefore, the effect of registration error can be greatest near stand edges. Some studies move the sample point away from edges into conditions that are more homogeneous.

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

This can be useful for developing training data in a digital classification. However, this same practice can reduce the scope of the accuracy assessment because some parts of the thematic map are excluded from the target population. Edge conditions can be very prevalent in detailed thematic maps for diverse landscapes. If the producer of the accuracy assessment avoids boundary conditions, such areas should be identified for the entire map with a GIS procedure. The extent and mapped composition of the excluded zone can be tabulated and reported in the assessment documentation, and the user of the assessment can judge the value of an assessment that is limited to interior conditions.

Alternative methods are available to mitigate the effect of registration errors on the quality of the accuracy assessment. In addition to producing the reference classification, the field crew or photo-interpreter can record data that indicate the likelihood of misregistration error:

1. Distance from the sample point to a stand boundary or other edge condition. The frequency of misclassification errors can be compared with the distance from an edge to gain insights into the nature of misregistration and misclassification errors.
2. Classification of stands based on internal heterogeneity relative to the scale of the pixel (e.g., dense and uniform fine-grained canopy, frequent gaps with coarse-grained canopy). If the apparent misclassification error is remarkably higher in heterogeneous stands, then misregistration error might be a significant problem.

2.2.2 Polygon map

Assume a Map Object is a 10 ha forest stand delineated on a thematic map through image segmentation or photo-interpretation. The entire stand is classified into one and only one category on the map.

The support area is the entire polygon that surrounds a sample point. The Reference Classifier is applied to the entire polygon from the map, not a portion of the polygon in the immediate vicinity of the sample point. It is important to use the polygon boundaries on the map when applying the Reference Classifier, and not use a special polygon delineation procedure that is only applied to each sample point. Otherwise, the analysis will not assess the thematic map being evaluated. Rather, it will assess an imaginary map that would have been produced if the special delineation procedure used for the accuracy assessment were applied to the entire thematic map.

2.2.3 Labeling Protocol for Reference Data

The next part of the Response Design is specification of the labeling protocol for the support region around each sample point. This could be

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

photo-interpretation with higher-resolution imagery, qualitative observations by a field crew, or sub-sampling and physical measurements by a field crew.

The labeling protocol needs to consider the size and configuration of the support region (as discussed above). If the support region is large (e.g., a polygon), then inexpensive methods might be necessary, such as photo interpretation or low-intensity stand examinations. If the support region is small (e.g., 1 ha or smaller), then more expensive classifications using tree measurements might be feasible. The advantage of the latter is more precise and repeatable reference classifications of forest conditions. The value of the accuracy assessment depends upon the credibility of the labeling protocol to the user of the assessment.

The quality of reference data also depends on compatibility of definitions. For example, there are a variety of definitions for forest and old-growth forest, and the response design must use definitions acceptable to users of the assessment. Otherwise, the investment in the accuracy assessment might yield little practical value. However, there can be advantages with a more detailed classification system for the reference data. For example, assume that the reference data separates "shrubby wetlands" from "forested wetlands", but the map groups both types of wetland into a single category called "wooded wetlands". The reference data statistically estimate the proportion of wooded wetlands that are actually shrubby and forested woodlands, even though these detailed categories are not separated on the map.

The quality of reference data is often a compromise between the ideal and the feasible. For example, interpretation of aerial photography is an inexpensive alternative to field classifications. Photo-interpretation is acceptable if the resolution of the photography is sufficient relative to the thematic classification system, the photography is available for the entire sampled population, the acquisition dates reasonably coincide with the remotely sensed imagery used to produce the thematic map, and these materials are acceptable to the user of the accuracy assessment

There are often variations among photo-interpreters when classifying the land cover for a reference point. A similar problem occurs with field crews who produce "ocular" classifications that are not based on physical measurements. To partially mitigate this source of uncertainty with the reference data, multiple interpreters can classify the same reference points, and a majority rule engaged to select the classification used in the accuracy assessment. Sample points that have different reference classifications among interpreters can be inspected more closely, and perhaps a different protocol used to determine a more reliable reference classification.

Secondary classifications of the sample point can also provide useful reference data. For example, a sample point in a forest with diverse conditions could be classified by an interpreter (in the field or from an aerial photograph) as a pine stand, but the interpreter could also record that the

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

stand might nearly meet the definition of a mixed pine/hardwood stand. Some assessments consider the map classification to be correct if it agrees with either the primary or secondary reference classification. This approach can be acceptable if well documented and apparent to users of the accuracy assessment. However, it is prudent to publish a complementary assessment, in which only the primary classification is used as reference data. This allows different users to choose the assessment that best meets their standards, and broadens the potential applications of the error matrix, such as calibration of areal estimates for misclassification bias (Czaplewski 1992).

Fuzzy-set classifications (e.g., Gopal and Woodcock 1994) are another type of reference protocol. An interpreter classifies each sample plot into categories such as "Correct", "Acceptable", "Not right" or "Very wrong". An interpreter needs to know the map classification before making the reference classification into a fuzzy set. This could influence the reference interpretation, perhaps leading to biased estimates of classification accuracy.

Metadata should document the quality of the reference data. If multiple interpreters are used to collect reference data, a then a summary of the variability among interpreters provides useful information for quantitative characterization. In addition, interpreters can classify each sample point based on their own confidence in the interpretation, for example: "Confident that the interpretation is correct", "Uncertain classification", or "Nearly a guess". Such information may aid in a detailed analysis of the accuracy assessment data.

2.3 Analysis

The analysis design specifies how to mathematically combine the data collected in Steps 1 and 2 to accurately infer accuracy assessment statistics, which are listed in the beginning of this Chapter, for the sampled population. Stehman and Czaplewski (1998) provide general guidance and key references. This Chapter only considers the two simple designs that can be reliably analyzed by an informed user: simple random point sampling, and stratified random point sampling based on map categories as strata. Table 5-8 provides the estimators for accuracy assessment statistics and their variances for these two designs.

The analysis must be compatible with the statistical design. Otherwise, estimates can be biased and misleading. For example, consider the stratified random sample described in Table 5-3. If these data were incorrectly analyzed as though they came from a simple random sample, then the biased estimate of producer's accuracy for old-growth would be 22/25=88.0 % (Table 5-3A) rather than the unbiased estimate of 2.8/7.3=38.4 %

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

Table 5-8. Estimators for simple random sampling and stratified random sampling, where each map category is a stratum.

Sampling estimation symbols, descriptions, and equations	
Symbol	**Description**
i	Row subscript, which designates the Map Classifier in the error matrix, and stratum *i* in stratified random sampling
j	Column subscript, which designates the Reference Classifier in the error matrix
n_{ij}	Number of sample points in row *i* and column *j*
$n_{i\bullet}$	Total number of sample points in row *i*
n	Total number of sample points
η_k	"Effective" sample size for estimate *k*, which is used to approximate the confidence interval with certain estimates
A	Total area of sampled population (known exactly from GIS)
$A_{\bullet j}$	Total area of sampled population in class *i* had all Map Objects been classified with the Reference Classifier (estimated from accuracy assessment sample)
$A_{i\bullet}$	Total area of sampled population in map class *i* (known exactly from GIS)
p_{ij}	Proportion of sampled population classified as category *i* by the Map Classifier and category *j* by the Reference Classifier
$p_{i=j}$	Proportion of sampled population in which the Map and Reference Classifiers agree (i.e., overall accuracy)
$p_{i\bullet}$	Proportion of sampled population classified as category *i* by the Map Classifier
$p_{\bullet j}$	Proportion of sampled population classified as category *j* by the Reference Classifier
$p_{j=Y\|i=X}$	Proportion of sampled population in reference category class *Y* given that the map classification is category *X*. For *X*=*Y*, $p_{(j=X\|i=X)}$ equals "user's accuracy."
$p_{i=X\|j=Y}$	Proportion of sampled population in map class *X* given that the reference classification is category *Y*. For *X*=*Y*, $p_{(i=Y\|j=Y)}$ equals "producer's accuracy."
var(*p*)	Variance of the estimated proportion *p*
Simple Random Sample	

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, \mathrm{var}\left(\hat{p}_{ij}\right) = \hat{p}_{ij}\left(1-\hat{p}_{ij}\right)/n$$

$$\hat{p}_{i\bullet} = \sum_j \frac{n_{ij}}{n}, \mathrm{var}\left(\hat{p}_{i\bullet}\right) = \frac{\hat{p}_{i\bullet}\left(1-\hat{p}_{i\bullet}\right)}{n}$$

$$\hat{p}_{\bullet j} = \sum_i \frac{n_{ij}}{n}, \mathrm{var}\left(\hat{p}_{\bullet j}\right) = \frac{\hat{p}_{\bullet j}\left(1-\hat{p}_{\bullet j}\right)}{n}$$

$$\hat{p}_{i=j} = \sum_i \frac{n_{ii}}{n}, \mathrm{var}\left(\hat{p}_{i=j}\right) = \frac{\hat{p}_{i=j}\left(1-\hat{p}_{i=j}\right)}{n}$$

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

Sampling estimation symbols, descriptions, and equations
$\hat{p}_{j=Y\|i=X} = \dfrac{\hat{p}_{ij}}{\hat{p}_{i\bullet}}$, $\mathrm{var}\left(\hat{p}_{j=Y\|i=X}\right) = \dfrac{\left(\hat{p}_{i\bullet} - \hat{p}_{ij}\right)\hat{p}_{ij}}{n\hat{p}_{i\bullet}^3}$
$\hat{p}_{i=X\|j=Y} = \dfrac{\hat{p}_{ij}}{\hat{p}_{\bullet j}}$, $\mathrm{var}\left(\hat{p}_{i=X\|j=Y}\right) = \dfrac{\left(\hat{p}_{\bullet j} - \hat{p}_{ij}\right)\hat{p}_{ij}}{n\hat{p}_{\bullet j}^3}$
Stratified random sample (stratum = *i*)
$\hat{p}_{ij} = \left(\dfrac{n_{ij}}{n_{i\bullet}}\right)\hat{p}_{i\bullet}$, $\mathrm{var}\left(\hat{p}_{ij}\right) = \dfrac{\hat{p}_{ij}\left(\hat{p}_{i\bullet} - \hat{p}_{ij}\right)}{n_{i\bullet}}$
$p_{i\bullet} = A_{i\bullet}/A$, $\mathrm{var}\left(p_{i\bullet}\right) = 0$
$\hat{p}_{\bullet j} = \sum_i \hat{p}_{ij}, \mathrm{var}\left(\hat{p}_{\bullet j}\right) = \sum_i \mathrm{var}\left(\hat{p}_{ij}\right)$
$\hat{p}_{i=j} = \sum_i \hat{p}_{ii}, \mathrm{var}\left(\hat{p}_{i=j}\right) = \sum_i \mathrm{var}\left(\hat{p}_{ii}\right)$
$\hat{p}_{j=Y\|i=X} = \dfrac{\hat{p}_{ij}}{p_{i\bullet}}$, $\mathrm{var}\left(\hat{p}_{j=Y\|i=X}\right) = \dfrac{\hat{p}_{j=Y\|i=X}\left(1 - \hat{p}_{j=Y\|i=X}\right)}{n_{i\bullet}}$
$\hat{p}_{i=X\|j=Y} = \dfrac{\hat{p}_{ij}}{\hat{p}_{\bullet j}}$, $\mathrm{var}\left(\hat{p}_{i=X\|j=Y}\right) = \left(\dfrac{\hat{p}_{ij}^2}{\hat{p}_{\bullet j}^4}\right)\mathrm{var}\left(\hat{p}_{\bullet j}\right) - \left(\dfrac{2\hat{p}_{ij} - \hat{p}_{\bullet j}}{\hat{p}_{\bullet j}^3}\right)\mathrm{var}\left(\hat{p}_{ij}\right)$
Effective sample size for each of the above estimators
$\eta_e = p_k(1-p_k)/\mathrm{var}(p_k)$, for k={ij; i=j $i\bullet$; $\bullet j$; i=X|j=Y; j=Y|i=X}
Conversion of proportions to areas for each of the above estimators
$\hat{A}_k = \hat{p}_k A, \mathrm{var}\left(\hat{A}_k\right) = \mathrm{var}\left(\hat{p}_k\right)A^2$, for k={ij; i=j $i\bullet$; $\bullet j$; i=X|j=Y; j=Y|i=X}
Cohen's kappa coefficient of agreement
See Hudson & Ramm (1987), Czaplewski (1994), Campbell (1996), Stehman (1996)

in Table 5-3B. Likewise, the biased estimate for the true area of old-growth would be (25/100)·1000000=250,000 ha from Table 5-3A, whereas the unbiased estimate from Table 5-3B is 73,070 ha. Table 5-8 compares the estimators for accuracy assessment statistics for both sampling designs.

Often overlooked is the effect of random sampling error on estimated accuracy statistics. Any sample estimate includes uncertainty because inference is made for the entire population based on a very small sample of that population. Sometimes this uncertainty can be large. For example, Tables 5-1 and 5-3B provide different estimates for the same sampled population. The extent of old-growth forest in the sampled population that is actually mapped as old-growth (i.e., producer's accuracy) is estimated to be 60.0 % from Table 5-1 and 38.4 % from Table 5-3B. Likewise, the true total area of old-growth forest in the sampled population is estimated at 100,000 ha with Table 5-1 and 73,070 ha with Table 5-3B. All of these are unbiased

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

because estimates from the two different sampling designs would converge on the same true value given a sufficiently large sample size. The differences among estimates are caused by different degrees of random sampling error. The magnitude of sampling error can be objectively estimated from "variance statistics" (Table 5-8) and communicated as confidence intervals Tables 5-4 to 5-8).

2.3.1 Confidence Intervals for Accuracy Assessment Statistics

Confidence intervals are important to interpretation of statistics from any probability sample. For example, the 90 % confidence interval for overall accuracy in Table 5-1 is 68 % to 83 % (i.e., there is a 5 % chance that the true accuracy is less than 68 %, and another 5 % chance that the true accuracy is greater than 83 %). If a similar estimate were obtained with a sample size of 20 points, then the 90 % confidence interval would be 54 % to 90 % (Table 5-5). If an overall accuracy of 70 % is required from the sampled portion of the map, then a sample size of 20 points is insufficient to determine success with much confidence. Another example is the estimated area of old-growth forest in Table 5-3B, which is 7 % or 70,000 ha. From Table 5-8, the estimated variance is var($p_{\bullet 2}$)=0.000672, which gives an effective sample size of $\eta_{\bullet j}$ =101 (Table 5-8). The approximate 90 % confidence interval is 3 % to 13 % of the sampled population (interpolated from Table 5-5), which equals 30,000 ha to 130,000 ha. If a similar estimate were made from a stratified random sample of 1000 points, then the 90 % confidence interval would be 6 % to 9 % or 60,000 ha to 90,000 ha. If confidence intervals are ignored, then there is a high risk of misinterpreting the results of an accuracy assessment.

2.3.2 Confidence Coefficient

Figure 5-1 and Tables 5-4 to 5-7 give confidence intervals for the 95 %, 90 %, 80 % and 50 % confidence coefficients. The choice of the confidence coefficient depends on the risks of incorrect inference to the user of the thematic map. If the user needs to be relatively certain that the true value is within the confidence interval, then the 95 % level would be a good choice. There is only a 2.5 % chance that the true value is less than the lower bound of the confidence interval, and another 2.5 % chance that the true value exceeds the upper bound of the interval. If more risk is acceptable, then the 50 % level could be used; there would be a 25 % chance that the true value is less than the lower bound of the confidence interval, and another 25 % chance that the true value exceeds the upper bound of the interval. For example, assume a mapped stratum is sampled with 50 points; and assume that the estimate of user's accuracy is 90 % (i.e., 45 of the 50 points are classified as forest with the reference data). The confidence interval at the 95

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

% level for the estimate of user's accuracy would be 78.2 % to 96.7 % (Table 5-4), while the confidence interval for the same estimate at the 50 % level would be 85.5 % to 93.2 % (Table 5-7).

2.3.3 Computation of Confidence Intervals

Confidence intervals can be interpolated from Tables 5-4 to 5-7. However, these functions are non-linear, and interpolations can be inaccurate, especially for sample sizes that vary considerably from those in the tables. Alternatively, several confidence interval calculators for proportions (i.e., the binomial distribution) are available on the Internet.

Table 5-8 provides approximate methods to estimate confidence intervals using an "effective sample size" (η_e) with Tables 5-4 to 5-7. This procedure will yield the exact interval, except for any statistic that is a ratio of two estimates (e.g., producer's accuracy). In this latter case, the confidence interval computation requires a Taylor-series approximation for the estimated variance, plus the beta-binomial distribution and specialized software for a numerical solution. In this case, use of an effective sample size with the binomial distribution is a convenient approximation. The effective sample size can be a non-integer, and is rounded to the nearest integer value.

I suggest using the following figure, which is more concise than previous ones. I originally envisioned the reader using the figure to interpolate confidence intervals for their own applications. That's why I made them big. However, I later added tables 5-4 to 5-7 to make the interpolations more accurate. Now, the purpose of the figure is to give reader a visualization of the tradeoffs between confidence intervals, sample size and confidence coefficient.

Suggest this figure be moved into Section 2.1.3 .

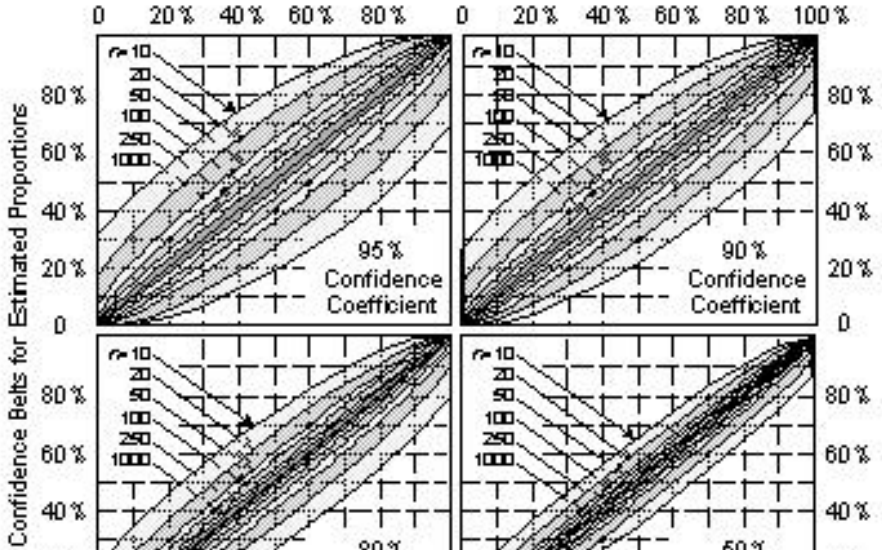

Figure 5-1. Confidence belts for estimated proportions in the cells and margins of an error matrix. The confidence coefficient is the probability that the true (but known) value is within the confidence belt. The effective sample size is denoted as *n* (see Table 5-8). Use Tables 5-4 to 5-7 to make more precise interpolations.

2.3.4 Cells that Equal Zero and Perfect Accuracy

Through random sampling error, a cell in the error matrix, usually representing a classification error, can have a value of zero. The true value is likely small, but greater than zero. Tables 5-4 to 5-7 include confidence intervals that cover such unobserved cross-classifications. For example, in Table 5-1, there were no cases among the n=100 sample points in which the map classification is old-growth forest and the reference classification is non-forest. Therefore, the estimated proportion of the map that is classified as old-growth forest and is truly non-forest is 0. However, from Table 5-5, the 90 % confidence interval for this estimate is 0 % to 3 %, or up to 30,000 ha; this means that there is an estimated 10 % chance that the true number of hectares so cross-classified exceeds 30,000 ha. A similar situation arises when all sample points are correctly classified, which is most likely observed with producer's or user's accuracy for rare category. For example, assume that the producer's accuracy is estimated at 100 % with an effective sample size of η_e=10. From Table 5-5, the approximate 90 % confidence interval would be 74.1 % to 100 %, meaning that there is a 10 % chance that the true producer's accuracy for that category is less than 74.1 %. See Lloyd (1999) for a more detailed summary of these issues.

2.4 Quality assurance

In addition to the usual care required during data collection and management, other steps can help assure the quality of an accuracy assessment. One very useful diagnostic tool is estimation of known values with the sample of reference points and the selected analysis design (i.e., Steps 1 and 3 above). If the sample estimates do not reasonably agree with the known values, then a procedural error is likely. For example, the sampled area of the map labeled as non-forest in Table 5-1 is exactly 549,020 ha, which is determined through a tally of all Map Objects in the GIS or image processing system. The same area estimated from the 100 sample points is 440,000 ha, with a 95 % confidence interval of 341,000 ha to 543,000 ha (interpolated with Table 5-5). This confidence interval does not cover the known true value; there is less than a 5 % chance that the true value (549,020 ha) is outside the bounds of the 95 % confidence interval (i.e., more than 543,000 ha or less than 341,000 ha). In one out of every 20 accuracy assessments, an unfortunate, but valid, random sample of points could cause this degree of apparent discrepancy. However, this discrepancy remains weak evidence for a potential procedural error. For example, the sampled population used to draw the sample might be defined differently than the sampled population used to tally the area statistics in Table 5-1.

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

Displaying the location of all sample points on a map base might reveal geographic areas that were unknowingly omitted from the sampled population. Alternatively, a stratified random sample might have been analyzed as though it was a simple random sample, or the thematic map might have been modified after the stratified sample was drawn. Whenever possible, diagnostic tests should be conducted before gathering expensive reference data in Step 2 to detect procedural errors in Steps 1 and 3.

In the case of stratified random sampling, the thematic map is used to define strata, and different comparisons must be used to detect discrepancies. For example, the stratified sample of points could be used to estimate the known areas of different soil types, elevation zones, or administrative areas in a GIS database.

2.5 Interpretation of results

The investment in a valid accuracy assessment has little value unless the results are effectively communicated to the user. This Chapter has already presented examples of informative statistics and their interpretations. The producer of an accuracy assessment should provide similar interpretations for their own error matrix. Stacked bar charts can provide a visual display of a large error matrix, and such displays can provide insights. The full error matrix or fuzzy-set contingency table should be published as metadata, thus maximizing options for alternative interpretations.

The terms “Producer’s Accuracy” and “User’s Accuracy” are prevalent in today’s remote sensing literature, but these terms are somewhat misleading. Both types of statistics are important to both the producer of a remotely sensed map and a user of that map. These and other conditional probabilities are more useful statistics than widely recognized. For example, Table 5-1 estimates that only 60 % of the old-growth forest, as defined in Step 2, is labeled as old-growth forest on the map; an estimated 40 % of the old-growth forest truly exists in the sampled population, but it is mislabeled as something else on the map, and its location is unknown on the map. As another example, Table 5-1 estimates that the old-growth forest category on the map is truly composed of 75 % old-growth forest and 25 % other types of forest. Therefore, the conditional probabilities from the error matrix document the composition of each thematic category as metadata.

Marginal proportions are also important metadata. For example, Table 5-3B estimates that the true extent of old-growth forest is 73,070 ha within the sampled population, with a 90 % confidence interval of 33,000 ha to 126,000 ha. However, only 41,634 ha are mapped as old-growth forest, and only 75 % of that mapped area is estimated to be truly old-growth forest in the sampled population. Therefore, analyses with the thematic map will likely underestimate attributes associated with old-growth forest. Both misclassification error and random sampling error cause these differences

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

between the map and estimates of the true exist conditions. Czaplewski (1992) discusses these "discrepancies" in more detail.

3. CONCLUDING REMARKS

There is a large body of remote sensing literature devoted to accuracy assessments. However, this chapter is limited to a few techniques that are simple and robust. The goal is to provide methods that can be applied by professionals who have no training in sample survey statistics. Large mapping projects often use elements that are more complex (e.g., cluster plots, two-phase and two-stage sampling, alternative stratification materials, multiple sampling frames) to reduce costs and improve efficiency (Stehman and Czaplewski 1998; Czaplewski 2000). However, complex designs require consultation with a statistician knowledgeable in sample surveys. Otherwise, there is high risk of using invalid methods that produce unreliable and biased estimates of accuracy assessment statistics. The bias can be large and can lead to unwise decisions.

Too often, the number of sample sites for reference data is inadequate for sufficiently precise estimates of accuracy assessment statistics. For example, assume the true user's accuracy for old-growth forest is 75 %. Unfortunately, this true value is never known in real-world applications, and an imperfect estimate from a sample is typically the best available information. With an effective sample size of 20 reference sites, there is about one chance in four that the estimated accuracy will be less than 65 %, even though the true accuracy is 75 %. However, with a sample of 50 reference sites, there is only about one chance in 10 that the estimated accuracy will be less than 65%. A closely related issue is precision of area estimates. For example, assume old-growth forest truly covers 120,000 ha of a 1,000,000 ha sampled population. With an effective sample size of 20 reference sites, there is about one chance in four that the extent of old-growth forest will be estimated at 50,000 ha or less. However, with a sample of 50 reference sites, the probability of this degree of underestimation decreases to one in 10. This chapter recommends that a hypothetical, although realistic, error matrix be constructed during the early planning phase of the accuracy assessment. If precision of hypothetical assessment statistics appears inadequate, then the number of reference sites should be increased, or the accuracy assessment should be omitted. There is no real value in expending scarce resources on an unreliable assessment.

The best data that can be realistically produced are imperfect estimates of classification accuracies and the true area of different forest conditions, and an imperfect map of where those conditions are located. The goal of image classification and accuracy assessment is to minimize these imperfections,

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

and reduce the risk of mis-informed decisions, within a reasonable budget. This Chapter provides methods that can help achieve this goal.

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. Remote Sensing of Forest Environments: Concepts and Case Studies. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p

REFERENCES

Campbell, J.B. (1996). *Introduction to Remote Sensing* (2rd ed.). Guilford Press, New York.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). John Wiley & Sons, New York.

Congalton, R. G. (1991). A review of accessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37, 35-46.

Czaplewski, R. L. (1992). Misclassification bias in areal estimates. *Photogrammetric Engineering and Remote Sensing*, 58, 189-192.

Czaplewski, R. L. (1994). Variance approximations for assessments of classification accuracy. *Research Paper RM-316* . Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station.

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

Czaplewski, R. L. (2000). Accuracy assessments and areal estimates using two-phase stratified random sampling, cluster plots, and the multivariate composite estimator. *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*, 79-100. Ann Arbor Press, Chelsea, MI.

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

Hudson, W. D., & Ramm, C. W. (1987). Correct formulation of the kappa coefficient of agreement. *Photogrammetric Engineering and Remote Sensing*, 53, 421-422.

Lloyd, C. J. (1999) *Statistical Analysis of Categorical Data* (3rd ed.). John Wiley & Sons, New York.

Lohr, S. L. (1999). Sampling: Design and Analysis. Brooks/Cole Publishing, Pacific Grove, CA.

Salant, P., & Dillman, D. A. (1994). *How to conduct your own survey*. John Wiley & Sons, New York.

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

Schreuder, H. T., Gregoire, T. G., & Wood, G. (1993*). Sampling methods for multiresource forest inventory*. John Wiley & Sons, New York.

Stehman, S.V. (1996), Estimating the kappa coefficient and its variance under stratified random sampling, *Photogrammetric Engineering and Remote Sensing*, 62, 401-407.

Stehman, S.V. (2002). Response to letter by G. Turk on chance correction and map evaluation. *Remote Sensing of Environment*, 82, 4.

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

Stehman, S. V. (2001). Statistical rigor and practical utility in thematic map accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, 67, 727-734.

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

Czaplewski, Raymond L. 2003. Chapter 5: Accuracy assessment of maps of forest condition: statistical design and methodological considerations, pp. 115-140. *Remote Sensing of Forest Environments: Concepts and Case Studies*. (Michael A. Wulder and Steven E. Franklin, Eds.) Kluwer Academic Publishers, Boston. 515p