[DC-04-019] Ground Verification and Accuracy Assessment

Abstract

Spatial products such as maps of land cover, soil type, wildfire, glaciers, and surface water have become increasingly available and used in science and policy decisions. These maps are not without error, and it is critical that a description of quality accompany each product. In the case of a thematic map, one aspect of quality is obtained by conducting a spatially explicit accuracy assessment in which the map class and reference class are compared on a per spatial unit basis (e.g., per 30m x 30m pixel). The outcome of an accuracy assessment is a description of quality of the end-product map, in contrast to conducting an evaluation of map quality as part of the map production process. The accuracy results can be used to decide if the map is of adequate quality for an intended application, as input to uncertainty analyses, and as information to improve future map products.

Keywords: error, error matrix, probability sampling, producer\'s accuracy, response design, sampling design, user\'s accuracy

Author & citation

Stehman, S. (2020). Ground Verification and Accuracy Assessment. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2020 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2020.1.14.

An earlier version can be found at:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Ground verification and accuracy assessment. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Acknowledgements: The author would like to thank one of the anonymous reviewers for their particularly helpful input.

Explanation

1. Definitions
2. Components of Accuracy Assessment
3. Response Design
4. Analysis
5. Sampling Design
6. Future Developments

1. Definitions

Reference class: best practical determination of the ground condition of a spatial unit

Response design: protocol for determining the reference class

Sampling design: protocol for determining which spatial units from the region mapped will serve as the basis for the accuracy estimates

Error matrix: square table of numbers from an accuracy assessment that shows the area of each combination of a mapped class with a reference class

User's accuracy of class K: the proportion of area labeled as map class K that also has K as the reference class (complement of commission error rate of class K)

Producer's accuracy of class K: the proportion of area labeled as reference class K that also has K as the map class (complement of omission error rate of class K)

2. Components of Accuracy Assessment

The fundamental steps required to produce an assessment of map quality are to select a sample of spatial units such as pixels from the region mapped, determine the reference class of each sampled unit, compare the reference class to the map class, and then estimate accuracy from the sample data. The sample units obtained to assess accuracy should be selected independently of the sample units used to train or develop the classification. A sampling protocol that should be avoided is having the same set of sampling units used for both training the classification and conducting the accuracy assessment as this typically results in overestimating accuracy. Data-splitting and cross-validation represent alternative valid approaches to determine map quality (Brus et al. 2011). Because most maps are typically provided in a raster format, the subsequent discussion will be limited to the case of a pixel being the spatial unit of the accuracy assessment.

Stehman and Czaplewski (1998) stated that three main components were required to conduct an accuracy assessment: 1) the **response design** is the protocol that assigns the reference class label or labels to a pixel and specifies how agreement between the reference class and the map class is defined; 2) the **analysis** specifies how accuracy will be quantified and the formulas for estimating these accuracy measures from the sample data; and 3) the **sampling design** specifies the subset of pixels to which the response design will be applied. In the next sections, details of these protocols are presented. Wickham et al. (2013) provide an example of a practical application in which the three protocols are integrated into an assessment of the accuracy of the National Land Cover Dataset of the United States.

3. Response Design

The response design is implemented to provide the best practical determination of the reference class for each sample pixel. Although obtaining "ground truth" may not be possible, the reference classification must be superior in accuracy to the classification of the map being assessed. Often human interpreters obtain the reference class labels, so a critical feature of the response design is that the interpreters should be unaware of the map

classification of the sample units they are interpreting.

A variety of sources of information may be used to determine the reference class, including ground visits, aerial photographs, and imagery from unmanned aerial vehicles. However, it is often too expensive and sometimes impractical to obtain such information. For example, in a monitoring program of land cover change from 1990 to the present derived from historical satellite imagery, collecting ground information for accuracy assessment would be impossible, and satellite imagery may be the only available information to determine the reference class. If the only information available to determine the reference class is the same imagery that was used to obtain the map classification, the required superior accuracy of the reference classification then often depends on expert human interpreters focusing intensive effort to assign the reference label to each sampled pixel. A key advantage of sampling is that it allows such intensive focus because of the much smaller number of sample pixels relative to all pixels comprising the entire map.

In the simplest case, a single reference class label is assigned to each sampled pixel. However, some pixels may be mixed having more than one class, as for example when a pixel contains a paved road passing between fields planted in crops. Reference class ambiguity can be addressed by assigning primary and secondary reference class labels, or by implementing a fuzzy class labeling protocol (Gopal and Woodcock 1994). The analysis (Section 4) is presented for the case in which a single reference class label is assigned, and this label will be compared to the map label of the pixel to determine agreement.

4. Analysis

The analysis of accuracy focuses on the error matrix and accuracy measures derived from this matrix (Foody 2002). In the example error matrix shown in Table 1, the rows represent the map classification, the columns represent the reference classification, and the cell entries represent the percent area of the region of interest (ROI). Entries on the diagonal of the confusion matrix represent correct classifications, and off-diagonal cells represent classification errors. The formulas for estimating accuracy and the standard errors depend on the sampling design. Olofsson et al. (2014) provide these formulas for simple random, systematic, and stratified sampling and Stehman (1997) provides formulas for cluster sampling.

A recommended good practice when estimating the error matrix from the sample is to report the percent or proportion of area for each cell rather than to report sample counts. If the error matrix estimated from the sample is reported in terms of percent or proportion of area, the formulas for estimating the accuracy measures are the same as those applied to calculate the measures from a population error matrix (i.e., the population error matrix would result from a census of reference classification of the ROI). The common accuracy measures derived from an error matrix in which the cell entries represent percent or proportion of area are the following: 1) overall accuracy, which is the sum of the diagonal entries of the error matrix; 2) user’s accuracy for class K, which is the diagonal cell entry representing class K divided by the row total of class K; and 3) producer’s accuracy for class K, which is the diagonal cell entry representing class K divided by the column total of class K.

Table 1. Example Error Matrix Estimated from a Sample. The cell entries represent the percent of area of the ROI that is mapped as the class indicated by the row label and has reference condition indicated by the column label. For example, 7% of the area of the ROI is mapped as water but has reference condition of natural surface.

		Reference Class				
	Class	Artificial surface	Natural surface	Water	Total	User's Accuracy (%)
Map Class	Artificial surface	6	1	1	8	75
	Natural surface	4	50	10	64	78
	Water	0	7	21	28	75
	Total	10	58	32	100	
	Producer's Accuracy (%)	60	86	66		Overall = 77%

Overall accuracy does not provide class-specific accuracy information, so reporting user's and producer's accuracies is strongly recommended. User's accuracy describes the percent of the area mapped as a specific class that is in reality that class and so is the complement of the percent of map commission error for that class. For example (Table 1), user's accuracy of artificial surface is 6/8=0.75 or 75% and commission error of artificial surface is, therefore, 1-0.75=0.25 or 25%. Producer's accuracy describes the percent of the area with ground condition of the class that is actually mapped as that class, and so is the complement of percent of map omission error for that class. For example, producer's accuracy of water is 21/32=0.66 or 66%, and omission error is 1-0.66=0.34 or 34%.

Liu et al. (2007) provide an excellent overview of other measures of accuracy. Few of the alternative measures reviewed by Liu et al. (2007) have led to enhanced understanding of accuracy, with kappa being the most commonly reported despite strong recommendations discouraging its use (Pontius and Millones 2011; Foody 2020). More recent work by Pontius and Santacruz (2014) separates the difference between map and reference classifications into three components, called Quantity, Exchange, and Shift. Pontius (2019) provides a case study example illustrating the promise of how analysis of the intensity of these components can yield additional insight into classification error.

The row and column totals of the confusion matrix provide important information regarding the distribution or composition of the classes in the ROI. These totals should be examined to confirm that they are consistent with the reality of the ROI. For example, from the Table 1 confusion matrix, a map user should assess whether the distribution of the map classes in the ROI (i.e., the actual percent composition of the classes in the full map) aligns with the map composition derived from the accuracy assessment. Row and column totals may be inconsistent with reality if, for example, sampling is limited to homogeneous areas or to easily accessed locations such as along roads. In such cases, classes that form large contiguous patches or that frequently occur near roads will show greater percent area than is true of the ROI. Another common situation in which the row and column totals of the error matrix are inconsistent with the actual distributions of the classes occurs when the analysis protocol is incorrect, such as by applying simple random sampling estimation formulas to a stratified sampling design. For example, if the sample sizes for all strata are equal but simple random sampling formulas are incorrectly used to estimate the error matrix, the row totals for all classes will be equal. Given that it is highly unlikely that the

ROI has all classes represented by equal percent area, an error matrix with equal row totals would be a strong indication that a stratified sample was not analyzed correctly. The error matrix produced by applying the proper stratified estimation formulas would have row totals that were consistent with map percent composition of the classes in the ROI. The ill-advised practice of normalizing an error matrix also creates the problem that the row and column totals of the error matrix do not match the reality of the class composition of the ROI (Stehman 2004).

5. Sampling Design

Sampling is a critical component of the data collection protocol because it is not practical to obtain the reference class for all pixels in the ROI. The decision of what sampling design to implement depends on discerning the strengths and weaknesses of the candidate sampling designs to achieve the objectives of the accuracy assessment while accommodating the prioritized desirable design criteria outlined in the next paragraph. Typical objectives are estimating overall accuracy and accuracy by class. Accuracy by subregion is often of interest if the ROI is large, for example accuracy by country if the ROI is a continent.

Desirable design criteria to consider when planning the sampling design include: 1) the sampling design satisfies the criteria defining a probability sampling design (see Section 4.4); 2) the sample is spatially well-distributed, sometimes also called spatially balanced; 3) accuracy estimates have small standard errors; 4) an unbiased estimator of variance is available rather than having to use an approximate variance; 5) the sampling design is easy to implement and the sample data easy to analyze; 6) the design is cost effective; and 7) the sample size can be increased or decreased even when data collection is in progress. These criteria will differ in importance for a given application making it necessary to decide which criteria have greater priority. With the objectives and prioritized desirable design criteria specified, the starting point for choosing a sampling design focuses on three main decisions: 1) Are the pixels sampled individually or sampled as clusters? 2) Are the pixels assigned to strata? and 3) What is the protocol used to select the sample pixels or clusters, simple random, systematic, or something else? Regarding the first two questions, a helpful mnemonic is, “Stratify for objectives, cluster for cost.” The issues determining the answers to these three questions are discussed in the following three sections.

5.1 Cluster Sampling

In the terminology of sampling, a cluster is a group of spatially contiguous pixels such as a 5x5 cluster of pixels or a 10km x 10km cluster of pixels. The starting point of cluster sampling is to partition the ROI into clusters. A sample of clusters is then selected, and all pixels in that cluster become part of the sample (one-stage cluster sampling), or a sample of pixels is selected from within each sampled cluster (two-stage cluster sampling).

Cluster sampling is almost always motivated by the need to constrain the sample spatially to reduce the cost of collecting the sample data. For example, the budget may allow for purchase of a limited number of very high resolution images or travel to a small number of field locations. Grouping the sampled pixels into a small number of clusters achieves the desired spatial constraint on the sample and the intended cost savings. A disadvantage of

cluster sampling is that it may result in larger standard errors than a sampling design of equivalent sample size in which the pixels are selected individually rather than in clusters. Typically pixels within each cluster are relatively homogeneous and therefore contain less information than pixels selected individually (i.e., not within clusters), and this translates to larger standard errors of the cluster sampling accuracy estimates. Consequently, the cost savings of cluster sampling must be such that a larger sample size of pixels can be obtained to compensate for the increased standard errors associated with cluster sampling. Standard error formulas for cluster sampling are more complex (Stehman 1997) than they are for other sampling designs.

5.2 Stratified Random Sampling

In stratified random sampling, pixels are grouped into strata and a simple random sample of pixels is selected from each stratum. Stratified sampling is most commonly implemented with the map classes defined as strata to achieve precise estimates of user's accuracy. Defining each map class as a stratum allows control of the sample size per stratum, so rare map classes can be allocated a sample size that is large enough to obtain a precise estimate of user's accuracy. Thus, the "stratify for objectives" rationale applies directly to estimating user's accuracy. Implementing stratified sampling to improve precision of estimated producer's accuracy is problematic because stratified sampling would require assigning each pixel in the ROI to a stratum based on the pixel's reference class. However, this is rarely feasible because the reference class is known only for the sample pixels, not for the ROI.

Stratification can also be used when the objectives require precise estimates of accuracy by subregion. Each subregion would be defined as a stratum, and the sample size allocated to each subregion specified to obtain the desired precision. Stratification by both map class and subregion is also possible in which case the map strata are constructed within each subregional stratum. Stratifying by map class within subregions may result in numerous strata and it is possible the overall sample size may not be large enough to allow adequate sample sizes in all strata. This would indicate that the budget for accuracy assessment is not sufficient to support all objectives of the project.

Strata used in a stratified sampling design become a permanent feature of the data and must be accounted for in the formulas used to estimate accuracy. For example, if the strata are defined by the map classes of one product and the reference sample is used to assess accuracy of a different product, each stratum will include pixels whose map class no longer matches the class associated with the stratum. The sample data obtained from a design using strata from one map can still be used to assess accuracy of a different map, but different formulas from those conventionally presented for stratified sampling must be used (Stehman 2014). While the choice of strata impacts precision of the accuracy estimators, the choice of strata will not result in biased estimators of accuracy when the appropriate stratified estimation formulas are used.

5.3 Simple Random vs. Systematic Sampling

Stratified sampling and cluster sampling are incomplete descriptions of a sampling design because it is also necessary to specify the protocol used to select the pixels from each stratum or the protocol used to select the clusters. The two most commonly used selection protocols are simple random and systematic. For simple random sampling with a sample

size of n units (e.g., pixels or clusters), all units have an equal probability of being included in the sample, and all pairs of units have an equal probability of jointly being included in the sample. Systematic sampling should be implemented with a randomized start, for example by selecting an initial sample pixel from the ROI with all pixels having an equal chance of being selected. Subsequent sample pixels are selected at fixed distances from the random starting pixel to create a regular grid of sample pixels. Similar to simple random sampling, all pixels have an equal probability of being included in the sample, but systematic sampling prevents many pairs of pixels (e.g., adjacent pixels) from ever being selected jointly in the same sample.

In terms of the desirable design criteria, systematic sampling yields a more spatially balanced distribution of the sample pixels that usually translates to more precise estimates of accuracy compared to simple random sampling. But a disadvantage of systematic sampling is that variances of the accuracy estimates must be approximated. Typically, simple random sampling variance estimator formulas are used, and this generally results in overestimating variance of the systematic sampling accuracy estimates. Simple random sampling has the advantage relative to systematic sampling of greater ease of changing sample size once the sample data collection has commenced. Both simple random and systematic sampling allow unbiased estimators of accuracy, but only simple random allows unbiased estimation of variance. Simple random sampling is preferable to systematic sampling if it is likely that the sample size may need to change after data collection has begun.

5.4 Probability Sampling

A critical element of a statistically rigorous accuracy assessment is that the sampling design must satisfy the conditions that define a probability sample. These conditions are that the probability of including a pixel in the sample (i.e., “inclusion probability”) must be known for the sampled pixels and the inclusion probability must be greater than zero for all pixels in the ROI. Probability sampling designs are necessary to invoke design-based inference, which is the form of statistical inference commonly applied in accuracy assessment (Stehman 2000). Use of a probability sampling design also ensures that the sample is representative of the ROI because of the known inclusion probabilities.

Precision of an accuracy estimate is quantified by its standard error. Small standard errors are therefore desirable, and the magnitude of the standard error is determined by the sampling design and sample size. A common misunderstanding is that the relative sample size (i.e., percent of the ROI sampled) is what determines precision, but in actuality it is the total sample size that determines the standard error. Consequently, even though the ROI may contain millions of pixels, precise and representative estimates of accuracy can be obtained from a probability sample that comprises an extremely small percent of the pixels in the ROI.

6. Future Developments

As satellite imagery continues to improve in temporal and spatial resolution, methods of accuracy assessment will evolve in tandem to address the new generation of map products. Monitoring change over time such as change in forest cover or change in surface water have led to response design protocols tailored to collect time series of reference data

(Cohen et al. 2010). To reduce the cost of reference data collection, volunteered geographic information (VGI) offers great potential, but use of such data must follow strict guidelines if the accuracy assessment is to be statistically rigorous (Stehman et al. 2018). Object-based accuracy assessments in which the assessment unit is a polygon are becoming more common, and specialized methods continue to be developed for these applications (Radoux et al. 2011; Ye et al. 2018). Lastly, maps of continuous variables such as biomass and percent impervious surface are becoming increasingly common and methods to describe accuracy of such products are needed (Riemann et al. 2010).

References

Brus, D. J., Kempen, B., and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62, 394-407.

Cohen, W. B., Yang, Z., and Kennedy, R. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync - Tools for calibration and validation. Remote Sensing of Environment, 114, 2911-2924.

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

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

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

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment, 148, 42-57.

Pontius, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing, 32(15), 4407-4429.

Pontius, R. G., Jr. (2019). Component intensities to relate difference by category with difference overall. International Journal of Applied Earth Observation Geoinformation, 77, 94-99.

Pontius, R. G., Jr., and Santacruz, A. (2014). Quantity, exchange, and shift components of difference in a square contingency table. International Journal of Remote Sensing, 35, 7543-7554.

Radoux, J., Bogaert, P., Fasbender, D., and Defourny, P. (2011). Thematic accuracy assessment of geographic object-based image classification. International Journal of Geographical Information Science, 25, 895-911.

Riemann, R., Wilson, B. T., Lister, A., and Parks, S. (2010). An effective assessment protocol

for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. Remote Sensing of Environment, 114, 2337-2352.

Stehman, S. V. (1997). Estimating standard errors of accuracy assessment statistics under cluster sampling. Remote Sensing of Environment, 60, 258-269.

Stehman, S. V. (2000). Practical implications of design-based sampling inference for thematic map accuracy assessment. Remote Sensing of Environment, 72, 35-45.

Stehman, S. V. (2014). Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. International Journal of Remote Sensing, 35, 4923-4939.

Stehman, S. V., and Czaplewski, R. L. (1998). Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. Remote Sensing of Environment, 64, 331-344.

Stehman, S. V., Fonte, C. C., Foody, G. M., and See, L. (2018). Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover. Remote Sensing of Environment, 212, 47-59.

Wickham, J. D., Stehman, S. V., Gass, L., Dewitz, J., Fry, J. A., and Wade, T. G. (2013). Accuracy assessment of NLCD 2006 land cover and impervious surface. Remote Sensing of Environment, 130, 294-304.

Ye, S., Pontius, R.G., Jr., and Rakshit, R. (2018). A review of accuracy assessment for object-based image analysis: From per-pixel to per-polygon approaches. ISPRS Journal of Photogrammetry and Remote Sensing, 141, 137-147.