R. Michael Hord*
William Brooner
Earth Satellite Corporation
Bethesda, MD 20015

Land-Use Map Accuracy Criteria

A statistical sampling procedure for determining the accuracy of classification, boundary line placement, and control point placement is described.

Introduction

Remote sensing particularly medium-and high-altitude aerial photography, is being used increasingly for the preparation of land-use inventory maps by a variety of users. Until recently, this increase of use, representing an important advance for the application of aerial photographic interpretation, has been gradual, due in part to the nature of technique and to product acceptance in user organizations.

Abstract: *In generating maps to meet prescribed accuracy standards from remote sensing data, validation is required. Cost considerations dictate that a sampling strategy for field checking be employed. The use of a sampling procedure to assess the accuracy of polygon classification, boundary line placement, and control point placement on land-use maps is described. Combining these measures into a single figure of merit with a confidence interval is suggested to foster acceptance in the user community.*

Some organizations have been using aerial photography for planning and land-use applications for several decades. Their applications have generally involved large-scale black-and-white photography, which has often served only as a base for plotting ground survey data, rather than exploiting the photography as an information source. Such users have not taken advantage of the inherent benefits of photo-interpretation, such as consistency of data, timeliness, or accuracy, nor of color, color infrared, multiscale, or other remote sensing techniques conventionally available today. Ground surveys, however, almost always lead to inconsistencies in detail over space as well as to extensive requirements for collection time and resources.

Acceleration of the acceptance of photo derived maps, however, requires addressing issues for which the remote sensing technologist is often unprepared. Included, for example are such issues as cost/benefit in relation to traditional approaches, and accuracy of results. For the sake of discussion, it is asserted that the costs of data acquisition and analysis have been demonstrated to be markedly lower with aerial photo interpretation than with ground surveys in all but the largest scale and most detailed site specific cases. Moreover, costs reduce as the scale of the data is reduced. Given the cost advantage, the discussion issue then becomes one of whether accuracy is sufficient when medium-and high-altitude photography is the principal data source. The remainder of this discussion focuses on the issue of quantifying map accuracy.

Accuracy of land-use interpretation is a complex issue, both in its definition and measurement. For example, an area delineated and classified as a particular category may be in error for one or more of three reasons: (1) classification error, (2) boundary line error, and (3) control point location error. Their interrelationship will be address-

* Mr. Hord is presently with the Institute for Advanced Computation, Alexandria, Va. 22314.

ed later. Each factor is of varying importance to the user, whose specification requirements differ, and each impacts the utility of the resulting products.

CLASSIFICATION ACCURACY ESTIMATION

Hierarchical classification schemes such as those used by the U. S. Geological Survey further complicate the first type of error. For example, two levels of detail are indicated, such as ,"11" which represents "urban residential land use." The first digit (Level I) is urban and built-up areas and the second digit (Level II) is residential areas. It can be seen, therefore, that an area which should be called 11 but is called 12 is in error at Level II; but is correct at Level I. In defining a classification accuracy estimate, the classification must either be correct at all levels reported, or the level of classification must be stated for the associated accuracy estimate (i.e., x per cent at Level I, y per cent at Level II, etc.). We assert that a single accuracy estimate will have the greatest unambiguous meaning to the majority of map users; their concern for accuracy implies concern for quality and is often not supported by an understanding of the complexities which arise in attempting to quantify the accuracy of land-use mapping. Therefore, a single value must refer to the lowest classification denominator: *for a classification to be correct it must be correct at all levels reported.*

A map is a graphic interpretation and the presentation of a complex surface that often contains abstractions. Without field checking the total map, exact accuracy cannot be verified; hence a sampling procedure must be employed to estimate classification accuracy.

Any sampling procedure to be employed must involve field work and must be statistically valid. It is important that users understand that any accuracy estimate based on sampling requires confidence intervals which are dependant on the number of sample points selected per map.

The procedure below is suitable, as an example, for estimating accuracy for one-acre sample points on land-use maps at 1:24,000 scale and covering 7-1/2 minute areas (approximately 56 square miles). The procedure involves the random selection of one-acre points with replacement, ground checking these points, and comparing the field observer's classification to that of the aerial photo-interpreter. The estimated accuracy of classification may then be computed within specified confidence intervals.

At each level of classification (e.g., Level I, Level II, etc.) it is assumed that there is a "true" category for each acre (or other land surface unit) on the map. (This assumption can, of course, be challenged. Land-cover/land-use phenomena are sometimes ambiguous to discrete classification, even for ground-based observers.) Each acre's assigned category is correct or incorrect (1 or 0) and the set of both is the population of concern. The mean, μ, of this population equals the sum of the population elements divided by the number, n, of these elements i.e,

$$\mu = (1/n) \sum_{i=1}^{n} x_i . \qquad (1)$$

μ is also p, the probability that any given acre has a correctly assigned classification. Since this is a binomial distribution, the variance is—

$$\sigma^2 = p\,(1 - p) = \mu\,(1 - \mu) . \qquad (2)$$

The confidence interval for μ can be calculated from the approximation[1]

$$P\,\left(- b < \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{N}}\right) < b\right) = 1 - \alpha \qquad (3)$$

where N is the item count of the sample, $\bar{x}$ is the sample mean, and 100 $(1 - \alpha)$ per cent is the confidence level of the interval. For a 95 per cent confidence interval, $\alpha = 0.05$ and, from the normal distribution tables, $b = 1.960$; hence

$$\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{N}}\right)^2 < (1.96)^2 \text{ ie } N\,(\bar{x}^2 - 2\bar{x}\,\mu + \mu^2)$$

$$< (1.96)^2\,\mu\,(1 - \mu) .$$

This is solved for upper and lower limits of μ.

Example 1:

If $N = 300$
and $\bar{x} = 0.98$,
then $0.9570 < \mu < 0.9908$.

That is, the true map accuracy is, with 95 per cent confidence, in the range 0.9570 to 0.9908. The sample accuracy is 0.98 from a sample check of 300 points.

Example 2:

If $N = 300$
and $\bar{x} = 0.96$,
then $0.9314 < \mu < 0.9770$.

Table 1 shows the map accuracy upper and lower 95 per cent confidence limits as a function of the number of samples and the accuracy value for these samples.

TABLE 1. MAP ACCURACY UPPER AND LOWER 95 PER CENT CONFIDENCE LIMITS (μ) AS A FUNCTION OF THE NUMBER OF SAMPLES (N) AND THE ACCURACY VALUES ($\bar{x}$) FOR THESE SAMPLES.

N	$\bar{x}$ (%)	μ (lower limit)	μ (upper limit)	N	$\bar{x}$ (%)	μ (lower limit)	μ (upper limit)
50	80	0.6896	0.8876	250	80	0.7461	0.8449
50	81	0.6808	0.8950	250	81	0.7568	0.8538
50	82	0.6920	0.9023	250	82	0.7677	0.8627
50	83	0.7034	0.9095	250	83	0.7785	0.8715
50	84	0.7149	0.9166	250	84	0.7895	0.8802
50	85	0.7624	0.9236	250	85	0.8005	0.8889
50	86	0.7381	0.9305	250	86	0.8115	0.8976
50	87	0.7500	0.9372	250	87	0.8227	0.9061
50	88	0.7620	0.9438	250	88	0.8339	0.9148
50	89	0.7741	0.9503	250	89	0.8452	0.9230
50	90	0.7864	0.9565	250	90	0.8566	0.9313
50	91	0.7989	0.9626	250	91	0.8681	0.9395
50	92	0.8116	0.9684	250	92	0.8797	0.9476
50	93	0.8246	0.9741	250	93	0.8914	0.9555
50	94	0.8378	0.9794	250	94	0.9034	0.9633
50	95	0.3514	0.9844	250	95	0.9155	0.9708
50	96	0.8654	0.9890	250	96	0.9280	0.9781
50	97	0.8799	0.9930	250	97	0.9407	0.9850
50	98	0.8950	0.9965	250	98	0.9541	0.9914
50	99	0.9111	0.9990	250	99	0.9683	0.9969
50	100	0.9287	1.0000	250	100	0.9849	1.0000
100	80	0.7112	0.8666	300	80	0.7511	0.8413
100	81	0.7222	0.8749	300	81	0.7618	0.8504
100	82	0.7333	0.8830	300	82	0.7726	0.8593
100	83	0.7445	0.8911	300	83	0.7834	0.8683
100	84	0.7558	0.8990	300	84	0.7943	0.8771
100	85	0.7672	0.9069	300	85	0.8052	0.8860
100	86	0.7786	0.9147	300	86	0.8162	0.8947
100	87	0.7902	0.9224	300	87	0.8272	0.9034
100	88	0.8019	0.9300	300	88	0.8383	0.9120
100	89	0.8137	0.9375	300	89	0.8495	0.9206
100	90	0.8256	0.9448	300	90	0.8608	0.9291
100	91	0.8377	0.9519	300	91	0.8722	0.9374
100	92	0.8500	0.9589	300	92	0.8837	0.9457
100	93	0.8625	0.9657	300	93	0.8954	0.9538
100	94	0.8752	0.9722	300	94	0.9072	0.9617
100	95	0.8882	0.9785	300	95	0.9192	0.9695
100	96	0.9016	0.9843	300	96	0.9314	0.9770
100	97	0.9155	0.9897	300	97	0.9440	0.9842
100	98	0.9300	0.9945	300	98	0.9571	0.9908
100	99	0.9455	0.9982	300	99	0.9710	0.9966
100	100	0.9630	1.0000	300	100	0.9874	1.0000
150	80	0.7289	0.8562	350	80	0.7549	0.8385
150	81	0.7398	0.8647	350	81	0.7656	0.8476
150	82	0.7508	0.8732	350	82	0.7763	0.8567
150	83	0.7618	0.8817	350	83	0.7871	0.8657
150	84	0.7730	0.8901	350	84	0.7979	0.8747
150	85	0.7842	0.8984	350	85	0.8088	0.8836
150	86	0.7955	0.9066	350	86	0.8197	0.8925
150	87	0.8068	0.9147	350	87	0.8307	0.9013
150	88	0.8183	0.9227	350	88	0.8418	0.9100
150	89	0.8299	0.9306	350	89	0.8529	0.9186
150	90	0.8416	0.9384	350	90	0.8641	0.9272
150	91	0.8534	0.9461	350	91	0.8754	0.9357
150	92	0.8654	0.9536	350	92	0.8868	0.9441
150	93	0.8776	0.9610	350	93	0.8983	0.9523
150	94	0.8899	0.9681	350	94	0.9100	0.9604
150	95	0.9026	0.9750	350	95	0.9219	0.9683

(Continued on next page)

TABLE 1. continued

N	$\bar{x}$ (%)	μ (lower limit)	μ (upper limit)	N	$\bar{x}$ (%)	μ (lower limit)	μ (upper limit)
150	96	0.9155	0.9815	350	96	0.9340	0.9760
150	97	0.9289	0.9876	350	97	0.9464	0.9834
150	98	0.9429	0.9931	350	98	0.9593	0.9903
150	99	0.9579	0.9976	350	99	0.9730	0.9963
150	100	0.9750	1.0000	350	100	0.9891	1.0000
200	80	0.7391	0.8495	400	80	0.7580	0.8363
200	81	0.7500	0.8583	400	81	0.7687	0.8454
200	82	0.7609	0.8670	400	82	0.7794	0.8546
200	83	0.7718	0.8757	400	83	0.7901	0.8636
200	84	0.7829	0.8843	400	84	0.8009	0.8727
200	85	0.7940	0.8929	400	85	0.8117	0.8817
200	86	0.8051	0.9013	400	86	0.8226	0.8906
200	87	0.8163	0.9097	400	87	0.8335	0.8995
200	88	0.8277	0.9180	400	88	0.8445	0.9083
200	89	0.8391	0.9262	400	89	0.8555	0.9170
200	90	0.8506	0.9343	400	90	0.8667	0.9257
200	91	0.8622	0.9423	400	91	0.8779	0.9343
200	92	0.8740	0.9501	400	92	0.8892	0.9428
200	93	0.8860	0.9578	400	93	0.9007	0.9511
200	94	0.8981	0.9653	400	94	0.9123	0.9593
200	95	0.9104	0.9726	400	95	0.9240	0.9674
200	96	0.9231	0.9796	400	96	0.9360	0.9752
200	97	0.9361	0.9862	400	97	0.9483	0.9828
200	98	0.9497	0.9922	400	98	0.9610	0.9898
200	99	0.9643	0.9972	400	99	0.9745	0.9961
200	100	0.9812	1.0000	400	100	0.9905	1.0000

To use this table, the user first determines the classification accuracy level to be achieved, such as 80 per cent. Then the number of samples to be checked are determined, such as 150 points randomly selected with replacement. The table then shows by the lower 95 per cent confidence limit that the minimum number of correct points required from the sample to achieve the specified 80 per cent accuracy within the 95 per cent confidence interval is, in this example, 131 correct points, so that $\bar{x}$ will be 87 per cent.

Table 1 also may be used to estimate classification accuracy. For example, given a sample of 250 points, of which 235 are found to be correct, then since $\bar{x}$ is 94 per cent the estimated lower limit of accuracy is 90 per cent with 95 per cent confidence. Again with 95 per cent confidence, the upper limit of accuracy is 96 per cent.

Another useful quantitative descriptor for land-use maps is the spatial complexity index defined as the average polygon ground area divided into the total area. Particularly as the digital multispectral-classification-generated land-use maps find greater acceptance and hence the occurrence of one-picture-element polygons increases, the spatial complexity index, which quantifies the spatial homogeneity of the map, has become an increasingly used concept. For example, on a map covering 36,000 acres with an average polygon area of 40 acres, the spatial complexity index would be expressed as 900.

BOUNDARY ERRORS

The second type of error, boundary lines, involves both locational tolerances and line widths relative to the scale of the final map. How wide is the map line in terms of ground distance? Line width data for standard KOH-I-NOOR pen point sizes is given in Table 2. Depending on the map scale, ground distances can be considerable even when within tolerance, a fact of no surprise to those familiar with cartographic representation.

Today, with the growing access to computers and to scanning densitometers, digital cartography is finding greater acceptance. In this context maps are often represented in raster format. Each point of a map is coded as a one or zero depending on whether it is or is not an element of a line. (For optimal utility the raster point diameter should approximate the line width.) If these points are spaced ten mils between centers, the array constituting the digital codification of a 20-inch by 30-inch

TABLE 2. GROUND LINE WIDTHS AT VARIOUS MAP SCALES FOR STANDARD KOH-I-NOOR PEN POINT SIZES.

		Ground Distance (feet)				
Koh-I-Noor Line	Actual Width (in)	1:24,000	1:63,360	1:100,000	1:250,000	1:500,000
6 × 0	0.005	10	26.4	41.67	104.0	208.0
5 × 0	0.0075	15	39.6	62.5	156.0	312.0
4 × 0	0.012	24	63.4	100.0	250.0	500.0
3 × 0	0.014	28	74.0	116.7	291.7	583.3
00	0.016	32	84.5	113.3	333.3	666.7
0	0.017	34	89.8	141.7	354.2	708.3
1	0.021	42	110.9	175.0	437.5	875.0
2	0.023	46	121.4	191.7	479.3	958.5
2½	0.028	56	147.8	233.3	583.3	1,166.7
3	0.037	74	195.4	308.3	770.8	1,541.7
4	0.052	104	274.6	433.3	1,083.3	2,166.7
6	0.067	134	353.7	558.3	1,395.7	2,791.4
7	0.068	136	359.1	566.7	1,416.7	2,833.3
8	0.09	180	475.2	750.0	1,875.0	3,750.0
9	0.10	200	528.0	833.3	2,083.3	4,166.7
10	0.15	350	792.0	1,249.9	3,125.0	6,250.0

map would consist of 2000 lines of 3000 points each a total of six million points.

A figure of merit is available[2] for specifying the degree of congruence between two such raster maps. Each point on map one and the corresponding point on map two is inspected to determine whether each is or is not an element of a line. Four totals are calculated: A, the total number of line points on map one for which the corresponding point on map two is also a line point; B, the total number of line points on map one for which the corresponding point on map two is not a line point; C, the total number of nonline points on map one for which the corresponding point on map two is a line point; and D, the total number of nonline points on map one for which the corresponding point on map two is not a line point.

Then
$$\begin{aligned} S_1 &= A + C \\ S_2 &= B + D \\ S_3 &= A + B \\ S_4 &= C + D \\ R &= A + B + C + D \\ E_A &= (S_1 S_3) / R \\ E_B &= (S_2 S_3) / R \\ E_C &= (S_1 S_4) / R \\ E_D &= (S_2 S_4) / R \\ X &= (A - E_A)^2 / E_A \\ &\quad + (B - E_B)^2 / E_B \\ &\quad + (C - E_C)^2 / E_C \\ &\quad + (D - E_D)^2 / E_D \end{aligned}$$

The larger X is, the better is the agreement between the two maps. For example, three maps in raster format each measuring 1,000 by 1,000 points are available. Which two are in better agreement if, in comparing maps one and two,

$A = 200{,}000, B = 500{,}000, C = 200{,}000$, and $D = 100{,}000$; in comparing maps two and three,

$A = 600{,}000$, $B = 100{,}000$, $C = 100{,}000$, and $D = 200{,}000$; and in comparing maps one and three,

$A = 300{,}000, B = 200{,}000, C = 300{,}000$, and $D = 200{,}000$?

The three values of X are—127,000; 274,000; and 100,000.

Hence, the best agreement is between maps two and three while the worst agreement is between maps one and three.

An alternative measure is $(A + D) / R$, a binomial variable comparable to $\bar{x}$ of the previous section.

CONTROL POINT LOCATION ACCURACY

Control point accuracy deals with the absolute geometric relationship of the map with respect to a universal frame of reference such as latitude and longitude, whereas boundary errors are associated with relative geometric fidelity.

For this discussion we will use a broad definition of the term control point: any identifiable map representation of a landmark. By this we mean to exclude points in the middle of homogeneous areas but include field corners, buildings, and river and road intersections, i.e., anything whose true ground position can be surveyed. The U.S. National Map Accuracy Standards state that for a map to be termed accurate, "for maps on publication scales larger than 1:20,000 not more than 10 per cent of the points tested shall be in error more than 1/30 inch, measured on the publication scale; for maps on publication scales of 1:20,000 or smaller, 1/50 inch. These limits of accuracy shall apply in all cases to well defined points only." Since we are dealing here with land-use maps that do not carry elevation information, vertical accuracy does not concern us. Hence for 1:100,000 scale maps 90 per cent of all ground locations are claimed to be accurate to within 167 feet, whereas 1:1,000,000 scale maps this ground location tolerance is 1667 feet. How many points need a cartographer check to have 95 per cent confidence that the map meets this standard? We may again use the binomial distribution here since each of the n control points on the map is either accurately placed or it is not. We want 90 per cent of all the points to be accurately placed and this will be so if the binomial probability p of correct placement is 0.9. Let S be the number of correct placements in m trials, i.e., after checking m map points on the ground. The expected value of S is mp and the variance of S is $mp\,(1\text{-}p)$.

The variable

$$y = \frac{S - mp}{\sqrt{p\,(1 - p)}} \tag{5}$$

is approximately normally distributed with zero mean and unit standard deviation if m is large.

$$P\left[\,-a < y < a\right] = 0.95 \text{ for } a = 1.96.$$

We may assess both limits on p by solving $y^2 = a^2$ for p; i.e.,

$$\frac{S^2 - 2mpS + m^2p^2}{p\,(1 - p)} = a^2. \tag{6}$$

Example 3: $m = 5$ and $S = 5$, i.e., we have ground checked five randomly selected points and each is correctly placed. The 95 per cent confidence interval for p is

$$0.866 < p < 1.00\cdot$$

That is, upon checking five points and finding them correct, you may be 95 per cent confident that the probability of correct placement for any point is greater than 0.866.

Example 4: If S = 25 and m = 25, then

$$0.994 < p < 1.00\ .$$

Here we may say with 95 per cent confidence that the map is at least 99.4 per cent accurate in control point placement.

Other Considerations

It is awkward to quote three values to specify map accuracy. Several composite figures of merit may warrant consideration. The authors suggest adoption of the root-mean-square (RMS) of the lower 95 per cent confidence limit of the three per cent measures presented earlier. Hence if the polygon classification accuracy is at least 90 per cent, the boundary accuracy is at least 95 per cent, and the control point correctness is at least 97 per cent, each at 95 per cent confidence, then the RMS accuracy of these is 94 per cent.

Obviously there is reason for dissatisfaction with these accuracy measures. One reason is that these three types of error are not independent. If a boundary is in error, then the adjacent polygons will be affected by having multiple land-use categories in some polygons as one example of this interdependency. Another is that the field observations are taken as the standard for quality control without recognizing that ground observations are frequently in error. Despite these drawbacks and others, the authors propose these evaluation methods with the realization that no figure of merit is ever entirely appropriate.

Earlier reference was made to the cost benefit advantage of maps generated from remotely sensed data. This can be valid only so long as the amount of ground checking for accuracy assessment is kept reasonable. In this regard the user is advised not to demand accuracy in excess of his real requirement. If the particular application for which a given map is generated need only be 80 per cent accurate in the above RMS sense, then specifying 98 per cent accuracy is not good practice.

Finally, it is noted that the time dimension has not been addressed in this discussion. How do these error measures change with age? Is it preferable to use a map with 80 per cent rms accuracy generated from imagery acquired two months ago or a map with 90 per cent RMS accuracy generated from year old imagery?

These and similar questions may be as

complex as the issue of accuracy measurement itself and generalizations may be hazardous in their simplicity. Land-use change *per se* is the issue here, and an aerial photograph may be regarded as an historical document which captures the landscape scene at only a single moment. The most suitable date assigned to a given land-use map is the date of photo acquisition rather than photo-interpretation unless field surveys can confidently update all areas of land-use change.

The dynamics of land-use change need examination; often the rate of change in seemingly dynamic areas is relatively low when considered in a county or regional context. From ground observations we can note, perhaps with alarm, the expansion of urban areas into adjacent hinterlands. Within central urban areas, the replacement process may be very visible to the local observer. But considered in the context of scale, new subdivisions may consume only relatively small amounts of available land per year, and many replacements may be new structures of the same previous land use. Such situations tend to reduce the necessity of two-month-old imagery given existing imagery of comparable quality that may be a year old.

To the contrary, field surveys for interpretation verification are complicated in areas of land-use change or conversion subsequent to the time of photo acquisition. What may have been an accurate interpretation of the photograph is not confirmed by field survey. The observation may lead to a recorded error in the classification accuracy estimation, penalizing the interpreter who performed his task properly and well, and reducing (erroneously) the quantified quality of the map product.

Conclusion

To foster the acceptance of land-use mapping from remote sensing, the cartographer must be able to specify the accuracy of his product. The three types of inaccuracy (area misclassification, boundary line error, and control point location error) have been addressed and a procedure for quantitatively specifying each type of error has been described.

The topic is by no means closed. For example, studies to ascertain the optimal accuracy requirement specification for various applications would examine the tradeoffs in the cost of field checking versus the cost of inadvertently accepting a map error.

Since the procedures are statistical in nature, some will mistrust them. Certainly counter examples can be produced, e.g., cases exhibiting good values for these figures of merit while the maps are obviously bad, at least aesthetically. Generally the cartographer's clients realize this. Maps characterized by good values for these figures of merit, even those generated from remote sensing data, should find ready acceptance. In many cases maps generated by using traditional methods may not score as well.

References

1. Brunk, H. D., *An Introduction to Mathematical Statistics,* Blaisdell, 1965, p. 177ff.
2. Hord, R. M., *Comparison of Edge and Boundary Detection Techniques,* AIPR Symposium, December 1971, Washington, D.C.