ALAN M. HAY
Department of Geography
University of Sheffield
Western Bank, Sheffield, S10 2TN, England

Sampling Designs to Test Land-Use Map Accuracy

Criteria for sample size are established and a stratified sampling design is described.

THERE ARE MANY research areas in which it is required that predictions of characteristics derived from other sources be tested by carrying out field observations. The most common example is the need to test interpretations of land use, vegetation type, or soil type which have been made on the basis of remotely sensed imagery (air photography, satellite photography, etc.). Although field data may have been used in the construction of an interpretation "key," there is still a need for a *post facto* exercise to determine the accuracy or frequency of error to which the interpretation is prone.

ABSTRACT: *In testing the accuracy of qualitative characteristics determined from remotely sensed data, five problems arise:*

I. What proportions of all decisions are correct?
II. What proportion of the allocation to a category is correct?
III. What proportion of the true category is correctly allocated?
IV. Is a category overestimated or underestimated?
V. Are the errors randomly distributed?

To tackle these questions it is necessary to determine sample size (always >50) and to adopt a stratified sampling design. The questions can then be answered using tabulated values for the binomial errors (Questions I - IV) and Poisson frequencies (Question V).

Similarly, there are circumstances in which easily collected diagnostic variables are used to predict other less easily observed characteristics. For example, field and air photo observations of aspect, slope, lithology, and vegetation might be used to "predict" soil type. Once again, the prediction method may have been based upon field data but its reliability as a method can only be ascertained by a *post facto* test using independently sampled field observations. In most cases, therefore, it is required to derive and analyze a table of the form in Table 1.

In such a table a number of questions are answered:

I. What proportion of all the sample predictions proved to be correct (incorrect)?

II. What proportion of the sample predictions of a single category proved to be correct (incorrect)?

III. What proportion of land truly (in the ground truth sense) in a category is correctly predicted?

IV. Is the net effect of II and III for predictions to overestimate or underestimate a given category?

V. If error occurs in either of the ways II and III is there any bias in these errors towards specific categories?

This problem may arise in a multi-category case where some categories are acknowledged to be very similar: In such a case, mis-classification between similar categories may be high although overall accuracy is quite high. This effect appears in Table 1 where many of the errors arise from an apparent confusion between E and F.

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 45, No. 4, April 1979, pp. 529-533.

TABLE 1. A FIELD TEST DATA TABLE

		Predicted characteristic:						
		A	B	C	D	E	F	
Characteristics identified in field testing	A	194	6	3	2	—	—	205
	B	3	80	1	1	—	—	85
	C	2	7	180	—	—	—	189
	D		5	5	197	—	—	207
	E	1	1	—	—	180	15	197
	F		1	1	—	15	65	82
		200	100	190	200	195	80	965

All these questions can be answered with complete confidence if the study is a total enumeration or a very large sample. But total enumeration or very large samples involve the very heavy burden of field observation which the prediction technique is presumably designed to avoid. The method must therefore focus upon the extent to which Questions I-V can be answered by recourse to sample data sets.

It should be noted that in some cases there may be sources of error other than the "prediction" system. For example, in certain types of remotely sensed imagery the matching of sites on the imagery with exact locations in the field is itself subject to error, which may then lead to identification of an apparently incorrect prediction. Similarly, a time interval between prediction and field survey may result in changes which are recorded as errors of prediction. No attempt is made to estimate such errors in the procedures described below.

SAMPLE SIZE

The question of sample size can be introduced with a simple example. Suppose that only ten sample points are checked and that the "results" indicate that all ten determinations were correct. The immediate reaction, which is quite common in some circles (Lins, 1976), is to assume that the method is 100 percent correct. However, sampling theory tells us that where there are ten trials the probability of all ten being correct is the 10th power of the true proportion of correct determinations: these probabilities are given in Table 2. On the other hand the result, 9/10 suggesting 90 percent correct, might arise from a situation where the true proportion was much higher (99 percent) or much lower (85 percent).

These results are derived by using the terms of the binomial expansion. In order to establish necessary sample sizes, it is necessary to fix required confidence limits: In this discussion it is assumed that the 95 percent level will be acceptable, but that all the guidelines given would need to be recalculated if different confidence limits were to be set.

By using this approach, it is possible to establish the range, at 95 percent confidence limits, within which the true proportion of errors probably lies for any specified sample size and success rate. These are tabulated for specific sample sizes by Hord and Brooner (1976) or can be presented in graphical form as in Figure 1 (see also Arkin and Colton, 1973, or Hill *et al.*, 1961). In that figure the actual percent accuracy achieved in the sample can be related to lower and upper bounds for the range of the probable true accuracy. It is worthwhile to stress that the true value may be higher or lower than the sample value; for example, the sample value of 45/50 (90 percent) might at the 95 percent confidence limits imply a true population

TABLE 2. ALTERNATIVE INTERPRETATIONS OF RESULTS FROM A SMALL SAMPLE (n = 10)

(a) If true proportion correct:	(b) Probability 10/10	(c) Probability of 9/10	(d) Probability of 9/10 or better (b + c)
99	0.90	0.09	0.99
95	0.60	0.32	0.92
90	0.35	0.39	0.74
85	0.20	0.35	0.55

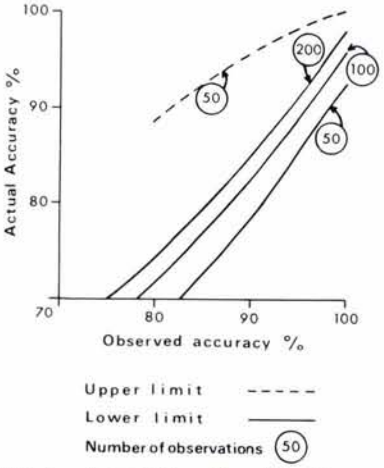

FIG. 1. A graph for estimation of accuracy from samples (source: Table 1, Hord and Broomer, 1976).

value as high as 95.7 percent or as low as 78.6 percent. The asymmetry of these graph values is also worthy of note. In tackling land-use sampling problems, some authors have attempted to use the standard error equation for binomial data, i.e.,

$$\mathrm{SE}\% = \sqrt{\frac{p\%\, q\%}{n}}\,.$$

For example, if the percentage correct ($p\%$) were 90 percent and n were 100, then

$$\mathrm{SE}\% = \sqrt{\frac{90.10}{100}} = 3\%$$

In such a case the estimates of p are supposed to be normally distributed with a mean of 90 percent and a standard deviation of 3 percent, i.e., at the 95 Confidence limits $p = 90 \pm 6\%$. (Comparable figures from the binomial expansion are 83-95%.) The key assumption in this method is that the errors will be *normally* (and therefore symmetrically) distributed. Although this assumption is acceptable when p and q are large and n is large (say 1000), when p or q is small (say less than 20 percent, and we hope that errors will indeed be less than 20 percent) the assumption is invalid. In conclusion, it is clear that any sample of less than 50 will be an unsatisfactory guide to true error rates, and in most cases minimum sample sizes of 50 or 100 are to be recommended.

SAMPLE DESIGN

The conclusion of the last section was that a minimum sample size of at least 50 would be necessary to test the accuracy of determinations. It must however be stressed that a sample of this size is necessary for *each* category or subcategory for which separate accuracy checks are needed. In any simple random areal sampling design this requirement will mean that the smallest category (in terms of area) determines the size of the total sample—yielding sample sizes for larger categories far in excess of the required number. An example of this effect appears in column (c) of Table 3. The best established solution to this problem is to sample each category separately using the categories already determined as the strata (column (f) in Table 3). This can however be combined with an overall sample in the following manner. Samples are randomly selected over the whole study area, the stratum in which each falls is identified and a running total for each stratum is maintained. Once any one of the strata (e.g., A) has a sufficient sample size, the *overall sample* is treated as complete. As sampling continues, further samples falling in the complete stratum are rejected, whereas those falling in other strata are retained until they too are filled. The resulting pattern will be similar to that shown in columns (d) and (e) of Table 3; however, due to sampling variation the proportions in column (d) will not always exactly reflect the proportions in column (b). It will be clear that such a method will not only yield a minimum sample for each category but will also yield an overall sample, *N*, which is big-

TABLE 3. A STRATIFIED SAMPLING DESIGN

(a)	(b)	(c)	(d)	(e)	(f)
Category	Proportion of study area %	Sample size in single random sample	Main Sample	Additional Sub-sample	Total Sub-sample
A	40	500	50		50
B	40	500	50		50
C	12	150	15	35	50
D	4	50	5	45	50
E	4	50	5	45	50
		1250	125		

ger than the required minimum. Its size can be estimated in advance as

$$N \bumpeq \frac{100}{S_{max}} \times n$$

where $S_{\max}$ is the percent of land in the largest stratum and
n is the minimum sample size for a stratum.

Because, in most cases, $S_{\max}$ is inversely related to the number of classes or categories, the greater the number of categories the greater will tend to be the value of N. Thus, estimates of overall accuracy will be more (usually much more) reliable for the whole area (Question I) than for individual categories.

In the discussion above it has been tacitly assumed that the categories are of equal interest. However, certain of the categories may be of great importance (for example, policy decisions will be made on the basis of the results) whereas other categories are of minimal interest. In such circumstances it may be desirable to increase the desired sample minimum (to 100, 200, etc.) for the important categories and to ignore other categories except to the extent that they appear in the overall random sample. In such a case the value of N will be given by the maximum value of the expression

$$N = \frac{100}{S_i} \times n_i$$

where S_i is the proportion in the ith stratum and n_i is the required sample size for that stratum.

Parenthetically it can be noted that in some studies it is important that accuracy estimates be given not only for *each* category but for sub-sections of the whole geographic area. Such an exercise may involve a spatially stratified sampling design to ensure that an adequate number (50 or 100?) of checks is made in each areal sub-section.

DATA ANALYSIS

TABLE 4. OVERALL PROPORTIONS OF ACCURACY: RANDOM SAMPLE

		A	B	C	D	E	
Ground Observations	A	48	—	—	—	—	48
	B	1	49	—	2	—	52
	C	1	—	13	—	1	15
	D	—	1	1	3	1	6
	E	—	—	1	—	3	4
		50	50	15	5	5	125

The result of such a sampling program will, therefore, be two tables analogous with Table 1. Using the same figures as in Table 3, we may inspect Tables 4 and 5. Results from Table 4 suggest an overall accuracy (Question I) of 116/125 (or 92.8 percent) suggesting a range of the true accuracy between 86 and 96 percent. In answer to Question III this table suggests that ground conditions A are correctly identified 48/48 (probable true range 93 to 100 percent) and ground conditions B are correctly identified 49/52, (94.2 percent; with a probable true range 89 to 99 percent). It becomes clear, however, that this overall sample is insufficiently large to answer Question III about categories C, D, and E. Although D appears to be badly identified 3/6 = 50 percent the sample size of six gives limits of 17 percent to 83 percent for the true accuracy at 95 percent confidence limits.

Question IV can be answered from Table 4, therefore suggesting that A tends to be overestimated and B tends to be underestimated (in each case by approximately 4 percent; 2/48 and 2/52). (There is also very flimsy evidence that D and E are misestimated by 1/6 and 1/4, respectively, but the small sample size makes these conclusions unreliable.) On the other hand, it is clear that, if the true frequency of B is indeed 52/125 (= 41.6 percent), the value of 50/125 is well within expected sampling error, and the incorrect sample predictions do not necessarily imply that there is a systematic underestimate of the overall proportions.

Question II can best be considered by using the samples from the strata, Table 5.

In Table 5 it is the columns which are directly interpretable (any effects in the rows are largely a consequence of the stratified sampling and should be treated with caution). The first three columns suggest that errors are low:

A.	48/50	Range	86-99%
B.	49/50	Range	89-100%
C.	47/50	Range	83-98%

TABLE 5. PREDICTION ACCURACY: STRATIFIED SAMPLE

	A	B	C	D	E
A	48	—	2	5	—
B	1	49	—	4	—
C	1	—	47	3	3
D	—	1	1	34	12
E	—	—	—	4	35
	50	50	50	50	50

The fourth column (D) suggests a major fault in the "predicting system." The sample success rate 34/50 could reflect a true success rate of as low as 53 percent and it is almost certainly no higher than 80 percent. Similar conclusions are reached about column E. There is, however, a distinction: In the case of D the incorrect predictions appear to be almost uniformly distributed among the other classes, whereas in the case of E there is only one major confusion, between E and D. The significance of these differences can be considered by using a simple probability test.

If the errors were distributed equally among all the incorrect categories with equal probability, the probability that any one cell has the value c is again given by the binomial function

$$f(c) = p^c q^{n-c} \frac{n!}{c!\,(n-c)!}$$

where in this case p is the probability of this cell being selected,

q is the probability of any other cell being selected,
n is the number of errors to be distributed, and
c is the number in the cell under consideration.

As usual in such studies it is necessary to estimate the probability that a given c or one even larger could have occurred under equiprobability. For larger examples the evaluation of the binomial function is tedious and use can be made of the Poisson approximation which is available in tabular form (3). In the case of column D in Table 5 there are 16 errors spread between four cells. The expected mean is therefore 4. Consulting the Poisson tables for a distribution with a mean of 4 gives

f(0)	f(1)	f(2)	f(3)	f(4)	f(5)	f(6)
0.018	0.073	0.147	0.195	0.195	0.156	0.104

. . . .

Summing the probabilities from f(0) to f(4) gives 0.628. Clearly, then, the probability of *five or more* is 0.372. Values as extreme as 5 arise frequently by chance and the value of 5 can be deemed not significant and we can conclude that an improved *positive identification* of D is needed. In column E, however, the mean is 3.5 and the high frequency is 12 (12 E characteristics are confused with D). The Poisson distribution gives a summed probability that the value lies between 0 and 11 as 0.999. Clearly, the probability that *12 or more* will appear in one cell is 0.001, so small that there is highly significant evidence that E is persistently misidentified as D. It is, therefore, in the *distinction between E and D* that the technique needs refinement.

CONCLUSIONS

The accuracy of determinations of qualitative characteristics should not only be interpreted in a probabilistic manner as established by Hord and Brooner (1976) but also should be based upon a correct sampling design. Basic to this design is that it should be stratified so that a minimum number of observations (50 is suggested in the paper) is performed for each characteristic requiring an accuracy estimate. If this design is adopted, it is possible not only to determine accuracy in general terms but also to identify underestimation, overestimation, and the presence of a significant frequency of misclassification between two categories by using simple tests based on the binomial distribution and its Poisson approximation.

ACKNOWLEDGMENTS

This problem was shown to me by Brian Lock, and useful comments upon drafts of the paper were made by Stan Gregory, Tom Welch and anonymous reviewers. Since this paper was submitted, J. L. van Genderen and B. F. Lock (1977) have published an example using a similar method (also developed by the present author).

REFERENCES

Arkin, H., and R. Colton. 1973. *Tables for Statisticians*, Barnes and Noble, New York.

Hill, H. P., J. L. Roth, and H. Arkin. 1962. *Sampling in Auditing*, Ronald Press, New York.

Hord, R. M., and W. Brooner. 1976 Land-Use Map Accuracy Criteria, *Photogrammetric Engineering and Remote Sensing*, Vol. 42, pp. 671-677.

Lins, H. F. 1976. Land-Use Mapping from Skylab S190B Photography, *Photogrammetric Engineering and Remote Sensing*, Vol. 42, pp. 301-307.

vanGenderen, J. L., and B. F. Lock. 1977. Testing Land-Use Map Accuracy, *Photogrammetric Engineering and Remote Sensing*, Vol. 43, pp. 1135-1137.

(Received March 26, 1977; revised and accepted January 9, 1979)