Remote Sensing: Statistical Testing of Thematic Map Accuracy

J.L. VAN GENDEREN[†]

Environment and Resources Consultancy, Fairey Surveys Ltd., Reform Road, Maidenhead, Berkshire SL6 8BU, England

B. F. LOCK

Department of Geography, Salisbury College of Advanced Education, Adelaide, Australia

P. A. VASS

Environment and Resources Consultancy, Fairey Surveys Ltd., Reform Road, Maidenhead, Berkshire SL6 8BU, England

In order to achieve wider acceptance among users of thematic maps derived from remote sensing data, the interpreter must be able to specify the accuracy of his product. This requires a valid sampling procedure to estimate classification accuracy. Although several alternative methods have been used in the past, none provide sufficient statistical justification for the allocation of sample points in each category of land use using remote sensing imagery. This paper describes a more detailed and more reliable method for determining the most appropriate (i.e., minimum,) sample size. The concept developed and described in the paper incorporates the probability of making incorrect interpretations at particular prescribed accuracy levels, for a certain number of errors, for a particular sample size. The remote sensing sampling strategy presented has the added advantage that it can easily be adapted for use with most forms of remote sensing imagery, including orbital data. It provides a reliable framework for testing the accuracy of any remote sensing image interpretation – based land use classification using the minimum number of sample points; thereby saving time and money, especially if it is employed in operational surveys where high specification accuracy levels need to be guaranteed.

1. Introduction

In recent years the development of techniques for collecting remotely sensed data has progressed very rapidly, but many problems still persist in the utilisation of the information. One of the main problems is the deficiency of appropriate techniques for establishing ground truth using satisfactory sampling techniques. (Kelly, 1970; Zonneveld, 1974; Allan, 1975). Therefore, in order to achieve wider acceptance among users of thematic maps, the interpreter must be able to specify the accuracy of his product. Due to time and cost constraints it would be virtually impossible,

[†] Corresponding author.

©Elsevier North-Holland Inc., 1978

0034-4257/78/0007-0001/$01.75

in the practical sense, to check completely each parcel throughout a region, and so a valid sampling procedure is required to estimate classification accuracy. This paper describes part of a project to develop an operational methodology for rapid production of small scale maps, from LANDSAT multispectral imagery, using inexpensive and unsophisticated techniques (van Genderen and Lock, 1976). As no satisfactory procedure could be located which provided directions for the interpretation, a new scheme was devised and tested. This contrasts with the usual practice of expressing the interpretation errors as a percentage of a subjectively derived number of subject sites by incorporating the probability of making incorrect interpretations at prescribed accuracy levels for a certain number of errors and for a particular sample size. The methodology was tested operationally in Murcia Province, South East Spain with the production of l/250,000 scale rural land use maps.

2. Sampling Procedure

Financial and temporal limitations, combined with the problem of adequately representing important minor classifications in the areal sample, have tended to focus the attention of researchers involved in resource surveys towards some form of stratified sampling technique rather than a strictly random sample. The major difference between the two approaches is that with the stratified random sampling the areal sample space is divided into strata and each stratum is treated as a separate sub-universe in which random sampling is employed. (Kelly, 1970).

Most researchers have stated that they have adopted a particular strategy without fully describing the methods they used for selecting sample sizes, the location and areas of sample sites,and the criteria adopted for accepting or rejecting the sites.

Curtis and Hooper (1974) demonstrated that the notion of ground truth site varies according to the user's objective and the size of the study area. They suggested that the allocation of sample plots can then be estimated by "the method of proportional allocation".

Zonneveld (1972) strongly emphasized that random sampling in land evaluation surveys tends to give too much prominence to the larger areas to the detriment of smaller ones. He believes that the selection of points, within a fully homogeneous area as indicated by interpreted patterns on the imagery, should be random on a stratified basis rather than using a simple random sample over the entire area.

In an attempt to evaluate land classification procedures using simulated space photographs, Rudd (1971) considered that stratified random sampling was most appropriate. After interpretation, the area of each category was measured and the smallest was assigned five sample points and the other categories were allotted sample points in proportion to their respective areas. No reason was given for adopting five sample points for the smallest area.

Dodt and van der Zee (1974) discussed the importance of ground checks in the identification of rural land use by photo-interpretation in order to verify all non-identifiable and ambiguous objects as well as sampling each category to ensure the accuracy of the interpretation. No indication of sampling method or suggestions on appropriate number of sampling sites was given, but they did include some criteria for establishing interpretation accuracy, including the 85-95% limits suggested by Anderson (1971).

Stobbs (1968) used a random sampling method to measure land use in Malawi. He calculated the number of points to be sampled on each photograph by using a pre-determined formula. Although it is one of the few published reports where the mathematical basis for determining the number of sample points is adequately detailed, the design parameters do not permit its extensive use. A similar sampling approach utilising the same formula was performed in Nigeria by Alford *et al.*, (1974).

Multi-stage sampling procedures have been used in studies associated with the production of thematic maps from LANDSAT imagery. Usually, the interpretation of the land use patterns on the orbital imagery becomes the first stratification in the multi-stage design. This is then accompanied by several stages of sub-sampling using low and/or high altitude photography and/or ground data acquisition to quickly and efficiently check interpretations. Unfortunately, there are insufficient published details about this type of sampling design applied to visual interpretation techniques.

In summary, stratified random sampling techniques have been readily accepted as the most appropriate method of sampling in resource studies using remote sensor imagery, so that smaller areas can be satisfactorily represented. But, the problem still remains on the selection of best sample size for each category. Several alternative methods have been used:

(a) stratify the region geometrically and then randomly select points within each square or rectangle (Berry and Baker, 1968). The number of points may be determined by utilizing formulae, (Stobbs, 1968; Alford *et al.*, 1974);

(b) stratify the region by interpreted land use categories, then estimate the total number of sample points that could be visited due to the constraints imposed by time and money, and subsequently distribute them proportionally by area (Zonneveld, 1974).

(c) stratify the region by interpreted category. Allocate a certain number of sample points to the category with the smallest area before distributing sample points to the other categories in proportion to their area (Rudd, 1974).

It is considered that the above methods do not provide sufficient statistical justification for the allocation of sample points in each category of a classification scheme utilizing LANDSAT MSS imagery. Consequently a more detailed and more reliable method for determing the most appropriate (i.e., minimum) sample size should be ascertained. To meet this requirement the following procedure has been developed and successfully employed in operational projects. A fuller description of this methodology may be found in Genderen, J.L. van and Lock, B.F. (1976).

2.1 Sample Size

The function of the ground truth survey in an operational system is to utilize a sound statistical sampling design which will test the correctness of the attribution by interpretation of specific sites to classes in the classification. That is, for any sample point, it should be shown whether the remote sensing allocation to a class within the classification is correct or in error.

Some of the main aspects that need to be considered in such a remote sensing sampling design are:

i. the frequency that any one land use type (on the ground) is erroneously attributed to another class by the interpreter (for example, in Table 1, 3/15 of A is erroneously attributed to other classes)

ii. the frequency that the wrong land use (as observed on the ground)is wrongly included in any one class by the remote sensing interpreter (for example, in Table 1 5/17 of A allocations are wrongly interpreted);

iii. the proportion of all land (as determined in the field) that is mistakenly attributed by the interpreter (for example, in Table 1 8/51 of all attributions are incorrect); and

iv. the determination of whether the mistakes are random (so that overall proportions are approximately correct) or subject to a presistent bias (for example, in Table 1 there may be a significant tendency to misattribute land use C (on the ground) to category A, i.e., 4/16.

Thus the successful design of a sampling and statistical testing procedure will allow an approximate answer to each of these aspects.

In order to determine the optimum sample size (defined as the minimum number of points that need to be checked in the field in order to meet a specification requirement of q accuracy) for a stratified random sample of a region which has been mapped by remote sensing techniques, it is necessary to consider, primarily, one category (stratum) which has been identified from remote sensing imagery. A sample of x points in that category can then be selected and the number of errors (f) checked in the field.

If such a procedure adopts a very small sample (e.g., $x = 10$) the number of errors would normally also be small (e.g., $f = 0,1,2$). However, the achievement of perfect results (i.e., $f = 0$) in such a small sample does not imply that the method is error free, as the result

TABLE 1
Matrix Showing Hypothetical Numbers of Sites
in Actual and Interpreted Land Use Categories

		LAND USE (on the ground)			
		A	B	C	Sum
LAND USE (interpreted from imagery)	A	12	1	4	17
	B	2	19	0	21
	C	1	0	12	13
	Sum	15	20	16	51

may occur by chance in a situation where a substantial proportion of the land use classification was in fact, erroneous. This fact is seldom appreciated by many image interpreters when checking the accuracy of the results of their remote sensing survey. The proportion of the interpretation which is in error would be identified in a very lengthy study, and is normally called p% (or p as a decimal fraction).

This is rarely possible due to time and cost in an operational survey of a large regional area and thus, sample populations are taken. The likelihood of making incorrect interpretations on insufficient samples must therefore be realized and accuracy levels adjusted accordingly.

The probability of making no interpretation errors when taking a sample of x from a remote sensing based classification, with real errors having a probability p, is given by the binomial expansion $(p + q)^x$ in which $q = I - p$.

The binomial expansion of $(q + p)^x$ is given by:

$$(p+q)^x = \underset{f=o}{\Sigma x} \; (f^x)\, p^f q^{(x\text{-}f)}$$

$$=P\ [0\text{ errors}] + P\ [1\text{ error}]$$

$$+ P\ [2\text{ errors}] + \ \ldots\ P\ [50\text{ errors}]$$

$$= P\ [50\text{ or fewer errors}]$$

The probability of f errors in x samples (i.e., P [f errors in x samples]) is given by:

$$P\ [f\text{ errors in }x\text{ samples}]$$

$$=(f^x)\, p^f q^{x\text{-}f}$$

where

$$p = P\ [\text{of making error}]$$
$$q = P\ [\text{of not making error}]$$

In the case of no errors in the interpretation, the last term of the binomial expansion is the only one of interest. (i.e., P [0 errors in x samples] = q^x). This was used to construct Table 2, whilst the binomial probability P [f errors in x samples] = $(f^x)\ p^f q^{x\text{-}f}$, was used in constructing Tables 3 and 4.

Table 2 shows the probability of scoring no interpretation errors in samples of varying sizes taken from a population with a range of real error proportions p. This table indicates that no error sample results can quite easily arise in small samples when the true error rate is high. Taking the conventional probability level of .95/.05 (95% / 5%), the table can be divided into two parts by a 'stepped' line. Above and to the left of the line, the probability of obtaining error free sample results decreases, while below and to the right of the line, it is possible to identify the high probability of obtaining an error free sample, that could only have occurred from a method that was relatively free of true errors.

Therefore, if the permissable error rate in the image interpretation is pre-determined, for example 85-90%, as suggested by the US Geological Survey Circular 671 (Anderson *et al.*, 1972) or as required in an operational job specification the sample size for each category (stratum) necessary for 85% interpretation accuracy should be at least 20, and for 90% accuracy, at least 30. Therefore, by using the table the minimum sample size required for checking any interpretation accuracy can be determined. It is a minimum because for any smaller sample size even a 'perfect' (i.e., error free ground check) result signifies very little. Tables 3 and 4 provide more detailed calculations of the probabilities of scoring errors in samples of varying sizes with specified interpretation levels of 85% and 90% respectively.

2.2 Sampling Strategy

It has been demonstrated that a minimum sample size of at least 30 is necessary for each land use type in order to meet the US Geological Survey Circular 671 criterion of 90% interpretation accuracy in land use surveys using remote sensing imagery. To locate these 30 points, random point sampling within a category (or stratum) can be performed by sampling, using random spatial coordinates. By this method, each random point is attributed to the interpreted land use type in which it falls until a sufficient number of points has been achieved in all categories. Extra points should be taken in order to ensure adequate coverage caused by the inability to reach a particular site due to unforeseen circumstances.

3. Establishment of Ground Truth

The structure of the classification scheme, when correctly established, has certain controlling criteria that tend to regulate the scope of the inquiry and determine the accuracy levels to which the imagery should be interpreted, the minimum size of the areal units and other aspects which can affect the nature of the overall ground truth procedure. These difficulties associated with the establishment of a satisfactory ground truth procedure in which the results of the interpretation can be checked in the field, were discussed in the preceding sections.

TABLE 2
Probability of Scoring no Errors in Samples of Varying Sizes from a Population with a Range of Real Proportions p

		Sample size										
		5	10	15	20	25	30	35	40	45	50	60
Specified interpretation accuracy	0.99											0.5472
	0.95						0.2146	0.1661	0.1285	0.0994	0.0769	0.0461
	0.90			0.2059	0.1216	0.0718	0.0424	0.0250	0.0148	0.0087	0.0052	0.0461
	0.85			0.0874	0.0388	0.0172						
	0.80		0.1074	0.0352								
	0.70	0.1681	0.0282									
	0.60	0.0778										
	0.50	0.0313										

——— stepped line indicates approximately 0.05 level of probability

TABLE 3
Probability of Scoring Errors in Samples of Varying Sizes From
a Population with Real Error Proportion of 85%
i.e. the Specified Interpretation Accuracy Level is 85%

Sample size \ Number of errors (*f*)	0	1	2	3	4	5
15	0.0874					
20	0.0388	0.1368				
25	0.0172	0.0759	0.1607			
30	0.0076	0.0404	0.1034			
35	0.0034	0.0209	0.0627	0.1218		
40			0.0365	0.0816		
45			0.0206	0.0520	0.0963	
50				0.0319	0.0661	0.1072
55				0.0189	0.0434	0.0781
60					0.0275	0.0544
65						0.0365

TABLE 4
Probability of Scoring Errors in Samples of Varying Sizes From a Population with Real Error Proportion of 90%, i.e. the Specified Interpretation Accuracy Level is 90%

Number of errors (*f*) Sample size	0	1	2	3
15	0.2059			
20	0.1216			
25	0.0718	0.1994		
30	0.0424	0.1413		
35	0.0250	0.0973		
40		0.0657		
45		0.0436	0.1067	
50		0.0286	0.0779	
55			0.0558	0.1095
60			0.0393	0.0844
65				0.0636
70				0.0470

——— stepped line indicates approximate 0.05 level of probability

3.1 Sampling Design

The pre-determined sample points are randomly distributed within each category by placing a large sheet of millimetre graph paper under the base map and using a random number table to generate the co-ordinates of the sample sites. These points are then plotted onto the preliminary land use map. The intersections of lines on the graph paper represent the centers of squares on the ground. The graph paper can be positioned so that the grid references of sites could eventually be transferred to the co-ordinate system used on the topographic maps.

The number of sites for each category was determined by considering the interpretation accuracy level prescribed in the criteria for adopting the classification system, and then by consulting the relevant tables. In practice, the optimum sample size is obtained by the addition of several more sites to the value derived from the tables as certain factors, e.g., bad weather and prohibited access may prevent detail at certain sites from being recorded. Also, the first one hundred sample points are maintained in order to obtain an approximate indication of the distribution of the strata and for the considerations of possible trends after completion of the field work.

Although this method removes much of the subjectivity involved in the allocation of sample sites to each category, it entails the generation of a large number of co-ordinates to be plotted in order to determine the categories into which the points fall. When the desired optimum number of points has been reached for a particular category, no more random numbers are added to the list for that category. Ideally, random sampling continues until the optimum number is reached for each category. However, in practice, the areal distribution and coverage of some categories may be too small to permit the generation of enough random points. If this occurs several alternatives are available to satisfy the need of an adequate field check. The investigator may visit all the areas in that category whilst field checking other sites, or use the number of sites that were generated after a tolerable period of random sampling and accept the fact that the accuracy level of that particular category may not meet the accepted level required for the survey. If these alternatives are not reasonable, then, the classification system may have to be readjusted so that the category is subsumed into the next level.

3.2 Analysis of Results

In order to verify the accuracy level of the interpretation of the LANDSAT multispectral imagery and to identify possible reasons for misinterpretation, the selection of appropriate ground data was considered an essential stage. The necessary information had to be in a form that could be collected rapidly and did not involve complicated and time consuming techniques. Therefore, a ground data collection sheet was designed so that as much relevant information as possible could be collected rapidly at each site in a manner that would facilitate subsequent data analysis.

On examination of the ground data collection results, several sites may have to be excluded due to inaccessibilty or misplotting. Then, the data for the sites which were incorrectly interpreted are investigated and comments on possible reasons for misinterpretation listed. On-site photographs are used as additional data sources for isolating causes of the incorrect interpretations. The predominant reasons for misclassification appears to be the occurrence of a site in a small area of a particular category which is located in a much larger area of another category, the size of the small patch not being large enough to meet the smallest mappable area requirements.

4. Conclusions

A scheme for systematically analysing the results of the interpretation of LANDSAT MSS imagery was successfully devised and tested, in the operational sense, during the production of a land use map of Murcia Province, South East Spain. Although the system was based on the commonly used stratified random strategy, one important aspect was developed in this study involving a method of determining the most appropriate sample size. This technique utilizes interpretation accuracy levels that are lower than the normally accepted standards adopted in conventional surveys using air photo-interpretation methods. This situation arises from the concept of incorporating the probability of making incorrect interpretations at particular prescribed accuracy levels, contrasting with the usual practice of expressing the interpretation errors as a percentage of a subjectively derived number of sample sites.

Consequently, it is believed that this approach offers a more meaningful explanation of the interpretation accuracy levels of the whole operation and within each category. Futhermore, it should prove to be very uesful in other types of operational remote sensing projects, saving time and money where stringent specifications need to be met; but prior to this study, it was not possible to check the accuracy of the work in any reliable, statistical manner.

References

Alford, M., E. Hailstone., J. Hailstone, and P. Tuley, (1974), The measurement and mapping of land resource data by point sampling on aerial photograhs, in *Environmental remote sensing: applications and acheivements* (E.C. Barrett and L.F. Curtis, Eds;) Edward Arnold London. pp. 113-126

Allan, J.A. (1975), *Land use in the Merida (Badajoz) region of Spain. An application of the LARS system 3 in a complex agricultural area using ERTS - 1 imagery.* Unpublished paper presented at the Second Annual General Conference of the Remote Sensing Society, National College of Agricultural Engineering, Silsoe, Bedfordshire, 9-11 Sept. 1975.

Anderson, J.R., (1971), Land Use Classification schemes, *Photogrammetric Engineering* **37**, 379-87.

Anderson, J.R., E.E. Hardy, and J.T. Roach, (1972), A land use classification system for use with remote sensor data. Washington, D.C., U.S. Geological Survey Circular 671.

Berry, B.J.L., and A.M. Baker, (1968), Geographic sampling, in *Spatial analysis: a reader in statistical geography* (B.J.L. Berry and D.F. Marble Eds.), Prentice Hall, Englewood Cliffs, N.J., pp. 91-100.

Curtis, L.F., and A.J. Hooper, (1974), Ground truth measurements in relation to aircraft and satellite studies of agricultural land use and land classification in Britain, *European Earth Resources Satellite Experiments,* Paris ESRO, 405-15 (ESRO-SP-100).

Dodt, J., and D. van der Zee, (1974) Identification of rural land use types, *I.T.C. Journal* 1974, 599-616.

Genderen, J.L. van and B.F. Lock, (1976), *A methodology for producing small scale rural land use maps in semi arid developing countries using orbital MSS imagery,* Final Contractors' Report - NASA-CR-151173, Dept. of Industry, London.

Kelly, B.W. (1970), Sampling and statistical problems, in *Remote Sensing with special reference to agriculture and forestry,* Washington, D.C., National Academy of Sciences 329-353.

Rudd, R.D. (1971), Marco land use mapping with simulated space photographs, *Photogrammetric Engineering* **37**, 365-72.

Stobbs, A.R. (1968), Some problems of measuring land use in underdeveloped countries: the land use survey of Malawi, *Cartographic Journal* **5**, 107-110.

Zonneveld, I.S., (1972), *Lectures on Vegetation Science (Ecology) and Vegetation Survey,* Enschede, I.T.C., (ITC Mimeograph Series).

Zonneveld, I.S. (1974), Aerial photography remote sensing and ecology, *I.T.C. Journal* 1974, 553-560.

Received 5 April, 1977: revised 21 June, 1977