Russell G. Congalton
Roy A. Mead
School of Forestry and Wildlife Resources
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

A Quantitative Method to Test for Consistency and Correctness in Photointerpretation*

Error matrices are analyzed using discrete multivariate analysis techniques as an aid in determining proper film/filter combinations, proper seasons, and appropriate interpreters.

Abstract: *A method has been developed to quantitatively test the degree of similarity between photointerpreters and/or photointerpretation variables, such as film/filter type, season, and scale. This method involves giving each photointerpreter the same set of photos to interpret, and then a similarity matrix is generated for each interpreter by comparing the interpretation results to the actual ground cover. It is assumed that the interpretation has been sampled in such a way as to assure that the resulting similarity matrix is representative of that interpretation. The similarity matrix is analyzed using a computer program called* KAPPA*, which implements a discrete multivariate analysis technique to determine if there is a significant difference between matrices. This technique allows the comparison of individual interpreters, a test for photointerpreter consistency, and the comparison of photointerpretation variables.*

Introduction

Photointerpretation, as defined by the American Society of Photogrammetry (1966), is "the act of examining photographic images for the purpose of identifying objects and judging their significance." Good, consistent photointerpretation depends on the experience and skill of the individual interpreter. The judgment involved is generally qualitative in nature and is, therefore, difficult to evaluate or compare with interpretations made by others. This paper suggests a method of quantifying photointerpretation results and gives a statistical method by which interpreters and other photointerpretation variables such as film/filter types can be compared.

The procedure proposed in this paper can test for significant differences between photointerpreters, test the consistency of the same interpreter over time, or test for significant differences between photointerpretation variables such as film type, season, and scale. Testing to see if interpreters are similar is useful when more than one interpreter is to work on the same project. If it can be determined that the delineations or identifications made by all the interpreters are not significantly different, then the project will yield uniform results. Also, it may be useful to test the same interpreter over a period of time to check for consistency in his interpretation skill. Lastly, it may also be important to determine if varying types of photography (film/filter combinations) or seasons (spring versus summer) of photography result in significantly different identifications.

Procedure

The first step in quantifying photointerpretation results is to cellularize the photographs. There are

* Presented, in part, at the 1981 Annual Convention of the American Society of Photogrammetry, Washington, D.C.

Photogrammetric Engineering and Remote Sensing,
Vol. 49, No. 1, January 1983, pp. 69-74.

0099-1112/83/4901-69$02.25/0

many ways that this may be accomplished. The identifications or delineations made by the interpreter can be digitized, and then a given cell size can be used to assign the majority land-use/land-cover category to each cell. A dot grid could also be used in which each dot defines the sample point. Obviously, the results of the assignment to categories will vary with different cell sizes or dots per square inch. Therefore, it is important to choose a cell size that is indicative of the objectives of the photointerpretation (i.e., the more detail required, the smaller the cell size). The minimum mapping unit used in the photointerpretation is a good choice for the cell size.

Once a grid system has been placed over the photo and each cell has been assigned to some category, then each cell is compared with its corresponding cell from another interpretation. If one of these interpretations is assumed to be correct (reference data), then comparison of the two sets of spatially defined cells yields a measure of "photointerpretation accuracy." This comparison is usually expressed in the form of an error matrix.

It will be assumed in this paper that the error matrices generated are representative of the actual photointerpretation results. This means that an appropriate sampling scheme was chosen and an adequate number of samples were taken when compiling the matrix. Because the technique proposed here analyzes the error matrices, it is necessary that this assumption hold for any results to be meaningful.

An error matrix is a square array set out in rows and columns which expresses the number of cells assigned to a particular land-cover type (photointerpretation) relative to the actual land cover (reference data). The columns represent the reference data while the rows indicate the photointerpreter assigned land-cover type (Table 1). The numbers in the error matrix are the actual tallies compiled by comparing the photointerpretation with the actual land-cover type on a cell by cell basis. As can be seen in Table 1, all correct classifications are indicated on the major diagonal of the error matrix. For example, in this case 16 cells were correctly classified as forest. Also note that the off-diagonal elements indicate misclassification, which is a combination of omission and commission error.

The specific method used to generate an error matrix is dependent upon what information is desired. If the similarity between two or more photointerpreters is to be determined, then each interpreter is given the same aerial photographs to interpret. A similarity matrix is tabulated for each interpreter by comparing his interpretation with a reference data set (assumed correct). If the test involves determining the consistency over time for a single interpreter, then a representative sample of the photos is selected for interpretation at the beginning of the project. At some later date, the remainder of the photos are interpreted and the two similarity matrices (Time I versus Reference Data and Time II versus Reference Data) are compared. Finally, if it is desired to measure the similarity of identifications made on different types of photography, a separate interpretation is performed on the same area for each set of photos by each interpreter and the appropriate similarity matrices are generated.

Statistical Analysis

Once the similarity matrices have been generated, a discrete multivariate analysis technique (Bishop *et al.*, 1975) is used to test the agreement between matrices. This technique is appropriate because the data are discrete and multinomially distributed. Parametric statistical techniques such as analysis of variance assume that the data are continuous and normally distributed. Chi square analysis is possible except that problems arise when cell values in the matrix equal zero. This problem does not affect the discrete multivariate analysis technique.

The statistic used to test the agreement between similarity matrices is called $\hat{K}$ (KHAT). KHAT is the

Table 1. Example Error Matrix for Three Cover Types

Photointerpretation	Reference Data: Pasture	Reference Data: Forest	Reference Data: Water
Pasture	22	8	1
Forest	4	16	0
Water	4	3	9

maximum likelihood estimate from the multinomial distribution and is a measure of the actual agreement minus the change agreement (Equation 1). The actual agreement is the cell value itself while the chance agreement is defined as the product of the marginals (row and column totals) for that cell.

$$\hat{K} = \frac{\sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r} (X_{i+} * X_{+i})}{N^2 - \sum_{i=1}^{r} (X_{i+} * X_{+i})} \quad (1)$$

where

r = number of rows and cols. in error matrix,
X_{ii} = number of obs. in row i and col. i,
X_{i+} = marginal total of row i,
X_{+i} = marginal total of col. i, and
N = total number of obs.

A KHAT value is calculated for each matrix and is a measure of how well the photointerpretation agrees with the ground verified reference data. The approximate large sample variance of KHAT, $\hat{\sigma}^2$, can then be used to construct a hypothesis test for significant difference between error matrices (Cohen, 1960). This test is possible because the large sample asymptotic distribution of KHAT is normal. The test statistic for significant difference is

$$\frac{\hat{K}_1 - \hat{K}_2}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}} \sim Z.$$

As mentioned above, this test statistic is normally distributed. this means that the values in a standard normal table at varying confidence levels can be used to determine if the two matrices are significantly different. For example, at the 95 percent confidence level, if the test statistic is greater than 1.96, one would reject the null hypothesis and conclude that the two matrices were significantly different.

This entire comparison process can be performed using a FORTRAN computer program called KAPPA. The program listing, documentation, and additional examples can be found in Congalton *et al.* (1982). Given the error matrices to be analyzed, the program calculates a KHAT value and variance for each matrix. This information is then used to perform a significance test on each pair of matrices. The program prints out the KHAT value, the variance, the test statistic for each pair of matrices, and the results of the test.

Example Data Analysis

The data used here to present examples of the analysis technique were taken from Lauer *et al.* (1970). Five photointerpreters of equal skill (experience) interpreted the same aerial photographs of Yosemite Valley, California. Two of these original matrices are shown here (Table 2) so that interested readers can verify the results and compare them to other techniques. Also, five film and filter combinations were used with a single interpreter. Finally, three maximum skill and three varying skill interpreters were given the same photos and asked to interpret them. The appropriate similarity matrices were compiled for each case.

Results and Discussion

None of the five intepreters tested on the photographs from Yosemite Valley produced significantly different interpretations (Table 3). The results were calculated at the 95 percent confidence level and indicate that these five interpreters could work together in this area without generating significantly different interpretations. Note that interpreters 3 and 5 were almost significantly different. Also note that interpreter 3 had the highest accuracy, while interpreter 5 had the lowest.

The results of the five different film and filter combinations are presented in Table 4. These film/filter combinations were (1) black-and-white infrared film with a Wratten 25 filter, (2) black-and-white infrared film with a Wratten 89B filter, (3) color infrared film, (4) enhancement X, and (5) enhancement. Y. Enhancement X was made by optically combining three narrow band-pass film/filter combinations (553, 682, and 754 nanometres) projected through red, blue, and green filters, respectively. While enhancement Y was made by optically combining black-and-white infrared film with three filters (Wratten 58, Wratten 25, and Wratten 89B) projected through green, green, and red filters, respectively (Lauer *et al.*, 1970). These results were also calculated at the 95 percent level. Note that the interpretations performed using EktaAero IR (color infrared) film yielded the lowest accuracy. Also, the IR-301/W25, IR/W89B, and enhancement Y combinations yielded significantly different interpretations from the EktaAero IR combination. However, all other combinations were not significantly different. This result indicates that the EktaAero IR film should not be used in this case. Any of the other four film/filter combinations would produce similar results, and therefore other criteria such as cost, availability, and ease of acquisition should be used to decide what film/filter combination is appropriate for this area.

Finally, the results of the maximum skilled interpreters versus the varying skilled interpreters are presented in Table 5. Each interpreter was asked to make general identifications between alfalfa, sorghum, cotton, and bare soil on photographs taken with black-and-white infrared film and a Wratten 25 filter. As can be seen from looking at the values of the test statistics, none of the

TABLE 2. TWO ORIGINAL ERROR MATRICES USED TO COMPARE PHOTOINTERPRETERS

Interpreter #1

Photointerpretation	Reference Data: Pine	Cedar	Oak	Cottonwood
Pine	35	14	11	1
Cedar	4	11	3	0
Oak	12	9	38	4
Cottonwood	2	5	12	2

Interpreter #2

Photointerpretation	Reference Data: Pine	Cedar	Oak	Cottonwood
Pine	32	15	5	3
Cedar	7	8	5	0
Oak	7	8	38	2
Cottonwood	6	7	15	1

TABLE 3. SUMMARY TABLE FOR FIVE INTERPRETERS OF YOSEMITE VALLEY PHOTOS

Interpreter	KHAT	Variance	Combination	Z Statistic	Result
1	0.31991	0.00288059	1,2	0.3465	NS
			1,3	−0.7320	NS
2	0.29420	0.00262762	1,4	0.9768	NS
			1,5	1.2535	NS
3	0.37485	0.00275198	2,3	−1.0997	NS
			2,4	0.6695	NS
4	0.24156	0.00355440	2,5	0.9521	NS
			3,4	1.6785	NS
5	0.21925	0.00356825	3,5	1.9572	NS
			4,5	0.2643	NS

TABLE 4. SUMMARY TABLE FOR FIVE FILM AND FILTER COMBINATIONS

Interpreter	KHAT	Variance	Combination	Z Statistic	Result
IR-301/W25 (1)	0.31991	0.00288059	1,2	0.2052	NS
			1,3	2.6833	S
IR/W89B (2)	0.30436	0.00286389	1,4	0.7870	NS
			1,5	−0.7390	NS
Ekta Aero IR (3)	0.12071	0.00263067	2,3	2.4775	S
			2,4	0.5779	NS
Enhancement *X* (4)	0.26163	0.00260384	2,5	−0.9514	NS
			3,4	−1.9477	NS
Enhancement *Y* (5)	0.37438	0.00255323	3,5	−3.5232	S
			4,5	−1.5702	NS

interpreters were significantly different. Note, however, that the three maximum skilled interpreters did have the three highest accuracies although they were not significantly different from the varying skill interpreters. This result indicates that maximum skill was not necessary to be able to identify the four categories of interest. It does not mean that varying skill is always as good as maximum skill. By increasing the detail of the identification, it is expected that the maximum

TABLE 5. SUMMARY TABLE FOR THE MAXIMUM SKILLED VERSUS VARYING SKILLED INTERPRETERS

Interpreter	KHAT	Variance	Combination	Z Statistic	Result
Varying skill (1)	0.55129	0.00314970	1,2	0.2618	NS
			1,3	0.1673	NS
			1,4	−1.3201	NS
Varying skill (2)	0.53027	0.00330095	1,5	−0.3019	NS
			1,6	−0.1381	NS
Varying skill (3)	0.53808	0.00309568	2,3	−0.0976	NS
			2,4	−1.5745	NS
			2,5	−0.5601	NS
Varying skill (4)	0.65240	0.00271537	2,6	−0.3970	NS
			3,4	−1.4996	NS
Varying skill (5)	0.57526	0.00315077	3,5	−0.4704	NS
			3,6	−0.3051	NS
			4,5	1.0072	NS
Varying skill (6)	0.56232	0.00321424	4,6	1.1698	NS
			5,6	0.1622	NS

skilled interpreters will perform significantly better than the varying skilled interpreters. This example demonstrates that for this level of detail it would not be necessary to hire the best interpreters at the highest salaries. Instead, good interpreters would do as well at these identifications and would cost less to employ.

Conclusions

The examples given in this paper demonstrate how photointerpretation results can be quantified using error matrices. These error matrices can then be statistically analyzed using a discrete multivariate analysis technique. The results of this analysis can aid in determining proper film/filter combinations, proper seasons, and appropriate interpreters. It can also help in evaluating and improving photointerpretation results. The technique proposed is simple and straightforward. Once the matrices have been generated, the rest of the analysis can be performed by a computer program. The ease by which the entire quantification process is performed should lead to its widespread use in photointerpretation projects.

References

American Society of Photogrammmetry, 1960. *Manual of Photographic Interpretation.* Falls Church, Virginia.

Bishop, Y. S., S. Fienberg, and P. Holland, 1975. *Discrete Multivariate Analysis: Theory and Practice.* MIT Press: Cambridge, Massachusetts. 575 p.

Cohen, Jacob, 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* Vol. 20, No. 1, pp. 37-40.

Congalton, R. G., R. G. Oderwald, and R. A. Mead, 1982. *Accuracy of remotely sensed data: sampling and analysis procedures.* Remote Sensing Research Report 82-1. Agristars Report RR-U2-04257. Coop Agreement 13-1134. Virginia Polytechnic Institute and State University. 83 p.

Lauer, D. T., C. M. Hay, and A. S. Benson, 1970. *Quantitative evaluation of multiband photographic techniques.* Final Report for Earth Observation Division, Manned Spacecraft Center, NASA, Contract No. NAS9-9577. 110 p.

(Received 7 August 1981; revised and accepted 30 September 1982)

International Symposium on Maps and Graphics for the Visually Handicapped

Washington, D.C.
10-12 March 1983

With the support of the Knights of Columbus (Washington State Chapter), the National Geographic Society, and the U. S. Geological Survey, the Association of American Geographers with the help of its co-sponsors (the U. S. National Committees of the International Cartographic Association and the International Geographical Union) has organized this International Symposium on Maps and Graphics for the Visually Handicapped, which will immediately precede the Annual Convention of the American Congress on Surveying and Mapping and the American Society of Photogrammetry.

The role which maps and other forms of graphic devices might play as aids in enhancing the life of visually handicapped individuals has become an important concern of cartographers, educators, geographers, and other specialists around the world. To explore this research area, the symposium will be organized around four major themes:

- Contexts within which maps and other graphics can serve as useful tools for the visually handicapped (work related, education, recreational, etc.).
- New Research Findings—The exchange of knowledge regarding perception, information content, symbol parameters, multi-model communication, and new experimental programs.
- Alternative techniques for producing tactual maps, audio tapes, and large print forms (molded plastics and embossed paper).
- Future research agenda related to improved design, reproduction, and use of maps and graphics for the visually handicapped.

For further information please contact

Particia J. McWethy
Executive Director
Association of American Geographers
1710 16th Street N. W.
Washington, DC 20009
Tele. (202) 234-1450

Joseph W. Wiedel
Chair, Steering Committee
Department of Geography
University of Maryland
College Park, MD 20742
Tele. (301) 454-6602