General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

NASA TECHNICAL MEMORANDUM

NASA TM X-73347

EVALUATION CRITERIA FOR SOFTWARE CLASSIFICATION INVENTORIES, ACCURACIES, AND MAPS

By Robert R. Jayroe, Jr.
Data Systems Laboratory

September 1976

NASA

George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.
NASA TM X- 73347		

4. TITLE AND SUBTITLE

Evaluation Criteria for Software Classification Inventories, Accuracies, and Maps

5. REPORT DATE
September 1976

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Robert R. Jayroe, Jr.

8. PERFORMING ORGANIZATION REPORT #

9. PERFORMING ORGANIZATION NAME AND ADDRESS

George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

10. WORK UNIT NO.

11. CONTRACT OR GRANT NO.

13. TYPE OF REPORT & PERIOD COVERED
Technical Memorandum

12. SPONSORING AGENCY NAME AND ADDRESS

National Aeronautics and Space Administration
Washington, D.C. 20546

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

This work was sponsored by the Office of Applications' Data Management Program.
Prepared by Data Systems Laboratory, Science and Engineering

16. ABSTRACT

The main tool for comparing remote sensing classification results with ground truth information is a contingency table derived from overlaying digital classification and ground truth maps. The purpose of this report is to explore methods of deriving a maximum amount of information from the contingency table and of modifying the contingency table to provide more information. This report contains 15 different statistical criteria derived from a contingency table that can be used to evaluate tabular classification results, which unfortunately provide little information on the visual characteristics of a classification map. Tabular results provide information relating mainly to how much rather than where, which is the purpose of a map. Therefore modifications are proposed to the contingency table which contain information on the spatial complexity of the test site, on the relative location of classification errors, on how well the classification maps agree with the ground truth maps, and which reduce back to the original information normally contained in a contingency table.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
Classification Remote sensing Classification Technique Evaluation	Categories 43 and 61

19. SECURITY CLASSIF. (of this report)	20. SECURITY CLASSIF. (of this page)	21. NO. OF PAGES	22. PRICE
Unclassified	Unclassified	27	NTIS

ACKNOWLEDGMENT

The author wishes to acknowledge Dr. R. Atkinson, Dr. H. Ramapriyan, Dr. B. Dasarthy, and Mr. M. Lybanon of the Computer Science Corporation, Huntsville, Alabama, for programming the software that provided the results for this report.

TABLE OF CONTENTS

LIST OF TABLES

TECHNICAL MEMORANDUM X-73347

EVALUATION CRITERIA FOR SOFTWARE CLASSIFICATION INVENTORIES, ACCURACIES, AND MAPS

I. INTRODUCTION

Considerable emphasis is now being given to the evaluation of image classification and compression techniques. This report describes the evaluation criteria and procedures that have been proposed and developed to focus attention on the existing state of the art and provide guidance for future research efforts. Although there are many criteria, e.g. costs, running times, computer resources, etc., that should be considered in evaluating techniques, the main emphasis of this report is concerned with statistical performance.

Assume that multispectral image data have been classified using a particular technique to produce a classification map (CM) and that the CM has been overlayed with a digital version of a ground truth map (GTM). The normal procedure is to produce a contingency table, such as shown in Table 1, and determine a percentage accuracy as a measure of the goodness of a classification technique. However, there would appear to be considerable risk involved in judging the merits of various classification techniques based upon this one number. Hence, one of the purposes of this report is to mathematically explore the contingency table to determine how much additional information can be extracted. However, it must also be kept in mind that the table only provides numerical results and contains relatively little information concerning the map producing abilities of the various classification techniques. The desired end result is that there will be a sufficient number of mathematical criteria that can be examined to ensure as much completeness in the evaluation as possible. Criteria and procedures similar to what is discussed in the report can also be adapted to evaluate compression and change detection analysis results.

The contingency tables used in this report resulted from a cooperative evaluation of classification techniques which involved Marshall Space Flight Center, Huntsville, Alabama, and the Tennessee State Planning Office, Nashville, Tennessee. Landsat data from the Bald Knob, Tennessee, Quadrangle were used as a test site and four sets of seasonal data were also included for multitemporal evaluation. All of the techniques discussed in this report are

supervised techniques and all used the same training areas for the classification results. The five classification results that are discussed include the Gaussian Maximum Likelihood which was used on one season of data as well as all four seasons simultaneously, the Linear Classifier Model which was also used on one season as well as all four seasons, and the Density Slicing Classifier which was used on only one season of data. The Linear Classifier uses hyperplanes to separate feature categories, while the Density Slicing Method selects a channel of data as well as a class interval in that channel to separate feature categories.

Section II describes contingency tables and tests derived from the tables in a general manner, and Section III describes the evaluation of the classification analysis results. Section IV describes a proposed approach for evaluating classification maps that reduces back to the normally used contingency table.

TABLE 1. GENERAL 5 BY 5 CONTINGENCY TABLE

GTM \ CM		1	2	3	4	5		
1	$n^{\pi}11$	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	e_1	$e^{\pi}1$
2	$n^{\pi}22$	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	e_2	$e^{\pi}2$
3	$n^{\pi}33$	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	e_3	$e^{\pi}3$
4	$n^{\pi}44$	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	e_4	$e^{\pi}4$
5	$n^{\pi}55$	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	e_5	$e^{\pi}5$
		o_1	o_2	o_3	o_4	o_5	N_T	%c
		$o^{\pi}1$	$o^{\pi}2$	$o^{\pi}3$	$o^{\pi}4$	$o^{\pi}5$	%I	N_c

II. MATHEMATICAL DESCRIPTION OF EVALUATION CRITERIA

Table 1 shows the general form of a 5 by 5 contingency table that is consistent in size with the tables used in Section III. The table indicates that there are five categories on the GTM being compared with five categories on the CM. The elements $n_{i,j}$ tell how many pixels in class j on the CM occur at the same locations as pixels in class i on the GTM. The symbol e_i is the number of pixels belonging to category i on the GTM and is the number that is expected to be obtained from the classification results. The symbol o_j is the number of pixels that were classified in category j on the CM or the number that is observed, which is usually different from what is expected. Mathematically speaking

$$e_i = \sum_j n_{i,j} \qquad \text{and} \qquad o_j = \sum_i n_{i,j} \quad . \tag{1}$$

The symbols π are probabilities of occurrences,

$$e\pi_i = e_i/N_T, \quad o\pi_j = o_j/N_T, \quad \text{and} \quad n\pi_{ii} = n_{ii}/N_T \quad , \tag{2}$$

where N_T is the total number of pixels. The symbols N_c, $\%c$, and $\%I$ are the number of correctly classified pixels, the classification accuracy, and inventory accuracy, respectively. These are computed using the following relations:

$$N_c = \sum_i n_{ii} \ , \quad \%c = 100(N_c/N_T) \ , \quad \text{and} \quad \%I = 100\left[1 - \sum_i \frac{c_i - o_i}{2N_T}\right] \ . \tag{3}$$

For the inventory accuracy, the number wrong is given by the summation of the absolute value differences, which has to be divided by two. The factor of two is necessary because if one pixel changes category two columns are affected on the contingency table and the pixels are in effect counted twice. The inventory accuracy can also be computed by choosing the smaller of e_i or o_i, summing over the categories, and multiplying by $100/N_T$ which gives the same result.

Two other tables can be generated from the actual contingency table; however, it is not necessary to do so because the actual table already contains the information. These two tables will be discussed to illustrate the concepts of randomness and optimumness.

The concept of randomness is illustrated using the maximum likelihood estimators. The likelihood of an observed sample of N_T being picked from an assumed population, i.e., e_i and o_j are given and remain constant under all conditions, is tantamount to replacing $n_{i,j}$ with $e_i o_j / N_T$ or $o_j \pi_i e_i$ in the contingency table. The only other quantities that change in the table are N_c, the number of correctly classified pixels, and %c. This result should hold true for any sample of size N_T picked from an assumed population and should be a random or "worst case" classification accuracy that is expected.

The optimum case classification accuracy that can be expected for a given inventory (e_i and o_j given) occurs when the classification accuracy equals the inventory accuracy. This is tantamount to replacing $n_{i,i}$ with the smaller of e_i or o_i on the diagonal, and the remaining $n_{i,j} (i \neq j)$ will either be zero or indeterminant. The only other quantities that are changed are again N_c and %c.

There are several statistical performance criteria that can now be calculated from the contingency table and these are discussed as follows:

1. The first criteria is the actual classification accuracy. The classification accuracies for the random and optimum cases provide upper and lower limits for the accuracy range, and a percent of optimum accuracy can be computed for a technique as a measure of how well it performed versus how well it could have performed.

The remaining criteria are concerned with chi-squared tests that are convenient to use because the table contains information related to what is expected and what is observed. The chi-squared tests and formulas for computing the chi-squared values relating to those tests are as follows:

2. Hypothesis: The distribution (o_j) of the classification inventory agrees with the distribution (e_j) of the ground truth inventory:

$$\chi_1^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} = \sum_j o_j^2/e_j - N_T = \frac{1}{N_T}\left[\sum_j o_j^2/e\pi_j - N_T^2\right]. \qquad (4)$$

3. Hypothesis: The distribution $(n_{i,i})$ of the correctly classified pixels agrees with the distribution of the ground truth inventory:

$$\chi_2^2 = \frac{1}{N_c}\left[\sum_i n_{i,i}^2/e\pi_i - N_c^2\right] \quad . \qquad (5)$$

4. Hypothesis: The distribution of the number of correctly and incorrectly classified pixels is optimum with respect to the given inventory and without regard to class:

$$\chi_3^2 = \frac{\left[\sum_i' \left(n_{i,i} + \frac{|e_i - o_i|}{2}\right) - N_T\right]^2}{N_T - \sum_i' \frac{|e_i - o_i|}{2}} + \frac{\left[N_T - \sum_i' \left(n_{i,i} + \frac{|e_i - o_i|}{2}\right)\right]^2}{\sum_i' \frac{|e_i - o_i|}{2}} \quad . \qquad (6)$$

These three chi-squared values should be as small as possible to satisfy the hypotheses, while the remaining chi-squared values to be discussed should be as large as possible so that the hypotheses will be rejected.

5. Hypothesis: The correctly classified pixels are randomly distributed:

$$\chi_4^2 = \frac{1}{{}_A N_c}\left[\sum_i' \frac{n_{i,i}^2 \, {}_R N_c}{o\pi_i \, c_i} - {}_A N_c^2\right] \quad , \qquad (7)$$

where ${}_A N_c$ and ${}_R N_c$ are the number of correctly classified pixels for the actual and random case.

6. Hypothesis: Each classification feature is randomly distributed among the ground truth features according to the classification inventory:

$$ {}_i\chi_5^2 = \frac{1}{e_i}\left[\sum_j \frac{n_{i,j}^2}{{}_o\pi_j} - e_i^2\right] , \qquad (8) $$

where i refers to the feature on the GTM and j refers to the feature on the CM.

7. Hypothesis: Each ground truth feature is randomly distributed among the classification features according to the ground truth inventory:

$$ {}_j\chi_6^2 = \frac{1}{o_j}\left[\sum_i \frac{n_{i,j}^2}{{}_e\pi_i} - o_j^2\right] , \qquad (9) $$

where i and j have the same meaning as in equation (8).

8. Hypothesis: The number of correctly and incorrectly classified pixels are randomly distributed without regard to class:

$$ \chi_7^2 = \frac{\left[\sum_i \left(n_{i,i} - {}_o\pi_i\, e_i\right)\right]^2}{\sum_i {}_o\pi_i\, e_i} + \frac{\left[\sum_i \left(n_{i,i} - {}_o\pi_i\, e_i\right)\right]^2}{N - \sum_i {}_o\pi_i\, e_i} . \qquad (10) $$

9. Hypothesis: The number of correctly and incorrectly classified pixels for a particular class are randomly distributed:

$$ {}_j\chi_8^2 = \frac{\left[n_{j,j} - {}_o\pi_j\, e_j\right]^2}{{}_o\pi_j\, e_j} + \frac{\left[n_{j,j} - {}_o\pi_j\, e_j\right]^2}{e_j - {}_o\pi_j\, e_j} , \qquad (11) $$

where j represents the class.

10. Hypothesis: The distribution of the classified pixels is independent of the ground tryth:

$$\chi_9^2 = N_T\left(\sum_{i,j} \frac{n_{i,j}^2}{o_i e_j} - 1\right) \quad , \tag{12}$$

11. The final criterion is the coefficient of contingency, which is similar to a correlation coefficient and is calculated from χ_8^2. The coefficient is given by

$$C = \left[\frac{\chi_8^2}{N_T(k-1)}\right]^{1/2} \quad , \tag{13}$$

where k is either the number of features on the GTM or CM, whichever is smaller.

For relatively comparing various classification techniques, the best values observed for all the chi-squared tests can be chosen as the expected chi-squared values. The actual observed chi-squared values for a particular technique can then be measured against what is expected by computing chi-squared values. The use of these criteria is illustrated in the next section.

III. EVALUATION RESULTS

Tables 2 through 6 are the contingency tables for the various techniques being examined, and u, t, a, f, and w are the feature categories urban, transportation, agriculture, forest, and water, respectively. The techniques are identified by the labels:

MLCM — Maximum Likelihood Classifier (Map)
LCM — Linear Classifier (Map)
MLMCM — Maximum Likelihood Multitemporal Classifier (Map)
LMCM — Linear Multitemporal Classifier (Map)
DSCM — Density Slicing Classifier (Map)

All of the classification programs are supervised techniques, and all programs were supplied the same training areas. The multitemporal programs used 16 channels of seasonal data rather than one season containing only 4 channels. Thus, all of the results have one season of data in common.

TABLE 2. CONTINGENCY TABLE FOR GTM VERSUS MLCM

GTM \ MLCM		u	t	a	f	w		
u	.2138	59	47	35	134	1	276	.0083
t	.1931	129	163	142	403	7	844	.0253
a	.6165	2325	751	5904	582	14	9576	.2868
f	.7945	[illegible]	2404	745	17488	238	22011	.6593
w	.4639	40	95	11	218	315	679	.0203
		3689	3460	6837	18825	575	33386	71.67
		.1105	.1036	.2048	.5639	.0172	81.94	23929

Table 7 lists the statistical criteria as a function of classification technique, and the numbers followed by an asterisk indicate the best numbers that were observed. The degrees of freedom (df) associated with each chi-squared value is also listed in Table 7. Of the 26 possible best numbers, MLCM has 6 of them, LCM has 3, MLMCM has 11, LMCM has 4, and DSCM has 2. By using the numbers followed by an asterisk as expected values, a chi-squared value can be computed for each technique that has n-1 or 25 df. These chi-squared values are listed in Table 8.

Tables 2 through 8 represent a considerable amount of information that needs an equal amount of discussion. First, for 1 df there is a 0.05 probability of finding a chi-squared value larger than 3.841 and a 0.01 probability of finding a value larger than 6.635. For 4 df the 0.05 and 0.01 chi-squared values are 9.488 and 13.277; for 16 df the 0.05 and 0.01 values are 26.296 and 32.0; and for 25 df the 0.05 and 0.01 values are 37.652 and 44.314. Using these values

TABLE 3. CONTINGENCY TABLE FOR GTM VERSUS LCM

GTM \ LCM		u	t	a	f	w		
u	.1811	50	36	41	149	0	276	.0083
t	.1789	101	151	164	427	1	844	.0253
a	.6587	1811	610	6308	842	5	9576	.2868
f	.8091	947	2180	897	17809	178	22011	.6593
w	.4271	28	62	17	282	290	679	.0203
		2937	3039	7427	19509	474	33386	73.71
		.088	.091	.2225	.5843	.0142	85.45	24608

TABLE 4. CONTINGENCY TABLE FOR GTM VERSUS MLMCM

GTM \ MLMCM		u	t	a	f	w		
u	.2101	58	49	44	124	1	276	.0083
t	.2927	101	247	175	316	5	844	.0253
a	.7049	1369	1126	6750	299	32	9576	.2868
f	.7641	1026	2718	1088	16819	360	22011	.6593
w	.5287	36	133	34	117	359	679	.0203
		2590	4273	8091	17675	757	33386	72.58
		.0776	.128	.2423	.5294	.0227	82.56	24233

TABLE 5. CONTINGENCY TABLE FOR GTM VERSUS LMCM

GTM \ LMCM		u	t	a	f	w		
u	.1775	49	35	43	148	1	276	.0083
t	.1765	89	149	160	443	3	844	.0253
a	.6638	1714	839	6357	650	16	9576	.2868
f	.8454	875	1267	1044	18609	216	22011	.6593
w	.4212	34	72	21	266	286	679	.0203
		2761	2362	7625	20116	522	33386	76.23
		.0827	.0707	.2284	.6025	.0156	88.01	25450

TABLE 6. CONTINGENCY TABLE FOR GTM VERSUS DSCM

GTM \ DSCM		u	t	a	f	w		
u	.2464	68	51	33	124	0	276	.0083
t	.1955	159	165	117	401	2	844	.0253
a	.5473	2767	609	5241	944	15	9576	.2868
f	.7497	1658	3085	595	16502	7	22011	.6593
w	.3608	43	44	12	335	245	679	.0203
		4695	3954	5998	18306	433	33386	65.86
		.1406	.1184	.1797	.5483	.013	77.45	21988

TABLE 7. STATISTIC VERSUS TECHNIQUE

Statistic \ Technique		MLCM	LCM	MLMCM	LMCM	DSCM
Random Classification Accuracy		43.11	45.23	42.29	46.55*	41.[illegible]
Actual Classification Accuracy		71.67	73.71	72.58	76.23*	65.80
Inventory or Optimum Accuracy		81.91	85.45	82.56	88.01*	77.45
Percent of Optimum Accuracy		73.32	70.81	75.22*	71.59	67.54
X_1^2	4 df	51574	26484	20191*	25701	84262
X_2^2	4 df	801	796	407*	929	1361
X_3^2	4 df	2264*	3707	2310	4390	2567
X_4^2	4 df	6102	4959	5833	4305	7773*
$_uX_5^2$	4 df	53	44	83*	53	46
$_tX_5^2$	4 df	101	103	251*	161	80
$_aX_5^2$	4 df	13941*	13545	13640	13725	12572
$_fX_5^2$	4 df	5968*	5751	5368	5906	5285
$_wX_5^2$	4 df	7865	8321*	7956	7348	6467
$_uX_6^2$	4 df	2264*	1673	886	1608	2230
$_tX_6^2$	4 df	151	170	220	250	399*
$_aX_6^2$	4 df	11202	11590	11951*	11236	10173
$_fX_6^2$	4 df	6459	6983	6981*	6775	5177
$_wX_6^2$	4 df	8080	8368*	7896	7335	6500
X_7^2	1 df	10834	10926	12557*	11618	7984
$_uX_8^2$	1 df	30	30	56*	33	26
$_tX_8^2$	1 df	73	79	205*	144	48
$_aX_8^2$	1 df	9068	10531	11164*	10304	8777
$_fX_8^2$	1 df	4760	4579	4867	5424*	2480
$_wX_8^2$	1 df	8015	8269*	7829	7275	6402
X_9^2	16 df	28139*	27767	27921	27173	24466
%c	4 df	45.90*	45.60	45.72	45.18	42.83

TABLE 8. TECHNIQUE VERSUS CHI-SQUARED VALUES USING TABLE 7 DATA

Technique	MLCM	LCM	MLMCM	LMCM	DSCM
Chi-squared	48950	5100	1597	6409	207832

and examining Tables 7 and 8 show that every single hypothesis was rejected and hardly any of the chi-squared values are even close to these numbers. An attempt was made to understand why the chi-squared values are so large by using the inventory from MLCM and computing the chi-squared values as a function of inventory accuracy. Equation (3) shows that the proportion of wrongly classified pixels for each category j is given by

$$\frac{|e_j - o_j|}{2N_T} \quad . \tag{14}$$

If it assumed that these proportions remain constant for any inventory accuracy, then Table 9 shows the inventory and chi-squared which result from this assumption.

TABLE 9. INVENTORY ACCURACY VERSUS CHI-SQUARED

Inventory Category, o_j							
u	t	a	f	w	%I	χ_1^2	Optimum χ_2^2
370	916	9500	21923	676	99.5	39.12	0.1187
465	989	9424	21835	673	99	158.21	0.4997
1221	1568	8817	21129	650	95	3953	13.685
2166	2293	8059	20247	622	90	15817	58.15
3110	3017	7300	19365	59.	85	35564	139
4055	3742	6542	18483	564	80	63239	263

Thus, it is not possible to accept the hypothesis that the distribution of the classification inventory is statistically significant when compared with the ground truth inventory even though the inventory is 99.5 percent correct. If the optimum classification accuracy is considered, then the chi-squared value is almost significant at 95 percent classification accuracy. Hence, it appears that the chi-squared tests are extremely strict, but because of this it also appears to be extremely good at relatively discriminating between the performance of various techniques.

Tables 7 and 8 show that different conclusions would be obtained if the techniques were judged on classification accuracy only versus a set of criteria. Presumably, the set of criteria provides for better judgment because it offers a more complete description of performance.

In Table 7, χ_1^2 shows that MLMCM benefited the most from the use of multitemporal data even through the classification accuracy increased less than 2 percent and the inventory accuracy less than 1 percent. This indicates that the inventory distribution improved considerably, and the inventory has to be relied on when there are no ground truth results. The inventory accuracy is usually higher than the classification accuracy because the misclassified pixels tend to cancel out not having classified enough pixels correctly.

The values for χ_2^2 show that the correctly classified pixels are better estimators of the ground truth inventory distribution than the classification inventory. Hence the error-cancelling effect of the correctly and incorrectly classified pixels is not all that good. The values for χ_3^2 also show that the distribution of correctly and incorrectly classified pixels is nowhere near optimum, but χ_7^2 shows that they are closer to being optimally distributed than randomly distributed. The values for χ_4^2 also show that the correctly classified pixels are closer to being optimally distributed than randomly distributed.

The values for χ_5^2 show that each feature is not randomly classified, although the categories urban and transportation are highly suspect. In all cases, the agriculture category is the least randomly classified even though it is not the most accurately classified or largest category. The values for χ_6^2 show that the ground truth category transportation is highly suspect of randomly occurring in places classified as other categories. This test was used primarily to determine if the number of misclassified pixels for a particular category were distributed or proportional to the population of the other ground truth categories. The values of χ_8^2 indicate that the number of correctly and incorrectly classified pixels for each category are not randomly distributed, but again the urban and transportation categories are suspect.

The values for χ^2_9 and $\%c$ show that the contingency table distribution does not indicate independence of the ground truth and classification results, but a 45 percent "correlation" is nothing to be proud of either. Hence, it appears that the classification performance was rather dismal for this test site. Although Table 8 indicates that MLMCM had the best performance, the chi-squared value is still too large when measured against the best possible performance of all the techniques.

There may be several reasons why the performance of the techniques is lower than expected. The first is that the best season may have not been chosen for those techniques that used only one set of data. Secondly, the test site is rather small (33386 pixels) as test sites go. The observation was made that the majority of classification errors occurred at the boundary of two or more different features and that the homogeneous areas were classified consistently accurately. Hence, if the test site had been expanded, it is expected that the misclassification would increase linearly and correct classification would increase proportionally to the area. Also better choices of training areas would probably be available. Expanding the site size would also provide a means of checking the stability of the statistics calculated for the 33386 pixel test site.

The discussion of these evaluation criteria and results provides a means of establishing a statistical base for determining the performance of various classification techniques on different types of data sets and for various remote sensing discipline applications. However, these criteria provide relatively little information concerning the goodness of a CM. The tabular results provide information only on how many of each category, whereas a map also provides this information as well as where this information is located. The next section addresses modification of the contingency table to provide information on the spatial complexity of the test site, on where misclassification errors occur, and on how well the CM agree with the GTM.

IV. PROPOSED STATISTICAL TESTS FOR EVALUATING CM AND/OR GTM

Although the previous tests contain relatively little information on the goodness of maps produced by various classification techniques, the tests can be adapted to provide some measure of map goodness. One possible clue as to what approach should be taken to adapt these tests is that CM with identical inventory and classification accuracies can appear quite different visually.

Thus, for two such CM, the best choice would appear to be to select the map whose homogeneous areas and boundaries coincide best with the GTM homogeneous areas and boundaries. A proposed quantitative approach to making this selection is to produce an 8 by 8 contingency matrix to replace each individual element in the contingency table of GTM versus CM. The model used to provide numbers for the contingency matrix in the contingency table is as follows.

Let x_{ij} be the reference sample on the CM and y_{ij} be the reference sample on the GTM at scan i and column j. Let $x_{i-1,j}$ and $x_{i,j-1}$ be two test samples adjacent to the reference sample on the CM at scans and columns i-1,j and i,j-1, respectively, and let $y_{i-1,j}$ and $y_{i,j-1}$ be the corresponding test samples on the GTM.

Several comparisons can be made between the reference and test samples on either the GTM or CM and between corresponding samples on the GTM and CM. For example, a vertical (horizontal) boundary would be indicated on the CM if pixel $x_{i,j}$ belongs to a different class than pixel $x_{i,j-1}$ ($x_{i-1,j}$). The same is true for the GTM if x is replaced by y. A homogeneous pixel area occurs when $x_{i,j}$, $x_{i-1,j}$, and $x_{i,j-1}$ belong to the same class on the CM. The same is also true for the GTM if x is replaced by y. A double boundary occurs when the reference sample disagrees with both test samples on either the CM or GTM. Comparisons also have to be made between the GTM and CM to determine how many agreements there are concerning the three corresponding pixels. In constructing the 8 by 8 matrix, the upper half will contain entries when the reference samples belong to the same class on the GTM and CM, the lower half will contain entries when the reference samples disagree on the GTM and CM, the left half of the matrix will contain entries when either or both of the test samples agree on the GTM and CM, and the right half of the matrix will contain entries when either or both of the test samples disagree on the GTM and CM. A pictorial description of the 8 by 8 contingency matrix is shown in the Figure and an explanation of the column and row labeling, as well as the entry values follows.

The row or GTM label definitions are:

1.1 — The reference samples agree on GTM and CM. There is no feature change in either the scan or column direction (homogeneous pixel area).

1.2 — The reference samples agree on GTM and CM. There is a feature change in the scan direction only (vertical boundary).

GTM/CM	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4
1.1	3	2	2	1	0	1	1	2
1.2	2	3,2	1	2,1	1	0,1	2	1,2
1.3	2	1	3,2	2,1	1	2	0,1	1,2
1.4	1	2,1	2,1	3,2,1	2	1,2	1,2	0,1,2
2.1	0	1,0	1,0	2,1,0	3	2,3	2,3	1,2,3
2.2	1,0	1,0	2,1,0	2,1,0	2,3	2,3	1,2,3	1,2,3
2.3	1,0	2,1,0	1,0	2,1,0	2,3	1,2,3	2,3	1,2,3
2.4	2,1,0	2,1,0	2,1,0	2,1,0	1,2,3	1,2,3	1,2,3	1,2,3

Reference Samples Agree

Reference Samples Disagree

Test Samples Agree

Test Samples Disagree

Figure. 8 by 8 contingency matrix.

1.3 — The reference samples agree on the GTM and CM. There is a feature change in the column direction only (horizontal boundary).

1.4 — The reference samples agree on the GTM and CM. There is a feature change in the scan and column directions (double boundary).

The row definitions of 2.1, 2.2, 2.3, and 2.4 are identical to the above, except that the reference samples on the GTM and CM disagree. The column or CM labels 1.1, 1.2, 1.3, 1.4 are identical to those previously defined, but 2.1, 2.2, 2.3, 2.4 refer to test sample disagreements. The entry values in the matrix range from zero to three. The left half of the matrix contains the number of agreements on the GTM and CM concerning the three pixel locations, and the right half contains the number of disagreements. For example, every time a 1.1 condition is encountered on the GTM and CM, all three pixels are in agreement and a three is added to the sum (which is initially zero for all elements) contained in matrix element 1.1, 1.1. Notice that 1.1, 2.1 and 2.1, 1.1 are impossible situations and always contain zero. In the case where 1.4, 1.4 is encountered, there can be 3, 2, or 1 agreements which are added to the sum in matrix element 1.4, 1.4 and there can be 0, 1, or 2 disagreements, respectively, which are added to the sum of matrix element 1.4, 2.4.

To construct the contingency table using the 8 by 8 contingency matrix, it is necessary to use only half of the 8 by 8 matrix for each table element. Thus, for the diagonal elements of the table, only the upper half of the matrix is used because the bottom half will be all zeros. For the off-diagonal elements of the table, only the lower half of the matrix is used because the top half will be all zeros. Hence, each element in the contingency table is replaced by a 4 by 8 contingency matrix. Notice that the original values of the single element contingency table can be obtained by adding the right half of the 4 by 8 matrix to the left half for each table entry, computing the sum of all of the elements of the resulting 4 by 4 matrix, and dividing by three. Therefore, the contingency matrix not only contains the same information as the contingency table, but it also contains a considerable amount of information related to the structure of the CM.

There are several types of map structure information that can be obtained from the 4 by 8 contingency matrices. By adding the right half of the 4 by 8 matrix to the left half and dividing all of the elements of the resulting 4 by 4 matrix by three, the 4 by 4 matrix will contain the number of homogeneous pixels, vertical boundaries, horizontal boundaries, and double boundaries on the diagonal elements for correctly classified pixels. The off-diagonal elements of the matrix contain the number of errors where feature changes occurred on the CM, but did not occur on the GTM or vice versa. Previous work done on identifying the major source of classification errors has indicated that the majority of misclassification occurs at a boundary between two or more different features. The matrix will help narrow down what type of boundaries produce the most errors. By not adding the right half to the left half of the 4 by 8 matrix, it is possible to determine for those elements having only two possible values, the number of events having each value. This is not possible with the matrix elements that can have three values.

By comparing a GTM with itself, the contingency table will contain only diagonal elements and these diagonal elements will contain 4 by 4 matrices (which are the upper left quarter of the original 8 by 8 matrices) that are themselves diagonal. These 4 by 4 diagonal matrices provide a means of determining the spatial complexity of each feature in terms of the number of observed homogeneous pixels and various types of pixel boundaries. By adding all of the 4 by 4 diagonal matrices, a general measure of spatial complexity can be obtained for the entire GTM independent of feature. These measures are the expected distributions that can be used in various χ^2 tests for comparing the CM (observed distributions) with the GTM to determine how well the spatial complexities agree. Thus, the comparing of spatial complexities provide a means of selecting the best CM from several maps that have similar inventory and classification accuracies. Comparisons can also be made between the various CM as well as comparing a CM with itself, if that type of information is desired.

Table 10 shows the contingency matrix for comparing the GTM with itself. For the urban category, which contains 276 pixels, there were 66 urban pixels (23.91 percent of the urban pixels) that had an urban pixel directly above it (previous scan, same column) and an urban pixel directly to the left of it (same scan, previous column). There were also 29 urban pixels (10.5 percent) that had an urban pixel directly above it and no urban pixel directly to the left (vertical boundary), 23 urban pixels (8.33 percent) that had an urban pixel directly to the left of it and no urban pixel directly above (horizontal boundary), and 158 urban pixels (57.24 percent) that had no urban pixels directly above or to the left (double boundary). In describing the features on the GTM, it could be said that the urban feature is 23.9 percent homogeneous, transportation is 4 percent homogeneous, agriculture is 77.5 percent homogeneous, forest is 73.1 percent homogeneous, and water is 18.3 percent homogeneous. There is a correspondence between the homogeneity of a feature and the feature classification accuracy in that the more homogeneous features appear to be more accurately classified.

TABLE 10. GTM/GTM CONTINGENCY MATRIX FOR EACH FEATURE

GTM \ GTM	u				t				a				f				w				Category Percent
	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	
u	66																				23.91
	29																				10.50
			23																		8.33
			1	58																	57.24
t					34																4.03
					1	52															18.90
						1	21														14.33
							5	37													63.62
a								74	22												77.51
									9	73											10.16
										6	92										7.22
											4	89									5.10
f											1	60	84								73.07
													23	38							11.07
														17	96						8.15
															16	93					7.69
w																1	24				18.26
																	1	86			27.39
																		1	19		17.52
																			2	50	36.81

Table 11 shows the contingency matrix for all of the GTM features combined.

TABLE 11. GTM/GTM CONTINGENCY MATRIX FOR ALL FEATURES

GTM / GTM	All Features				Category Percent
	1.1	1.2	1.3	1.4	
1.1	2 37 30				71.07
1.2		37 78			11.31
1.3			27 51		8.23
1.4				31 27	9.36

The table indicates that the entire map is 71.1 percent homogeneous, which corresponds very closely with the classification accuracies presented in Tables 2 through 6. Thus, it appears that the homogeneity percentage for the GTM could be used as a good estimate of expected minimum classification accuracy. Table 11 also indicates that it may be worthwhile to consider using spatial information in the classifier because 91 percent of the pixels belong to the same feature as the previous pixel in the same scan or same column.

Table 12 shows the contingency matrix of MLCM/GTM for each feature. The diagonal and row percentages for correct classification are obtained by ratioing the diagonal elements and row sums of the diagonal matrices in Table 12 with the elements in Table 10. For forest, the diagonal percentages show that for 64.2 percent of the time, the reference pixel was corrected classified when the previous pixel in the same scan and same column were also correctly classified as forest. For the case where the reference pixel and the previous pixel in the same column were correctly classified as forest, but the previous pixel was categorized as belonging to another feature, the success was only 16.8 percent. In the case of a horizontal boundary for forest, the success in correct classification was only 13.5 percent, and for the case of a double boundary for forest the success was only 11.4 percent. This situation seems to be typical for large homogeneous areas, indicating that the interior pixels tend to be more correctly classified than the transition or boundary pixels between two or more features. If the constraint is removed that the previous pixels in the same scan and same column on the CM have to agree with the class configuration of the corresponding pixels on the GTM, then the row percentages show that for forest and when there is no feature change in the previous pixels on the GTM, the reference pixel on the CM is correctly classified 86.2 percent of the time.

The situation appears to be different for highly linear features such as transportation/communication (t). In this case, the diagonal and row percentages are higher when a feature change is present in the previous pixels. This is probably due to high data contrast between roadways and power line right of ways versus forested areas.

TABLE 12. MLCM/GTM CONTINGENCY MATRIX FOR EACH FEATURE

GTM \ MLCM		u				t				n				f				w				Diagonal Percentages	Row[a] Percentages
		1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4		
u	1.1	7	12	5	1	3	2	1	5	4	1	4	4	5	7	1	4	0	0	0	0	10.69	37.87
	1.2	2	2	2	2	0	3	0	2	4	2	3	1	2	1	1	2	0	0	0	0	6.89	27.58
	1.3	3	2	1	3	1	3	0	1	1	3	3	1	0	0	0	1	0	0	0	0	4.34	39.13
	1.4	4	5	2	6	3	6	5	12	1	1	1	1	63	22	12	12	0	0	0	1	3.79	10.75
t	1.1	0	2	1	4	0	2	0	4	1	1	2	1	4	0	7	5	0	0	0	0	0.00	17.67
	1.2	2	5	4	7	4	8	4	13	3	3	1	3	48	10	17	14	0	0	0	2	5.26	17.10
	1.3	4	4	5	12	7	11	7	10	5	4	1	0	19	7	12	12	0	0	0	1	5.78	28.92
	1.4	15	20	16	28	14	29	13	40	39	20	25	26	133	33	44	38	0	2	0	2	7.44	17.87
n	1.1	371	475	331	546	27	91	74	206	3571	765	503	264	34	35	35	92	0	0	0	2	48.11	68.75
	1.2	39	76	47	116	14	34	35	92	96	180	18	50	45	37	47	42	0	2	0	3	16.49	35.35
	1.3	35	43	45	78	6	29	13	43	138	29	102	47	18	21	11	33	1	0	0	0	14.73	45.64
	1.4	16	34	11	62	7	21	13	46	38	45	16	40	43	27	25	37	3	3	0	0	8.17	26.83
f	1.1	56	97	75	245	145	311	232	764	50	50	45	88	10 323	1649	1089	798	7	16	6	37	64.18	86.16
	1.2	30	71	54	125	37	111	61	200	56	35	55	49	666	410	149	244	9	21	19	16	16.81	61.07
	1.3	28	43	30	67	26	51	52	126	23	38	18	27	614	176	242	199	5	15	5	17	13.47	68.54
	1.4	22	64	36	93	30	64	50	150	58	41	44	38	393	208	114	193	15	23	9	8	11.39	55.63
w	1.1	0	0	2	4	0	3	1	11	0	0	1	2	4	9	4	20	37	11	13	2	29.83	50.80
	1.2	0	2	3	2	1	8	2	10	0	0	2	2	4	24	5	17	30	60	7	11	32.25	56.06
	1.3	0	1	0	6	1	2	4	7	0	1	0	0	2	5	2	13	17	9	28	19	23.52	61.34
	1.4	0	1	4	15	5	8	11	19	0	0	2	1	36	28	20	27	14	20	12	25	10.00	28.40

a. For correct classifications only.

It also appears that the effect of banding can be observed by examining the diagonal and row percentages change for the 1.2 and 1.3 cases. If the class configuration is preserved on the CM and GTM (diagonal percent), the classification accuracy is higher for 1.2 (vertical boundary). However, if the class configuration is ignored on the CM, the classification accuracy is higher for 1.3 (horizontal boundary on GTM). Both situations are supported by the fact that banding is observed as a horizontal phenomenon produced by data changes in the vertical direction.

Table 13 is a summary of the information in Table 12 for all features. The diagonal and row percentages were obtained by ratioing the diagonal elements and row sums of Table 13 with the elements of Table 11. The total diagonal percentage was obtained by ratioing the sum of the diagonal elements in Table 13 with the sum of the diagonal elements in Table 11. The diagonal and row percentages indicate essentially the same results as previously mentioned. However, it is interesting to compare three types of classification accuracy based upon different constraints. For MLCM Table 2 shows that if the total number of pixels for each feature on the CM (regardless of where they occur) are compared with the total number of pixels for each feature on the GTM, then the inventory accuracy is 81.94 percent. If the constraint is added that the CM features pixels are correct if they agree with the GTM feature pixels at the same location, then the classification accuracy is 71.67 percent. If a constraint is added that feature changes on CM and GTM have to agree together with the correctly classified pixels, then a measure of the map accuracy is 45.77 percent as indicated by the total diagonal percentage in Table 13.

TABLE 13. MLCM/GTM CONTINGENCY MATRIX FOR ALL FEATURES

MLCM / GTM	All Features				Diagonal Percentages	Row Percentages
	1.1	1.2	1.3	1.4		
1.1	13938	2439	1610	1069	58.73	80.30
1.2	818	660	180	317	17.46	52.27
1.3	779	227	380	278	13.81	60.48
1.4	463	307	159	304	9.72	39.43
	Total Diagonal Percentage				45.77	

APPROVAL

EVALUATION CRITERIA FOR SOFTWARE CLASSIFICATION INVENTORIES, ACCURACIES, AND MAPS

By Robert R. Jayroe, Jr.

The information in this report has been reviewed for security classification. Review of any information concerning Department of Defense or Atomic Energy Commission programs has been made by the MSFC Security Classification Officer. This report, in its entirety, has been determined to be unclassified.

This document has also been reviewed and approved for technical accuracy.

for W. T. POWELL
Director, Data Systems Laboratory

☆U.S. GOVERNMENT PRINTING OFFICE 1976—740-049/19 REGION NO. 4