GT Index: A Measure of the Success of Prediction

GÖKSEL TÜRK*

Environmental Research Group, EARTH, Inc., 1044 E. Halcyon, Tucson, Arizona, 85719

In order to evaluate the results of a remote sensing study, ground truth data often are collected and a cross-classification table is prepared. However, the commonly used Percent Correct Classification criterion may be misleading. An operationally meaningful measure of success of prediction is proposed. An example using corn blight data demonstrates the value of this new index and casts some doubts on earlier conclusions based on the percent correct classification criterion.

1. Introduction

In order to evaluate the results of a remote sensing (RS) study, ground truth (GT) data are often collected and a cross-classification table is arranged. If all items are correctly classified, then nondiagonals of the cross-classification table will be empty. However, this degree of accuracy is rarely, if ever, obtained. It therefore becomes desirable to define a criterion for measuring the success of a study. The percent correct classification (PCC) has been used most commonly. Unfortunately, it may not measure what it seems to measure.

About a century ago, Finley (1884) evaluated his tornado predictions using PCC and gave himself an impressive 98.18%. His data are summarized in Table 1.

Had all cases been classified as "no tornado," an even higher score would have resulted, because of the many cases of "no tornado" in the sample. This problem was soon pointed out, and led to debates over the issue of how to evaluate the success of prediction.

In a letter to the editor of *Science*, a solution to the controversy was offered by the famous scientist, philosopher, and logician C. S. Peirce (1884). He assumed that the observed cross-classification table (Table 1) is a mixture of two cross-classification tables. One belongs to an infallible predictor and the other is the

TABLE 1 Cross-Classification Table for Finley's Tornado Predictions and Occurrences During April 1884

	Occurrence		
Prediction	Tornado	No Tornado	Totals
Tornado	11	14	25
No Tornado	3	906	909
Totals	14	920	934

Finley's tornado prediction PCC score 98.18%
"No tornado" prediction PCC score 98.50%

*Permanent address: Yidiz, Posta Cad. 22/13, Gayrettepe, Istanbul, Turkey.

0034-4257/79/010065+11$01.75

result of predictions made by "an utterly ignorant person" who guesses randomly using some definite proportions of "yeses" and "noes." Therefore, the fraction expressing the predictions of the infallible one should be a measure of the "approximation to complete knowledge." Regarding the correct predictions made by an utterly ignorant person, he wrote "the second witness may know *how often* he ought to answer 'yes,' but I give him no credit for that, because he is ignorant *when* he ought to answer 'yes'" (original emphasis).

Most classification programs used in pattern recognition (e.g., computer classification of RS data) in fact have some provision to adjust their answer to the distributions of the GT categories (i.e., they will say "yes" as often as they ought to). Now should we give them credit for this, or separate the infallible fraction? In full agreement with Peirce, this author thinks we should separate the infallible part; since what is important is the skill in predicting and not *apparent* accuracy.

2. The Prediction Matrix

Following Peirce's basic conceptualization, let us assume the observed cross-classification to be a mixture of infallible and random predictors.[1] For the infallible predictor, the prediction matrix (i.e., transition probability matrix) will be the identity matrix. On the other hand, the random predictor will have an assignment probability R_j for jth RS category (i.e., items, regardless of their GT category, have an assignment probability R_j for the jth RS category). If the infallible predictor is used θ percent of the time, then the random predictor will be used $(1-\theta)$ percent of the time. The observed cross-classification table will be the sum of θ times the infallible prediction matrix and $(1-\theta)$ times the random prediction matrix (Fig. 1). Therefore:

$$P_{ij} = \begin{cases} \theta + (1-\theta)R_j & \text{for } i = j \text{ (diagonals)} \\ (1-\theta)R_j & \text{for } i \neq j \text{ (nondiagonals)} \end{cases} \tag{1}$$

where θ is the fraction of the time during which the infallible predictor is used. R_j is the probability for any item, regardless of its origin (i.e., its GT category), to be assigned into jth RS category by the random predictor. This may be the correct category, as well. P_{ij} is the probability of joint occurrence for cell (i,j).

This model turns out to be equal to the Markov model studied by Goodman (1964), who provided techniques to test the goodness-of-fit and to estimate the model parameters, namely θ and R_j.

For many applications, it may be too restrictive to require that all categories have the same predictive success (i.e., the same θ). Furthermore, whether all categories are predicted with the same degree of success is an interesting question. Thus, if the common θ requirement (i.e., the fraction of the time during which the infallible predictor must be used equally for all categories) is relaxed, then the model becomes:

$$P_{ij} = \begin{cases} \theta_i + (1-\theta_i)R_i & \text{for } i = j \\ (1-\theta_i)R_j & \text{for } i \neq j \end{cases} \tag{2}$$

where θ_i is the θ value for the ith GT category and all other symbols as previously defined.

[1] C. S. Peirce (1884) states that he had an extension to the case having more than two categories, but the author was not able to locate this extension if it was ever published.

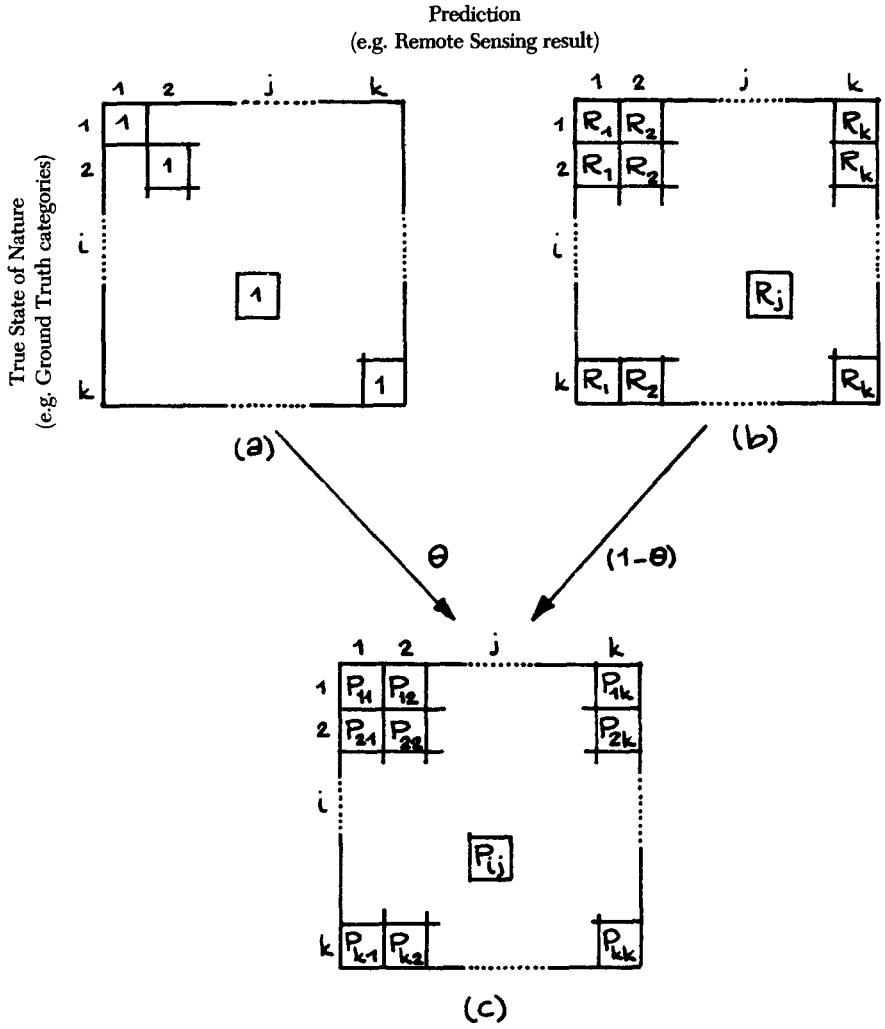

FIGURE 1. Prediction matrix of infallible predictor's (a) and of random predictor's (b) produces observed cross-classification table (c).

Notice that the nondiagonals of the observed cross-classification table are *not* a mixture but are due to random prediction only. If the data fit the model, they should satisfy the "quasi-independence" of nondiagonals.

"Quasi-independence" is essentially an extension of the independence concept of probability (see Goodman, 1968, and others). Similar to the multiplication theorem of probabilities, joint occurrence probability P_{ij} for cell (i,j) which belongs to the quasi-independent set (i.e., the remaining cells after nonrandom cells are deleted from the table) should be equal to the product of a row parameter a_i and a column parameter b_j, which both are normalized to one. That is:

$$P_{ij} = a_i \cdot b_j \tag{3}$$

Goodman (1968) and others provide a method of "iterative proportional fitting" (See Appendix A for a brief description) to estimate parameters a_i and b_j. Then a

goodness-of-fit test can be carried out using expected versus observed frequencies.

If correct classifications do not contain any "lucky guesses," as the PCC criterion assumes, then the random part of the observed cross-classification table can be thought of as due to random assignment of misclassified items. If R_j retains its meaning as defined above, then the probability of assignment into the jth RS category, given the fact that the item is misclassified (i.e., assignment into correct category is excluded), will be $R_j/(1-R_i)$, which follows directly from the definition of conditional probability. Consequently, the model that does not allow any provision for "lucky guesses" will be:

$$P_{ij}^{*} = \begin{cases} A_i & \text{for } i = j \\ (1-A_i)\cdot(R_j/(1-R_i)) & \text{for } i \neq j \end{cases} \quad (4)$$

where A_i the proportion of correctly classified item from ith GT category; P_{ij}^{*} the joint occurrence probability for cell (i,j) for this model; and R_j and R_i are as previously defined.

3. The GT Index[2]

We have already argued against giving credit to the "lucky guesses," and θ_i's were indicated as a good measure of the success of prediction. The model given by Eq. (2) is considered the true underlying process. If Eq. (4) is set equal to Eq. (2), by simple arithmetic, it can be shown that:

$$\theta_i = (A_i - R_i)/(1-R_i) \quad (5)$$

Substituting the estimate of A_i and R_i, an estimate of θ_i can be obtained. This estimate, $\hat{\theta}_i$, is multiplied by 100, and the result is dubbed as the GT index. Equation (5) provides another interpretation for θ_i; that is, it is a normed index of deviation from a completely random assignment:[3]

$$\theta_i = \frac{\text{(Actual correct classification)} - \text{(lucky guesses)}}{\text{(100\% correct classification)} - \text{(lucky guesses)}}$$

This definition of θ_i, in fact, is nothing more than the proportion of agreement *corrected for chance agreement.*

Obviously, the estimate of A_i multiplied by 100 is equal to the PCC. Thus, it follows that:

$$A_i = \theta_i + (1-\theta_i)\cdot R_j \quad (6)$$

indicating that A_i values (and consequently PCC) are inflated by a chance component that is proportional to $(1-\theta_i)$, the fraction of those times during which the random predictor is used. The more the random predictor is used, the greater the inflation in PCC value. Therefore, the PCC has a built-in mechanism that covers up some of the failure in the prediction process. This fact is demonstrated in Section 4.1 using a published RS result.

In the ideal case (i.e., when all items are correctly classified), all nondiagonals of the cross-classification table will be empty. Thus, "the more items on diagonals, the better" can be adopted as a guiding principle.

All misclassification must be random. If this is not the case, then the classification procedure is confusing some cate-

[2]Short for "Ground Truth" index. Its being the initials of the author seems accidental.

[3]If the RS result is more in error than that obtained by chance (i.e., random assignment), then θ_i will be negative.

gories with others more often than would be expected by chance. This means that the classification can be *improved* by reducing the confusion between these categories. How this reduction may be accomplished is another matter. Sometimes a change in the classification rule may be enough, and sometimes an improvement yet to be realized in the remote sensing system may be required.

Randomness of misclassification is cardinal to the meaningfulness of the proposed model. If the majority of current remote sensing studies do not satisfy this requirement, it should be considered as a deficiency of current practice rather than as a defect of the proposed model. The author is perfectly aware of the negative implications of this judgement for most of the RS studies done so far, because the validity of their results might be in doubt due to the possibility of hitherto unnoticed problems stemming the use of PCC criterion. One of the important purposes of this paper will have been accomplished, however, if it prevents future studies from having such defects and if it generates interest for studying the problems of evaluating the results of RS studies and their solutions. This paper is only a modest beginning contribution.

Biases in misclassification indicate that there is more information yet to be uncovered. The "structure of confusion" will show which categories should be more clearly distinguished from others. By finding a way to discriminate better between these categories the classification will be improved.

Obviously, there will be cases in which randomness of misclassification cannot be obtained despite every reasonable effort of researchers. This has to be accepted philosophically. However, when the requirement of "randomness of misclassification" is not satisfied, calculation of the GT index may be inappropriate since its value will be biased in some unknown manner.

If we can assume that in a given GT category a certain proportion of items will *always* be correctly identified and the rest will be distributed randomly into all categories including the correct category, then the same model is obtained. The GT index is the estimate of the proportion of those items that will *always* be classified correctly. This provides another conceptualization for the model.

4. The Blight of Corn Leaf and of PCC

Bauer et al. (1971) studied the possibility of corn leaf blight identification using 1/60,000 scale color infrared film with 10× magnification with the help of a photographic key. Airphoto results were checked against corn leaf blight severity classes obtained by field survey.

Using a FORTRAN program written by author, their data were reanalyzed. Reanalysis of their Table 3 (rearranged as Table 2 of this paper) demonstrates the problem associated with the PCC index and the value of the GT index. It also indicates a hitherto undetected problem in their study. Of course, no criticism of their valuable work is intended. Since the proposed GT index was not available to them and deficiency of the PCC index was not clarified at that time, none of the following remarks can and should be construed against their otherwise careful study.

Their results, shown in Table 2, indicate PCC values above 70% for any cate-

TABLE 2 Bauer et al.'s Observed Frequencies for Corn Leaf Blight Study. Classified as (Remote Sensing Result)

Ground Truth Classes	Others	Slight and Mild	Moderate	Very Severe	Severe	Total	PCC
Others	148	1	8	2	0	159	93.0
Slight & mild	0	50	15	3	0	68	73.5
Moderate	1	6	39	7	1	54	72.2
Severe	1	0	6	25	1	33	75.8
Very severe	1	0	0	1	6	8	75.0
Totals	151	57	68	38	8	322	83.2

Source: After Bauer et al. (1971), Table 3.
Note: In Bauer et al.'s Table 3, "others" (i.e., other than corn fields) follows the "very severe corn leaf blight" category. Here "others" is considered as the "no corn leaf blight" category and made the first row and column of the table.

gory and an overall PCC above 80%. This looks rather impressive. However, the hypothesis of no corn field in the area (so there is no corn leaf blight either) will result in an overall PCC value of 48%. Considering the effort required for implementation of this hypothesis, 48% PCC is a respectable figure. Of course, such a strategy will be self-defeating for the purpose of corn leaf blight detection. The point is, PCC value strongly depends upon the relative abundances of each category within the sample at hand.

Since there is an apparent tendency towards the diagonals, the "quasi-independence" of nondiagonals hypothesis was tested first. Expected frequencies obtained by iterative proportional fitting are remarkably close to the observed ones and a chi-square test did not reject the hypothesis[4] (See Table 3).

Bauer et al. (1971) remarked that "incorrect classification of the corn fields was primarily into adjacent severity classes." The implication is that the most similar classes are confused with each other more often than not. However such an apparent tendency in Table 2 is an artifact due to the decreasing order of marginal frequencies. In the face of the good fit to the quasi-independence model, their contention is untenable.

The reviewer pointed out that when arguing the validity of Bauer et al.'s contention, fairness requires more concern with Type II errors (i.e., accepting quasi-independence when it is false), and at the 10% level, quasi-independence can be rejected. However, the departure from quasi-independence is mostly due to the confusion between two dissimilar categories; namely, between non-corn areas and areas of very severe corn blight. In fact, this single cell makes up more than half of the χ^2 value of 20.941.

An analysis of residuals following Haberman (1973) indicates the same point (see Table 3). Confusion between these two categories is understandable if most corn plants were destroyed. However, this can hardly be considered as "incorrect classification...into adjacent (sic) severity classes."

Using Eq. (5) and estimates of A_i's and

[4]The method of iterative proportional fitting is suitable when "quasi-independence" is assumed or tested. Therefore, one should be more concerned with not rejecting the quasi-independence when it is true (i.e., not to commit Type I error) if he wants to calculate GR indexes as often as possible. Thus, the confidence level is chosen accordingly. Sometimes, other considerations may require more concern with Type II errors.

TABLE 3 Comparison of Observed Frequencies with Expected Frequencies Under Quasi-independence. Standardized Residuals[a] are Given in Parentheses

	Others	Slight and Mild	Moderate	Severe	Very Severe
Others					
Observed frequency		1.00	8.00	2.00	0.00
Expected frequency		1.16	7.70	1.87	0.26
		(−0.15)	(0.11)	(0.09)	(−0.51)
Slight & mild					
Observed frequency	0.00		15.00	3.00	0.00
Expected frequency	0.78		13.48	3.28	0.46
	(−0.88)		(0.41)	(−0.15)	(−0.68)
Moderate					
Observed frequency	1.00	6.00		7.00	1.00
Expected frequency	1.78	4.66		7.51	1.06
	(−0.58)	(0.62)		(−0.19)	(−0.05)
Severe					
Observed frequency	1.00	0.00	6.00		1.00
Expected frequency	0.37	0.97	6.44		0.22
	(1.04)	(−0.99)	(−0.17)		(1.66)
Very severe					
Observed frequency	1.00	0.00	0.00	1.00	
Expected frequency	0.08	0.21	1.38	0.34	
	(3.27)	(−0.46)	(−1.17)	(1.15)	

χ^2 due to diagonals	672.465**	with 5 d.f.	
χ^2 due to nondiagonals	20.941	with 11 d.f.	Quasi-independence *not* rejected.
χ^2 Total	693.406**	with 16 d.f.	

[a]Standardized Residual = (Observed Frequency − Expected Frequency)/$\sqrt{\text{Expected Frequency}}$. Standardized residuals are approximate standard normal deviate (z), and chi-square associating with each cell is square of its standardized residual.

R_j's, estimates of θ_i's and then the corresponding GT indexes were obtained. Table 4 provides a comparison of GT indexes with corresponding PCC values.

This comparison indicates the following facts:

1. PCC figures are inflated to varying degrees as expected.

2. Inflation in PCC of the "moderate corn leaf blight" category (Severity Class 3) is extremely high. In fact, the GT index for this category is about one fifth of the PCC value. This category represents some sort of dividing line between cases where infection is confined to the lower leaves (i.e., less than 10 percent of upper leaf area is affected) and cases where most of the upper leaf area is affected (i.e., more than 30 percent of upper leaf area). This may explain the difficulties experienced by the interpreter. PCC value indicates about the same level of diagnostic ability, whereas the GT index clearly shows the difficulty experienced due to the borderline characteristic of this category.

TABLE 4 Comparison of GT Indexes with PCC Values

	Categories				
Criterion	Others	Slight and Mild	Moderate	Severe	Very Severe
PCC	93.08	73.53	72.22	75.76	75.00
GT index	92.80	70.54	15.05	71.01	74.41
Inflation[a]	0.28	2.99	57.17	4.75	0.59
Relative inflation[b]	0.30	4.24	379.75	6.68	0.79

[a]Inflation = PCC minus GT index
[b]Relative inflation = Inflation as percent of GT index

3. "If *only* the lower leaves are severely damage after the milk stage is reached, yields are *reduced less*" (Bauer et al., 1971, pp. 694–95, emphasis added). Thus, recognition of the Severity Class 3 may be important for plant protection purposes. If so, the usefulness of the proposed identification technique should be seriously questioned.

4. The data do not contain any "noninfected" areas, and the first two severity classes are lumped together. If the separation of "infected" and "noninfected" areas, or the separation of the first two classes is desired for one reason or another, then the GT indexes should be recalculated using appropriate cross-classification tables. These new values will not necessarily be the same, since lumping the categories together inflates the number of correctly classified individuals. This dependency of its value upon the class definition is not a defect but a virtue of the proposed GT index. It is, in fact, desirable that a measure of prediction success reflect the classes *as defined for the data*. The fact that the definition of the classes can affect the degree of prediction success naturally means that careful attention should be given to the class definitions in light of the expected uses of the final conclusions. Obviously, one can increase the prediction success by combining the categories that are hard to discriminate from each other, but the price of this apparent increase in prediction success is to have less meaningful categories for the intended uses.

5. Discussion and Conclusion

An operationally meaningful measure of the success of prediction is proposed. It is interpreted as the fraction of the time during which infallible prediction is made for a given GT category. It can also be thought of as the proportion of individuals in a given GT category that will be correctly and surely identified by the RS under consideration.

If there are some biases in misclassification (i.e., some categories are confused with each other more often than what is expected by chance), then there is some information yet to be uncovered. The "structure of confusion" that is indicated by biases will provide some guidance for the improvement needed in classification by indicating those categories that should be better differentiated from each other. How this can be accomplished is another matter.

When biases exist, the data will not satisfy the requirement of "quasi-independence" and calculation of the GT

index may be inappropriate, since its estimate will be biased an unknown manner. Of course, it is still possible to consider the GT index as a "proportion of agreement corrected for chance agreement," provided chance agreement is defined as the random process defined by row and column marginals of off-diagonals. This is because iterative proportional fitting provides an estimate for a random matrix with given row and column marginals. But nonrandomness of off-diagonals implies that another process is at work, and that is why calculation of the GT index may not be appropriate.

When randomness of off-diagonals is accepted, the GT index gives the proportion of agreement corrected for chance agreement; thus, it is independent from the composition of the sample at hand (i.e., not affected by relative abundances of categories within the sample.) This is its main virtue over PCC, which depends upon the sample composition. And this property of the GT index provides an opportunity to study the effects of certain factors on the predictive ability of a given RS procedure. Investigators can profitably study the change in values of the GT index with regard to the factors under consideration.

For instance, if separability of categories depends upon the time of the year, then, for any given category, the GT index will be a function of time, and using this information the best time for RS operation can be determined. Similarly, different classification rules can be evaluated by comparing corresponding GT indexes. Obviously, differences to be evaluated can be any other conceivable factor (e.g., type of sensor, weather condition, stage of plant growth) or combination of factors.

In short, the GT index opens up possibilities for studying the following questions, which are important in an RS context:

1. What is the degree of prediction for each category?
2. Are these successes about the same for all categories?
3. What is the effect of a certain factor or combination of factors on the success of prediction for each category?

These are the promises of GT index methodology for RS studies. Of course, the proof of the pudding is in the eating. How much it will live up to its promises can only be determined after accumulation of enough empirical evidence about its performance. The example above simply indicates that the use of the GT index can significantly contribute to RS studies. Whatever the final judgement on the value of the GT index, if this paper can facilitate the recognition of the problem associating with the PCC criteria, of the need for better evaluation techniques for the result of RS studies, and of the importance of the randomness of off-diagonals, then it will accomplish its main purposes.

Appendix A

Here, the calculation of expected frequencies of quasi-independent cells, f_{ij}^{*}, by iterative proportional fitting is described for the case where the diagonals are deleted. This procedure can be extended for rectangular (rather than square) tables or for the case where some other specified set of cells (rather than diagonals) have been deleted. For details and further reference, see Goodman's paper mentioned in the text.

First, replace the entries in the deleted cells (here, the diagonals) of the table by zero; and obtain new marginal totals of rows, f_i^0, and of columns, f_j^0.

Next, calculate two sets of intermediate quantities, U_i's and V_j's, which will be used later in calculating f_{ij}^* and row and column parameters, P_i and R_j (P_i and R_j correspond to a_i and b_j of Eq. (3) of the text; however, b_j corresponds to R_j throughout the paper, so this symbolism is preferred).

Calculation of U_i's and V_j's by iterative proportional fitting is as follows:

Step 1. Set initial values of U_i's equal to corresponding marginal totals corrected for deleted cells; that is:

$$U_i^{(0)} = f_i^0 \qquad \text{for all } i\text{'s}$$

Step $2m$. ($m = 1, 2, \ldots$.) i.e., even-numbered steps

$$V_j^{(2m-1)} = f_j^0 \Big/ \left[U_\cdot^{(2m-2)} - U_j^{(2m-2)} \right]$$

where $U_\cdot^{(2m)} = \sum_{i=1}^{k} U_i^{(2m)}$ and superscripts indicate the iteration number; k is the number of rows (columns).

Step $2m+1$. ($m = 1, 2, \ldots$.) i.e., odd-numbered iterations

$$U_i^{(2m)} = f_i^0 \Big/ \left[V_\cdot^{(2m-1)} - V_i^{(2m-1)} \right]$$

where

$$V_\cdot^{(2m-1)} = \sum_{j=1}^{k} V_j^{(2m-1)}$$

The iterative steps are continued for $m = 1, 2, \ldots$ until the desired accuracy is obtained. Then f_{ij}^* and R_j are calculated as follows:

$$f_{ij}^* = U_i \cdot V_j$$

and

$$R_j = V_j \Big/ \sum_{j=1}^{k} V_j$$

Since both observed and expected frequencies are known, a chi-square, χ^2, or likelihood-ratio, G^2, test can be carried out. Both statistics have an asymptotic chi-square distribution with $(k^2 - 3k + 1)$ degree-of-freedom. In general, degree-of-freedom is equal to degree-of-freedom for the complete table minus number of deleted cells. Test statistics are:

$$\chi^2 = \Sigma (f_{ij} - f_{ij}^*)/f_{ij}^*$$

$$G^2 = 2\Sigma \log(f_{ij}/f_{ij}^*)$$

where the summation is over quasi-independent (nondeleted) cells (here, over the nondiagonals).

Acknowledgments

All calculations were done by a CDC 6400 computer at the University Computer Center of the University of Arizona. The reviewer contributed to the improvement of the presentation.

References

Bauer, M. E., Swain, P. H., Mroczynski, R. P., Anuta, P. E., and MacDonald, R. B. (1971), Detection of southern corn leaf blight by remote sensing techniques. *Proc. 7th Int. Symp. Remote Sens. Environ.* 17-21, May 1971, Vol. I, pp. 693–704.

Finley, J. P. (1884), Tornado predictions. *Amer. Meteorol. J.* 1:85–88.

Goodman, L. A. (1964), The analysis of persistence in a chain of multiple events. *Biometrika* 51:405–411.

———, (1968), The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with and without missing entries. *J. Amer. Stat. Assoc.*, 63:1091–1131.

Haberman, S. J. (1973), The analysis of residuals in cross-classified tables, *Biometrics* 29:205–220.

Peirce, C. S. (1884), The Numerical Measure of the Success of Predictions. *Science* 4:453–454.

Received 30 November 1976; revised 21 April 1978.