L. Daniel Maxim
Leigh Harrington
Everest Consulting Associates, Inc.
Princeton, NJ 08550

The Application of Pseudo-Bayesian Estimators to Remote Sensing Data: Ideas and Examples

New estimators may offer improvements over conventional alternatives in the analysis of confusion matrices.

Abstract: *Analyses of data from contingency tables is frequently required in remote sensing applications. In conventional procedures, requisite probabilities are estimated using the principle of maximum likelihood. There are, however, recent improvements to these methods. This paper examines, with several numerical illustrations, the application of pseudo-Bayesian methods to contingency-table analysis of problems in remote sensing. Additionally, a procedure is proposed and illustrated for generating reasonable prior distributions based upon the maximum-entropy concept. This procedure enables the experimentor to reflect partial prior knowledge within a rigorous and (given computer access) easy-to-use framework. The principal contribution of this effort is the organization and presentation in a unified manner of several methodologies which have been developed separately in the literature.*

Introduction

Remote sensing applications often involve the analysis of data from bi- or multivariate contingency tables. One of the most common situations is in the development of "confusion matrices" to characterize the performance of remote sensing systems for detection and identification tasks. Table 1a, for example, shows raw data from Ulaby *et al.* (1980) for crop classification using both like- and cross-polarized L-band radar imagery of an area in Huntington County, Indiana. In this illustration, radar imagery of fields categorized as woods, pasture, corn, or soybeans was analyzed to develop classification decision rules. When applied to a ground truth calibration data set, these classification decision rules produced the confusion matrix or contingency table shown in Table 1a. Thus, of 105 total fields, four were correctly identified as woods, four of 40 soybean fields were misclassified as corn, etc. Let i and j ($i = 1, \ldots ,r$; $j = 1, \ldots ,r$) be indices representing crop type, x_{ij} be the number of fields of type i classed as type j, $x_{i\cdot} = \Sigma_j\, x_{ij}$ the total number of fields of type i in the sample, $x_{\cdot j} = \Sigma_i\, \mathbf{x}_{ij}$ the total number of fields classified into category j, and $N = \Sigma\Sigma\, x_{ij}$ be the total number of fields in the sample. These data are used for illustrative purposes throughout the paper because they serve as a concrete example and also because the small dimensionality of the problem (a 4 by 4 matrix) enables the reader to follow the computations with ease.

A related remote sensing application that also involves the use of contingency table analysis is the problem of evaluating map accuracy. Relevant papers that address this application are those of van Genderen and Lock (1977), van Genderen *et el.* (1978), Ginevan (1979), Hay (1979), and an excellent and rigorous contribution by Card (1982). The rows in the contingency table represent the true classification of cells/quadrats/pixels/objects, while the columns represent the assigned categories displayed in the map for these same cells. With some obvious adjustments to interpretation, all that follows in this paper is relevant to the map accuracy evaluation problem.

0099-1112/83/4905-0649$02.25/0

TABLE 1. DATA FROM AN EXPERIMENT BY ULABY *et al.* (1980)

		ASSIGNED CATEGORY				
		Woods	Pasture	Corn	Soybeans	Subtotal
(a) RAW DATA, x_{ij}						
	Woods	4	0	6	5	15
TRUE	Pasture	0	11	0	2	13
CATEGORY	Corn	0	0	25	12	37
	Soybeans	0	1	4	35	40
	Subtotal	4	12	35	54	105
(b) CONDITIONAL PROBABILITIES, θ^*_{ij}						
	Woods	0.267	0	0.400	0.333	1.000
TRUE	Pasture	0	0.846	0	0.154	1.000
CATEGORY	Corn	0	0	0.676	0.324	1.000
	Soybeans	0	0.025	0.100	0.875	1.000
	Subtotal	0.267	0.871	1.176	1.686	4.000
(c) UNCONDITIONAL PROBABILITIES, p_{ij}						
	Woods	0.038	0	0.057	0.048	0.143
TRUE	Pasture	0	0.105	0	0.019	0.124
CATEGORY	Corn	0	0	0.238	0.114	0.352
	Soybeans	0	0.010	0.038	0.333	0.381
	Subtotal	0.038	0.114	0.333	0.514	1.000

Topics of analytical interest in the analysis of these data include, *inter alia:*

- *Data reduction.* How can these data be efficiently summarized? As one example, Table 1b shows the conditional identification probabilities denoted by θ_{ij}, the relative frequency that a field of type i is classified as type j. In Table 1b, the θ_{ij}'s are given by their usual estimates, $x_{ij}/x_{i.}$. As a second example, Table 1c shows the unconditional probabilities p_{ij}, that a field selected at random from the population is of type i and classified as type j. Here also these values are given by their usual estimates, x_{ij}/N. (These are maximum likelihood estimates under certain assumptions, as discussed in Kendall and Stuart (1967, Vol. II, p. 548), and are the unique minimum variance unbiased estimates in any event (see Bishop *et al.* 1975, p. 407).) Note that these unconditional probabilities are of value only if the ground truth fields are chosen at random or adjustments are made in the case of a stratified sample. See Card (1982) for discussion of these and related issues. Ulaby *et al.* (1980) unfortunately do not specify the sampling procedure used.

A related problem that arises often in connection with automated decision rules, such as produced by linear discriminant analysis, is that of estimating the error rates to be expected in practice rather than those observed on training data. See the work of Glick (1978) for useful discussion and background.

- *Hypothesis testing.* Do these data differ significantly from another set? For example, in the same reference Ulaby *et al.* present classification results developed using like- or cross-polarized data alone. Does the use of both types of polarization significantly improve classification accuracy? If these data come from different areas, or were taken at different time periods, or using alternative sensors, are the observed differences statistically significant or could they have come about by chance alone? In these latter situations the data would be presented as a multivariate contingency table and additional subscripts defined to represent season, area, sensor type, etc. Relevant literature on the general problem of analysis of categorical data includes works of Fleiss (1973), Bishop *et al.* (1975), Fienberg (1977), and Kendall and Stuart (1967).
- *Estimation*—Given imagery over a new area with observed numbers of fields $x_{.j}$, what are the best estimates of the true number of fields of each type in the new area? Relevant analytical models can be found in Maxim *et al.* (1981) and Bauer *et al.* (1978).

CHARACTERISTICS OF AND DIFFICULTIES WITH THESE DATA

There are two characteristics of these data worth noting in the present context:

- These matrices often contain a fairly "large" number of cells (though only 16 in this example) with a "small" number of observations per row (from 13 to 40 in this example).
- In consequence, the density of cells with zero counts is often quite high. In Table 1a, for example, six out of 16 cells (about 38 percent) contain zero counts. Such zeros may truly reflect zero probabilities, but can also reflect "sampling zeros"—cells which have a non-zero probability but are nonetheless empty due to sampling variability. In truth, not all zeros are alike; some zeros may be "smaller" than others. The rate of misclassification of soybeans as woods (zero out

of 40), for example, may be smaller than the fraction of woods classified as pasture (zero out of 13), but these differences are not reflected in the maximun-likelihood estimates of the conditional probabilities shown in Table 1b.

The probability of a sampling zero can be quite high, particularly if the true conditional probability is low and the number of observations is small. For example, there are 20 to 1 odds for a sampling zero if the true probability is 0.005 and the same size is 10, and the odds of a sampling zero are greater than 2 to 1 even if the sample size is as much as 80. The probability of a sampling zero out of N independent trials with true probability θ is $(1-\theta)^N$.

Another useful question to ask is, how large a sample size is necessary to have a specified probability, g, that no sampling zero will result? For one cell the minimum sample size is

$$N = \frac{ln\ (1-g)}{ln\ (1-\theta)} \tag{1}$$

where ln is the log to the base "e." Thus, for example, the sample size must be at least 460 to have a 90% chance of avoiding a sampling zero when the true cell probability is 0.005. This minimum sample size rises steeply as θ approaches zero. Moreover, even these computations *underestimate* the actual likelihood because only one cell is considered here and confusion matrices typically contain many cells. To judge from published data, sample sizes for row totals in confusion matrices are seldom of sufficient magnitude to reduce the likelihood of sampling zeros to "tolerable" values when θ is small.

Of course, the observed zeros in the table might well reflect a genuine zero probability (though upon reflection, a probability that is *exactly* zero seems unlikely in this context), so a more descriptive term for the cell zeros is "problematic zeros," a term suggested by Fienberg and Holland (1970).

Why Are Sampling Zeros a Problem?

There are several reasons why sampling zeros constitute a problem in the analysis of contingency-table data. These include:

- The phenomenon noted above that "not all zeros are the same." In some applications, for example, studies on health effects of smoking, alternative surgical procedures, etc., small differences in probability can be highly important and, therefore, every effort should be made to extract the most information possible from the data. (See Bishop and Mosteller (1969) for one illustration.)
- In some hypothesis testing or model development situations, zeros significantly complicate or even prevent the use of certain analytical techniques. For example, Quirk and Scarpace (1982) note the difficulty of using chi-square tests when some cells have zero frequencies. Bishop *et al.* (1975) note the difficulties occasioned by sampling zeros in model-fitting using the method of weighted least squares. Finally, the use of certain transformations of the data, e.g., the logarithmic, may be impossible if there are zeros in the data.
- In "scaling-up" experimental results with misclassification errors, use of a sampling zero in place of the correct non-zero probability can significantly affect numerical results. Maxim *et al.* (1981), for example, provide the following estimate of the true number of fields of "type 1," y_1, from the observed number of fields of type 1 and type 2, x_1 and x_2, misclassification probabilities denoted by α and β, respectively, and the detection probability p as

$$y_1 = \frac{(1-\beta)\,x_1 - \beta\,x_2}{p\ (1-\alpha-\beta)}. \tag{2}$$

If the values for x_1, x_2, p, α, β were 90, 460, 0.5, 0.2, and 0, respectively, y_j would be about 225. But what if β were really 0.1 rather than 0, as assumed? The corresponding y_1 would be only 100, less than one-half of the previous value. The sensitivity of the estimate to β arises because x_2 is large relative to x_1. This condition may not have occurred in the ground truth sample—particularly if the ground truth sample was purposive rather than random. (It is a well-known result that the discrimination power of a comparative test is maximized if equal sample sizes are chosen (see Fleiss (1973).)

For these and other reasons, investigators have sought plausible ways to deal with the problems of sampling zeros. Many of these approaches appear to have a distinctly *ad hoc* character. For example, sparce rows are often combined to eliminate zeros (or even small values). Fienberg and Holland (1970) and Good (1965) summarized various other suggestions. Several of these approaches avoid zero counts by the simple expedient of adding "pseudo counts" to the cells; illustrative procedures include adding 1 count to all cells, 1/4 count to only the empty cells, 1/2 count to all cells, or to only the empty cells, etc.

Pseudo-Bayesian Estimates: A Useful Approach?

One class of approaches that has some proven theoretical as well as practical merit is the use of so-called pseudo-Bayesian estimates. Operationally, the use of pseudo-Bayesian estimators can be succinctly described by the following steps:

- Select a set of r^2 prior unconditional probabilities λ_{ij} for all i and j. These λ_{ij} values may be specified exogenously, be data dependent, or developed by a procedure described later in this paper.
- Compute/estimate/specify a weighting factor, K. Options for estimation of K are described in the following section.
- Compute "smoothed" cell estimates

$$m^*_{ij} = N\,p_{ij}{}^* = \frac{N}{N+K}\,(x_{ij} + K\,\lambda_{ij}). \tag{3}$$

> Smoothed conditional probabilities θ_{ij}^* or unconditional probability follow directly from the smoothed cell estimates.

Estimates, p_{ij}^*, derived from Equation 3 are weighted linear combinations of the maximum-likelihood estimates of the unconditional probabilities, $p_{ij} = x_{ij}/N$, and the λ_{ij} values, i.e.,

$$p^*_{ij} = \frac{N}{N+K} p_{ij} + \frac{K}{N+K} \lambda_{ij}. \qquad (4)$$

From Equation 4 it can be seen that whenever $K > 0$ and $\lambda_{ij} > 0$, then $p^*_{ij} > 0$ even if $p_{ij} = 0$, i.e., cell zeros will be removed. (As a technical point, estimates given by Equations 3 or 4 are termed pseudo-Bayesian because, while the form of Equation 4 is similar to a Bayesian estimator, the parameters K and λ_{ij} are replaced by point estimates rather than distributions and the values may be data dependent.) Numerical examples will follow, but first it is appropriate to comment on the parameters K and λ_{ij}.

The Choice of K

From examination of Equation 4 it can be seen that K functions as a weighting coefficient. Specifically, the estimate p^*_{ij} given in Equation 4 is exactly what would be obtained if λ_{ij} were an independent estimate of the cell probability based on a sample size of K. K is thus the "pseudo sample size" associated with the prior estimates $\lambda_{ij.}$ A large value for K (in comparison to N) causes the prior information to be heavily weighed relative to the experimental data; small values the opposite.

Mathematically, any value of $K \geq 0$ is admissible, and indeed several more or less *ad hoc* suggestions for choice of K have been ventured; among these $K = N^{0.5}$, $K = r^2/2$ (see Bishop *et al.* (1975)). One particular choice that has both theoretical and practical merit is

$$\hat{K} = \frac{N^2 - \sum\sum x_{ij}^2}{\sum\sum x_{ij}^2 - 2N\sum\sum x_{ij}\lambda_{ij} + N^2\sum\sum \lambda_{ij}^2}. \quad (5)$$

(For the technically minded, $\hat{K}$ as defined in Equation 5 is the maximum-likelihood estimate of the value of K that minimizes the quadratic risk function of the estimator Equation 4. See Bishop *et al.*, 1975, for a discussion of risk functions and the derivation of Equation 5. Estimates given by Equation 4 using $\hat{K}$ given by Equation 5 may be biased, but, depending upon the choice of λ_{ij} and the p_{ij} values, of these estimates can be a "substantial improvement" in terms of the quadratic risk (see Bishop *et al.*, 1975, p. 407) over the traditional estimate.) Yet other estimates of K have been proposed and are appropriate under certain circumstances; however, $\hat{K}$, as defined above, is often a reasonable choice and, in any event, had received the widest attention.

The Choice of λ_{ij}

The λ_{ij} values represent the *a priori* estimates of the p_{ij} values. While any sequence of λ_{ij} values that satisfy the basic conditions of a probability distribution (viz. $0 \leqslant \lambda_{ij} \leqslant 1.0$ and $\Sigma\Sigma\ \lambda_{ij} = 1.0$) are mathematically acceptable, the objective is to select values as close to the true but unknown probabilities as possible.

In the specific context of the remote sensing illustration identified in Table 1, there is often exogenous information available that can be used to estimate or at least to narrow the domain of reasonable choices for λ_{ij}. For example,

- Data may be available from another experiment that is expected to yield similar classification probabilities.
- If the present sensor is designed to be an improvement over an earlier design or sensor and image-enhancement technology combination, earlier data might be used to set bounds upon the performance of the present sensor.
- Prior land use maps or other aerial survey data may enable estimation of the row totals $p_{i\cdot}$ if the fields/cells/quadrats are selected at random. (See Card (1982) for a discussion of this idea in a different context.)

What is needed is a systematic technique for incorporation of all available prior information in the selection of the λ_{ij} values. Our proposed choice is described below.

Beginnings of a Systematic Procedure: Maximum Entropy Distributions

A concept that is highly useful in many areas of statistics, information theory, communications theory, statistical thermodynamics, etc., is that of *entropy*. In the specific context of a bivariate contingency table, the entropy, E, of the prior estimates is defined by the equation,

$$E = -\Sigma\Sigma\lambda_{ij} \ln \lambda_{ij}, \qquad (6)$$

and depends upon all the prior cell probabilities. Loosely, the entropy is a measure of the "randomness" of the assignment. As noted, because the λ_{ij} come from a probability distribution, they must also satisfy the constraint that the sum of the cell probabilities is unity (i.e., $\Sigma\Sigma\ \lambda_{ij} = 1.0$). Absent detailed knowledge, it is reasonable to select a prior distribution that maximizes the "randomness" of the assignment. The distribution that maximizes the entropy can be found by solving the optimization problem,

$$\text{Max } E = -\Sigma\Sigma\ \lambda_{ij}\ ln\ \lambda_{ij} \qquad (7)$$

$$\text{Subject to } \Sigma\Sigma\ \lambda_{ij} = 1.0. \qquad (8)$$

The solution to the above problem is well known and requires setting the λ_{ij} all equal to a common value, $1/r^2$ in the case of an $r \times r$ matrix (see Jaynes (1957a, 1957b, 1963) or Kullback (1967)). In other words, this prior distribution is uniform and reflects the principle that, absent other knowledge, all cells should be assumed to be "equally likely." Thus, this prior distribution has equal pseudocounts in each cell. The following example illustrates requisite computations for pseudo-Bayesian estimates using this result.

Example 1

Table 2 shows results of the use of the maximum entropy prior on the data of Ulaby *et al.* (1980) given in Table 1. In this example, r = 4 and the λ_{ij} values are, therefore, all equal to $1/r^2$ or 1/16. The calculated value of $\hat{K}$ from Equation 5 is 5.7824, small in comparison to N, and thus the weighting coefficient for the actual data, $(N/(N + K))$, is 0.9478. Such a weighting is in accord with intuition, because if little is known *a priori*, it is reasonable that the raw data should figure significantly in the final results. The value of the entropy of the prior, $-\sum\sum \lambda_{ij}\, ln\, \lambda_{ij}$ is 2.77, while that of the smoothed data $-\sum\sum p^*_{ij}\, ln\, p^*_{ij}$ is 1.98. Thus, as might be expected, the structure of the actual data reduces the entropy of the resulting distribution.

Table 2a contains the smoothed cell counts (note that these are no longer necessarily integers), given by the equation

$$m^*_{ij} = N\, p_{ij}^* = \frac{N}{N + K}\,(x_{ij} + K\, \lambda_{ij}); \qquad (9)$$

Table 2b shows the smoothed conditional probabilities, given by the equation

$$\theta^*_{ij} = m^*_{ij}/m^*_{i\cdot}; \qquad (10)$$

and finally, Table 2c displays the smoothed unconditional probabilities, given by the equation

$$p^*_{ij} = m^*_{ij}/N. \qquad (11)$$

(Throughout this paper, several places of accuracy will be retained. This is done both to facilitate the reader following the computations and to avoid round-off errors in these calculations. Final smoothed estimates can be adjusted as desired.) Referring first to Table 2a, note that, after smoothing, the cell counts no longer sum to the original row totals. This is because no such constraint was placed upon the prior. Indeed, as noted, the prior "pseudo data" have all row and column totals equal. It is only because the smoothing coefficient $(N/(N + K))$ weights the actual data so heavily that the smoothed row totals are so similar to the values of the original data set. Secondly, note that the smoothed counts are all greater than zero. (This is true whenever it is also true of the prior distribution.) Note also that all cells that originally had zero counts, now have non-zero counts which are equal to a common value $N\, K/(r^2)\,(N + K)$; the same is true for the p^*_{ij} values (which differ only by the scaler $1/N$),

Table 2. Estimates with Uniform Prior (Maximum Entropy Without Row or Cclumn Constraints)

			ASSIGNED CATEGORY			
		Woods	Pasture	Corn	Soybeans	Subtotal
(a) SMOOTHED COUNTS m^*_{ij}						
	Woods	4.1338	0.3425	6.0394	5.0816	15.5973
TRUE	Pasture	0.3425	10.7684	0.3425	2.2381	13.6915
CATEGORY	Corn	0.3425	0.3425	24.0376	11.7162	36.4388
	Soybeans	0.3425	1.2903	4.1338	33.5157	39.2823
	Subtotal	5.1613	12.7437	34.5533	52.5516	105
(b) SMOOTHED CONDITIONAL PROBABILITIES, θ^*_{ij}						
	Woods	0.265	0.022	0.387	0.326	1.000
TRUE	Pasture	0.025	0.787	0.025	0.163	1.000
CATEGORY	Corn	0.009	0.009	0.660	0.322	1.000
	Soybeans	0.009	0.033	0.105	0.853	1.000
	Subtotal	0.308	0.851	1.177	1.664	4.000
(c) SMOOTHED UNCONDITIONAL PROBABILITIES, p^*_{ij}						
	Woods	0.039	0.003	0.058	0.048	0.149
TRUE	Pasture	0.003	0.103	0.003	0.021	0.130
CATEGORY	Corn	0.003	0.003	0.229	0.112	0.347
	Soybeans	0.003	0.012	0.039	0.319	0.347
	Subtotal	0.049	0.121	0.329	0.500	1.000

ADJUSTMENT PARAMETERS:
$K = 5.7824 \qquad \lambda_{ij} = 1/16 \;\; \forall\, i,j$
OTHER COMPUTATIONS:
$-\sum\sum \lambda_{ij}\, ln\, \lambda_{ij} = 2.77 \qquad -\sum\sum p^*_{ij}\, ln\, p^*_{ij} = 1.98$

though not of the θ^*_{ij} as the smoothed-row totals differ.

Thus, these pseudo-Bayes estimates remove the (presumed) sampling zeros. However, though improved, these estimates are still not entirely satisfactory. Some specific objections that can be raised include:

- There is no particular reason to alter the original row totals of the data; indeed, perhaps these should be retained.
- Though the raw estimates need improvement, the choice of the "equally likely" hypothesis may be equally unsatisfactory—the experimenter may know more about the situation than this distribution would indicate.
- Adjusted cell counts of cells with "original zeros" are all identical.

Yet, the maximum entropy approach has a certain intuitive appeal as well as theoretical justification. Can these deficiencies be remedied?

In Search of Improvements

Referring to the first of the above objections, a natural extension to the method is to impose additional constraints upon the prior. If row (and column) totals are believed correct, for example, then a prior distribution can be selected that maximizes the entropy of the prior, but subject to the constraint that row and/or column totals match those for the experimental data. This requires solving a new optimization problem,

$$\text{Max } E = -\sum\sum \lambda_{ij} \, ln \, \lambda_{ij} \qquad (12)$$

Subject to

$\sum_j \lambda_{ij} = p_{i\cdot} \forall i$ (row totals maintained) (13)

$\sum_i \lambda_{ij} = p_{\cdot j} \forall j$ (column totals maintained) (14)

$\sum\sum \lambda_{ij} = 1.0$ (probabilities sum to unity). (15)

It can be shown that the resulting prior distribution is given by (see Good (1965)),

$$\lambda_{ij} = p_{i\cdot} \, p_{\cdot j}, \qquad (16)$$

a distribution identical to the independence hypothesis in contingency table analysis. This, too, has been suggested as a candidate prior (see Bishop *et al.* (1975) and indeed employed in a recent remote sensing study (see Quirk and Scarpace (1982)), but it is interesting to note that, rather than being another arbitrary (albeit simple) choice, it can be developed by a direct extension of the maximum entropy idea.

For the case where the prior is given by $\lambda_{ij} = p_{i\cdot} \, p_{\cdot j}$, the entropy of the prior E is 2.35. Note that this value is less than that for the equal-probability case, reflecting the additional information contained in the row and column totals. Likewise, computations indicate the value of K is higher, 12.69, compared to 5.78 in the first example. Still, even in this case, actual data are highly weighted (.8922) in comparison to the prior, reflecting the lack of certainty of the *a priori* information.

The smoothed counts will now sum to the row and column totals of the original data, a direct consequence of the additional constraints placed on the prior. As well, the smoothed counts corresponding to cells with "original zeros" are now positive and non-zero; however, unlike the smoothing shown in Table 2, the smoothed values for these cells are not all identical; another consequence of the row and column constraints. These are all improvements to the original smoothing of Example 1. Even so, the assumptions that underlie the smoothing given by this model can still be challenged. Perhaps of most controversy are the classification (conditional) probabilities implied by the maximum entropy solution. For if $\lambda_{ij} = p_{i\cdot}p_{\cdot j}$, then the corresponding θ_{ij} (given by $\lambda_{ij}/p_{i\cdot}$.) is simply the column total $p_{\cdot j}$. In other words, the classification probabilities do not differ row-to-row (although they do differ column-to-column) and are numerically equal to the column probability sums—a dubious assertion when applied to a photo-classification matrix. Thus, though the prior resulting from solution of Equation 12 improves upon that of Table 2, it is still not an entirely satisfactory reflection of likely prior knowledge.

Trying Again

Further improvements can be effected by the same mechanism used above—adding constraints to the maximum entropy formulation. Consideration of the expected properties of misclassification matrices in remote sensing applications suggests that

- The probability of correct classification, θ_{ii}, or the diagonal elements of the conditional probability matrix ought to be "higher" than off-diagonal elements. Absent other specific knowledge, it is reasonable to assume initially that these diagonal elements are all equal, and given this assumption, a natural estimate of their common value is the observed proportion of cells correctly classified. For the data given in Table 1, this proportion is (4 + 11 + 25 + 35)/105 or 0.7143. A constraint that specifies this is

$$\theta_{ii} = \sum x_{ii}/N \; \forall \; i. \qquad (17)$$

- Absent knowledge to the contrary, it is also reasonable to assume that misclassification is symmetric. Thus, for example, if 10 percent of cornfields are misclassified as soybean fields then, under the symmetry assumption, 10 percent of the soybean fields are misclassified as cornfields (later this condition is relaxed). Mathematically,

$$\theta_{ij} = \theta_{ji} \quad i \neq j. \qquad (18)$$

- As before, it is reasonable to require that the prior proportion of fields in each category match the observed proportion in the data

$$\Sigma_j \lambda_{ij} = p_{i\cdot}, \qquad (19)$$

i.e., the row-sum constraint, though in our view there is no compelling reason to impose a column-sum constraint.

Combining the above constraints and recasting them in λ_{ij} form, the maximum entropy optimization problem becomes,

$$\text{Maximize E} = -\sum\sum \lambda_{ij} ln\ \lambda_{ij}, \qquad (20)$$

$$\text{subject to} \quad \sum \lambda_{ij} = p_{i\cdot} \qquad \forall\, i \text{ (row-sum constraint)} \qquad (21)$$

$$\lambda_{ij} = p_{i\cdot}\ (\sum x_{ii}/N) \quad \forall i \text{ diagonal element constraint} \qquad (22)$$

$$\lambda_{ji} = \lambda_{ij} \frac{p_{j\cdot}}{p_{i\cdot}} \quad i \neq j \text{ symmetry constraint} \qquad (23)$$

$$\sum\sum \lambda_{ij} = 1.0 \qquad (24)$$

Numerous computer codes exist which can solve non-linear optimization problems such as that posed above and (provided feasible solutions exist) the problems are well-behaved; that is, the domain of feasible solutions is convex, the objective function is concave and smooth (see Fiacco and McCormick (1968) for one method of solution).

Example 2

Table 3 shows the solution to the above optimization problem in terms of both conditional and unconditional probabilities. Shown also are computations of the maximum entropy and $\hat{K}$. Table 4 shows the resulting smoothed estimates based upon the data. Several points are noteworthy:

- In terms of conditional probabilities, the maximum-entropy prior allocates the misclassification probability equally among the non-diagonal elements. Coupled with the symmetry constraint, the effect is to set all non-diagonal elements to a common value (0.0952 in this instance). Diagonal elements are, of course, constrained to be equal to the common estimate 0.7143. (See Table 3a.)
- The λ_{ij} values for the prior sum to specified row totals, though column totals differ from the experimental data. The entropy of the prior, 2.1843, is smallest of all the priors advanced, reflecting the additional structure imposed by the constraints. (Note that a prior consisting of only one non-zero element, and that equal to unity, would have zero entropy.) The $\hat{K}$ value, 35.3, is largest among the priors considered—indeed, the $\hat{K}$ value is negatively associated with the entropy of the prior.
- Referring to Table 4, note the effect of the smoothing. The conditional probability matrix (Table 4b) retains much of the structure of the original data matrix (Table 1b), but the effects of smoothing are also evident; zeros are removed and the diagonal elements smoothed towards the average fraction correctly classified.

Table 3. Maximum Entropy Prior with Symmetric Misclassification Matrix and $\theta_{ii} = \theta_{jj}$

		ASSIGNED CATEGORY				
		Woods	Pasture	Corn	Soybeans	Subtotal
(a) CONDITIONAL PROBABILITIES, θ_{ij}						
	Woods	0.7143	0.0952	0.0952	0.0952	1.000
TRUE	Pasture	0.0952	0.7143	0.0952	0.0952	1.000
CATEGORY	Corn	0.0952	0.0952	0.7143	0.0952	1.000
	Soybeans	0.0952	0.0952	0.0952	0.7143	1.000
	Subtotal	1.000	1.000	1.000	1.000	4.000
(b) RESULTING UNCONDITIONAL PROBABILITIES, $\lambda_{ij} = p_i \cdot \theta_{ij}$						
	Woods	0.1020	0.0136	0.0136	0.0136	0.1428
TRUE	Pasture	0.0118	0.0884	0.0118	0.0118	0.1238
CATEGORY	Corn	0.0335	0.0335	0.2517	0.0335	0.3524
	Soybeans	0.0363	0.0363	0.0363	0.2721	0.3810
	Subtotal	0.1836	0.1718	0.3134	0.3310	1.000
(c) ANCILLARY COMPUTATIONS						

$$\sum x_{ij}^2 = 2{,}213 \qquad -\sum\sum\lambda_{ij} ln\ \lambda_{ij} = 2.184$$

$$\sum \lambda^2_{ij} = 0.1639023 \qquad N^2 \sum \lambda_{ij}^2 = 1807.023$$

$$\sum \lambda_{ij} x_{ij} = 17.9531 \qquad -2N \sum x_{ij}\ \lambda_{ij} = -3770.151$$

$$K = \frac{11{,}025 - 2{,}213}{2{,}213 - 3770.151 + 1807.023} = 35.266$$

Of the estimates presented thus far, Table 4 appears the most reasonable. But, can the smoothing of Table 4 be further improved? To do so would require more technical knowledge about the characteristics of the radar system and the response of the crops. Certainly the maximum entropy framework is sufficiently flexible to incorporate additional prior knowledge. This is conveniently done with the constraint set. Several illustrations are provided below;

- If it were known, *a priori,* that there were "structural zeros" or "logical zeros" in the data matrix (e.g., that is was *impossible* to misclassify crop k as crop l), constraints of the form $\theta_{kl} = 0$ will ensure this result in the smoothed values, if true for the data. The objective function (Equation 20) is modified by deleting the kl^{th} term.
- If misclassification probabilities were known not to be symmetric, constraints can be devised to incorporate this knowledge.
- Suppose it were known that a confusion potential existed within a certain set of crops denoted by the symbol A, and within yet another set denoted by the symbol B, but that misclassification between sets was not possible. Constraints of the form,

$$\theta_{ij} = 0 \text{ if } i \in A \text{ and } j \in B, \tag{25}$$

would impose this logical structure. Again, the objective function is altered to delete requisite terms.

- Inequality constraints can also be used to model *a priori* knowledge. If discrimination between crops i and j were known to be at least as difficult as between crops k and l, for example, a constraint of the form

$$\theta_{ij} \geqslant \theta_{kl} \tag{26}$$

would ensure that the prior reflected this knowledge. If experience suggested that this misclassification was twice as likely, then the appropriate constraint would be $\theta_{ij} \geqslant 2\theta_{kl}$, etc. Referring to the observed matrix of conditional probabilities (shown in Table 1b), the high misclassification probabilities for woods as corn or soybeans compared to woods classed as pasture might be thought genuine, rather than simply an artifact of the data. If so, the misclassification probabilities of the prior could be constrained to accommodate this belief.

- Prior experience might be limited to more aggregated knowledge, such as overall estimates of misclassification rates expressed as errors of comission or omission). For example, the error of commision is the probability that a cell is classified as type j when it is not j. This quantity is given by,

$$\sum_{\substack{i=1 \\ i \neq j}}^{r} \lambda_{ij} \text{ or equivalently} \sum_{\substack{i=1 \\ i \neq j}}^{r} \theta_{ij}\, p_{i\cdot} \tag{27}$$

Constraints that specify upper bounds for Equation 27 can be added to the optimization problem.

- Finally, it is possible to place constraints upon the final smoothed estimates themselves, i.e., the

TABLE 4. ESTIMATES USING MAXIMUM ENTROPY PRIOR WITH SYMMETRIC MISCLASSIFICATION, DATA ESTIMATED θ_{ii}, AND CONSTRAINED ROW TOTALS

		ASSIGNED CATEGORY				
		Woods	Pasture	Corn	Soybeans	Subtotal
(a) SMOOTHED COUNTS, m^*_{ij}						
	Woods	5.687	0.3590	4.8505	4.1019	15
TRUE	Pasture	0.3115	10.5681	0.3115	1.8087	13
CATEGORY	Corn	0.8844	0.8844	25.392	9.8673	37
	Soybeans	0.9583	1.7069	3.9526	33.3835	40
	Subtotal	7.8412	13.518	34.507	49.161	105
(b) SMOOTHED CONDITIONAL PROBABILITIES, θ_{ij}						
	Woods	0.379	0.024	0.323	0.273	1.000
TRUE	Pasture	0.024	0.813	0.024	0.139	1.000
CATEGORY	Corn	0.024	0.024	0.686	0.267	1.000
	Soybeans	0.024	0.043	0.099	0.835	1.000
	Subtotal	0.451	0.904	1.132	1.514	4.000
(c) SMOOTHED UNCONDITIONAL PROBABILITIES, p^*_{ij}						
	Woods	0.054	0.003	0.046	0.039	0.143
TRUE	Pasture	0.003	0.101	0.003	0.017	0.123
CATEGORY	Corn	0.008	0.008	0.242	0.094	0.352
	Soybeans	0.009	0.016	0.038	0.318	0.381
	Subtotal	0.074	0.128	0.329	0.468	1.000

ADJUSTMENT PARAMETERS:
$\hat{K} = 35.266$ λ_{ij} as shown in Table 3

> m^*_{ij} values, if appropriate. Incorporation of these constraints, however, complicates the mathematics of the optimization problem because of the appearance of K in the constraints. (No such problem emerges if K is specified in advance, but if K is computed by Equation 5 the constraints will contain quadratic forms.)

All that is necessary to incorporate these is a sufficiently flexible non-linear optimization computer code, prior knowledge, and some imagination.

UNKNOWN STATISTICAL PROPERTIES OF THESE ESTIMATES

At present, the statistical properties of the resulting pseudo-Bayesian estimators have only been rigorously explored for selected cases. These cases, for the most part, are highly idealized and simple (see Bishop *et al.* (1975))—and, moreover, results tend to depend upon the true (but unknown) population parameters. In particular, little is known about the properties of the constrained maximum-entropy priors; though logic suggests that *if the prior knowledge is accurate,* then the properties of the estimates ought to compare favorably with conventional alternatives. Moreover, results from the simple cases examined to date are encouraging and suggest that further examination will prove fruitful. Of particular interest is the observed property that the risk of the pseudo-Bayesian estimates is often appreciably less than that of conventional estimates when probabilities are significantly different from unity and when the dimensionality of the tables increases (Fienberg and Holland, 1970). So, even though the original rationale for these estimates centered on the difficulties occasioned by zero counts, these estimators may be much more broadly useful. In any event, few would argue against the wisdom of reflecting all prior knowledge in the development of these estimates.

The ideas advanced herein for generating prior distributions have theoretical and practical justification and, at least with the aid of computers, are easy to implement. But only time and further work will prove the worth of these estimators. In particular, it would be interesting to compare the methods suggested here with some of those suggested in Glick (1978 and referenced papers). It is particularly appropriate to close with a quote from this paper,

> "The task of estimating probabilities of correct classification confronts the statistician simultaneously with difficult distribution theory, questions intertwining sample size and dimension, problems of bias, variance, robustness, and computation costs. But coping with such conflicting concerns (at least in my experience) enhances understanding of many aspects of statistical classification—and stimulates insight into general methodology of estimation."

Remote sensing specialists are encouraged to examine these ideas and follow developments in this area.

ACKNOWLEDGMENTS

The authors wish to thank Ms. Mary Kennedy, Dr. Nancy David, and Dr. James Campbell for stimulating discussions and comments on an earlier draft of this paper. The reviewers' comments added clarity and focus to this paper. The reviewers also brought the work of Glick to our attention.

REFERENCES

Bauer, M. E., M. M. Hixson, B. J. Davis, and J. B. Etheridge, 1978. Area Estimation of Crops by Digital Analysis of LANDSAT Data, *Photogrammetric Engineering and Remote Sensing,* Vol. 44, No. 8, pp. 1033-1043.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland, 1975. *Discrete Multivariate Analysis,* MIT Press, Cambridge, Mass.

Bishop, Y. M., and F. Mosteller, 1969. Smoothed Contingency-Table Analysis, *The National Halothane Study,* ed. J. P. Bunker *et al.*, chapter IV-3, pp. 237-286. Report of the Subcommittee on the National Halothane Study of the Committee on Anesthesia, Division of Medical Sciences, National Academy of Sciences-National Research Council, National Institutes of Health, National Institute of General Medical Sciences, Bethesda, Md., Washington, D.C., U.S. Government Printing Office.

Card, Don H., 1982. Using Known Map Category Marginal Frequencies to Improve Estimates of Thematic Map Accuracy, *Photogrammetric Engineering and Remote Sensing,* V. 48, No. 3, pp. 431-440.

Fiacco, A. V., and G. P. McCormick, 1968. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques,* John Wiley and Sons, New York.

Fienberg, S. E., 1977. *The Analysis of Cross-Classified Categorical Data,* MIT Press, Cambridge, Mass.

Fienberg, S. E., and P. W. Holland, 1970. Methods for Eliminating Zero Counts in Contingency Tables, in *Random Counts in Models and Structures,* ed. by G. P. Patil, The Pennsylvania State University Press, University Park, Pennsylvania.

Fleiss, J. L., 1973. *Statistical Methods for Rates and Proportions,* John Wiley and Sons, New York, New York.

Ginevan, M. E., 1979. Testing Land-use Map Accuracy: Another Look, *Photogrammetric Engineering and Remote Sensing,* V. 45, No. 10, pp. 1371-1377.

Glick, N., 1978. Additive Estimators For Probabilities of Correct Classification, *Pattern Recognition,* V. 10, pp. 211-222.

Good, I. J., 1965. *The Estimation of Probabilities,* MIT Research Monograph, No. 30, MIT Press, Cambridge, Mass., p. 73.

Hay, A. M., 1979. Sampling Designs to Test Land Use Map Accuracy, *Photogrammetric Engineering and Remote Sensing,* V. 45, No. 4, pp. 529-533.

Jaynes, E. T., 1957a. Information Theory and Statistical Mechanics, *Physical Review,* 106, pp. 620-630.

———, 1957b. Information Theory and Statistical Mechanics, *Physical Review,* 108, pp. 171-190.

———, 1963. New Engineering Applications of Information Theory, in *Proceedings First Symposium Engineering Applications of Function Theory and Probability,* ed. by J. L. Bogdanoff and F. Kozin, John Wiley and Sons, New York, New York.

Kendall, M. G., and A. Stuart, 1967. *The Advanced Theory of Statistics,* in three volumes, Hafner Publishing Company, New York, New York.

Kullback, S., 1968. *Information Theory and Statistics,* Dover Publications, Inc., New York, New York, p. 7.

Maxim, L. D., L. Harrington, and M. Kennedy, 1981. Alternative Scale-Up Estimators for Aerial Surveys Where Both Detection and Classification Errors Exist, *Photogrammetric Engineering and Remote Sensing,* Vol. 47, No. 8, pp. 1227-1239.

Quirk, B. K., and F. L. Scarpace, 1982. A Comparison Between Aerial Photography and Landsat for computer Land-Cover Mapping, *Photogrammetric Engineering and Remote Sensing,* V. 48, No. 2, pp. 235-240

Ulaby, F. T., P. P. Batlivala, and J. E. Bare, 1980. Crop Identification with L-Band Radar, *Photogrammetric Engineering and Remote Sensing,* Vol. 46, No. 1, pp. 101-105.

van Genderen, J. L., and B. F. Lock, 1977. Testing Land Use Map Accuracy, *Photogrammetric Engineering and Remote Sensing,* V. 43, No. 9, pp. 1135-1137.

van Genderen, J. L., B. F. Lock, and P. A. Vass, 1978. Remote Sensing: Statistical Testing of Thematic Map Accuracy, *Remote Sensing of Environment,* V. 7, pp. 3-14.

(Received 20 May 1982; revised and accepted 30 January 1983)