

PENNSSTATE



Center for **S**tatistical **E**cology and **E**nvironmental **S**tatistics

Analytic Solution of the Regularized Latent Truth Model for Binary Maps
By G. P. Patil, and Charles Taillie

¹Center for Statistical Ecology and Environmental Statistics
Department of Statistics
The Pennsylvania State University
University Park, PA 16802

EPA Project Officer: N. Phillip Ross

Prepared with partial support from the United States Environmental Protection Agency Cooperative Agreement Number CR-825506. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Technical Report Number 2000-0601
TECHNICAL REPORTS AND REPRINTS SERIES
June 2000



Department of Statistics
The Pennsylvania State University
University Park, PA 16802

G. P. Patil
Distinguished Professor and Director
Tel: (814)865-9442 Fax: (814)865-1278
Email: gpp@stat.psu.edu
<http://www.stat.psu.edu/~gpp>

Analytic Solution of the Regularized Latent Truth Model for Binary Maps

G. P. Patil and Charles Taillie

Center for Statistical Ecology and Environmental Statistics

Department of Statistics

The Pennsylvania State University

University Park, PA 16802

Abstract. Consider two maps having the same spatial extent and the same mapping categories but where each map is subject to classification error. An overlay of the maps yields a (dis)similarity matrix whose (i, j) -entry is the areal proportion placed into category i by the first map and into category j by the second map. Patil and Taillie (2003) have proposed a latent truth model which specifies the dissimilarity matrix in terms of the true (but unknown) proportions for the mapping categories and the unknown error rates for the two maps. The number of parameters in the model exceeds the degrees of freedom in the dissimilarity matrix. However, a method of regularization is applied to effectively reduce the dimension of the parameter space and to permit model fitting. From the fitted model, one obtains estimates for the true mapping proportions as well as estimated error matrices for each of the maps. This paper considers binary maps and obtains explicit expressions for the fitted parameters of the regularized latent truth model.

Keywords: Accuracy assessment; Binary maps; Dissimilarity matrix; Error matrix; Latent truth model; Regularization.

Prepared with partial support from the United States Environmental Protection Agency Cooperative Agreement Number CR-825506. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

1 Introduction

Consider a raster map whose pixels have been classified into K categories using two independent classification methods which we refer to as the I -method and the J -method. We suppose that both methods are attempting to arrive at the same underlying “truth” but each classification is potentially subject to error. Let π_{ij} be the fraction of pixels assigned to category i by method I and to category j by method J . Here, i and j range from 1 to K and $\boldsymbol{\pi} = [\pi_{ij}]$ is the (dis)agreement matrix for the two classification methods.

Patil and Taillie (2003) have proposed a latent truth model in order to estimate and study the accuracy of the two classifications, as opposed to their agreement or disagreement. The model supposes that there is some latent “true” classification T . Let p_t , $t = 1, \dots, K$, be the proportion of pixels which are assigned to category t by classification T . Also, for given t , let α_{ti} be the conditional probability that a pixel whose true class is t is assigned to category i by method I . Let β_{tj} be the corresponding conditional probability for method J . Each of the matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is row stochastic. The latent truth model for $\boldsymbol{\pi}$ is given by

$$\pi_{ij} = \sum_t p_t \alpha_{ti} \beta_{tj}, \quad i, j = 1, \dots, K. \quad (1)$$

This model has more parameters on the right hand side of equation (1) than there are degrees of freedom on the left hand side. Therefore, some method of regularization has to be applied in order to effectively reduce the parameter space before the model can be fitted.

This paper studies the case of binary maps with $K = 2$ mapping categories. For this special case, Patil and Taillie (2003) suggest that the regularized forms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be taken as

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 - \delta & \delta \\ \delta & 1 - \delta \end{bmatrix} \quad (2)$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{bmatrix}, \quad (3)$$

The parameters δ and ϵ determine the error rates for the I and J classifications, respectively. The model requires that $0 \leq \epsilon, \delta \leq 1$, but in practice ϵ and δ need to be fairly small (say, $\epsilon, \delta < \frac{1}{2}$) or else the two landcover categories become completely confused. The regularized model with $K = 2$ has 3 independent equations with 3 unknown parameters (p, δ, ϵ). The goal of the present paper is to obtain explicit expressions for these three parameters in terms of the π_{ij} .

2 Solution of the Regularized Model for $K = 2$

We have

$$\pi_{ij} = \sum_{t=1}^2 p_t \alpha_{ti} \beta_{tj}, \quad i, j = 1, 2. \quad (4)$$

where π_{ij} is known (or estimated) and

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 - \delta & \delta \\ \delta & 1 - \delta \end{bmatrix}$$

$$\beta = \begin{bmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{bmatrix}.$$

Let $p_1 = p$ and $p_2 = 1 - p$ so there are three unknown parameters p, δ, ϵ which have to be estimated by fitting the model specified in equation (4). Write the 2×2 table $[\pi_{ij}]$ as

$$\begin{array}{c|c} \text{C} & \text{A} \\ \hline - & - \\ \hline \text{B} & 1 \end{array}$$

where the entries A, B, C represent the 3 degrees of freedom in the 2×2 table. Note that $0 \leq A, B \leq 1$ and $\min(A, B) \geq C \geq A + B - 1$. Patil and Taillie (2003) have suggested that the latent truth model for general K be fitted by minimizing the “distance” between the right and left hand sides of the regularized model equation. In the present case, we see that the number of independent equations is exactly the same as the number of degrees of freedom in the data matrix. Accordingly, we ask if parameters can be estimated by solving equation (4) exactly.

Theorem 1. The system of equations (4) is equivalent to the following set of equations:

1. $A = p(1 - 2\delta) + \delta \equiv p + \delta(1 - 2p)$
2. $B = p(1 - 2\epsilon) + \epsilon \equiv p + \epsilon(1 - 2p)$
3. $C = p(1 - \delta - \epsilon) + \epsilon\delta.$

Proof. For item (1), we have

$$\begin{aligned} A &= \pi_{1+} = \sum_{t=1}^2 p_t \alpha_{t1} \quad \text{since} \quad \beta_{t+} = 1 \\ &= p(1 - \delta) + (1 - p)\delta = p(1 - 2\delta) + \delta. \end{aligned}$$

The derivation of item (2) is similar to that of item (1). For item (3), we have

$$\begin{aligned} C &= \pi_{11} = \sum_{t=1}^2 p_t \alpha_{t1} \beta_{t1} = p(1 - \delta)(1 - \epsilon) + (1 - p)\delta\epsilon \\ &= p(1 - \delta - \epsilon + \epsilon\delta) + \delta\epsilon - p(\delta\epsilon) = p(1 - \delta - \epsilon) + \delta\epsilon. \end{aligned}$$

A solution (p, δ, ϵ) of the three equations in Theorem 1 is said to be *valid* if $0 \leq p, \delta, \epsilon \leq 1$. There are instances in which non-valid solutions exist. Also, solutions need not be unique. In fact, if (p, δ, ϵ) is a solution then so is $(1 - p, 1 - \delta, 1 - \epsilon)$, and if either of these solutions is valid so is the other. We call these paired solutions as *complementary*.

We are going to give necessary and sufficient conditions for the existence of a valid solution. This is a very complicated issue in full generality so we first study the nondegenerate situation in which neither A nor B equals $\frac{1}{2}$. Observe, from the first two equations of Theorem 1, that if there is a solution with $p = \frac{1}{2}$, then $A = B = \frac{1}{2}$; therefore, nondegeneracy excludes the possibility of a solution with $p = \frac{1}{2}$.

3 Solution in the Nondegenerate Case

Lemma 1. Suppose $A \neq \frac{1}{2} \neq B$ and that (p, δ, ϵ) is a solution. Then, $p \neq \frac{1}{2}$ and

1. $\delta = \frac{A-p}{1-2p}$
2. $\epsilon = \frac{B-p}{1-2p}$
3. $C - AB = p(1-p)\frac{1-2A}{1-2p}\frac{1-2B}{1-2p}$
4. $(\frac{1}{2} - p)^2 \cdot Q = p(1-p)$ where $Q = 4\frac{C-AB}{(1-2A)(1-2B)}$.

Proof. Parts (1) and (2) are immediate from (1) and (2) of Theorem 1. Also from Theorem 1, we have that

$$\begin{aligned} AB &= [p(1-2\delta) + \delta][p(1-2\epsilon) + \epsilon] \\ &= p^2(1-2\delta)(1-2\epsilon) + p[\delta(1-2\epsilon) + \epsilon(1-2\delta)] + \delta\epsilon \\ &= p^2(1-2\delta)(1-2\epsilon) + p[\delta + \epsilon - 4\epsilon\delta] + \delta\epsilon. \end{aligned}$$

Subtracting this from $C = p[1 - \delta - \epsilon] + \delta\epsilon$, gives

$$\begin{aligned} C - AB &= p[1 - 2\delta - 2\epsilon + 4\epsilon\delta] - p^2(1-2\delta)(1-2\epsilon) \\ &= p(1-2\delta)(1-2\epsilon) - p^2(1-2\delta)(1-2\epsilon) \\ &= p(1-p)(1-2\delta)(1-2\epsilon) \end{aligned} \tag{5}$$

But, by part (1) of this lemma, we have

$$1 - 2\delta = 1 - 2\frac{A-p}{1-2p} = \frac{1-2p-2(A-p)}{1-2p} = \frac{1-2A}{1-2p}.$$

Similarly, $1 - 2\epsilon = \frac{1-2B}{1-2p}$. Substituting these into equation (5) gives part (3) of the lemma; part (4) is a rearrangement of part (3). This completes the proof.

The lemma provides a roadmap for obtaining solutions. The quadratic in part (4) is solved to obtain values for p and these values are substituted into parts (1) and (2) to obtain δ and ϵ . Reversing the arguments in the proof of Lemma 1 shows that this does produce solutions—however, the solutions do not have to be valid.

Theorem 2. Suppose $A \neq \frac{1}{2} \neq B$. There is a valid solution if and only if each of the following conditions holds (Q is defined in part (4) of Lemma 1):

1. $Q \geq 0$
2. $1 - 2\min(A, B) \leq \sqrt{\frac{1}{1+Q}}$
3. $2\max(A, B) - 1 \leq \sqrt{\frac{1}{1+Q}}$.

In this case, there are exactly two solutions; these solutions are both valid, are complementary, and are given by

$$p = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{1}{1+Q}} \quad (6)$$

$$\delta = \frac{A-p}{1-2p} \quad (7)$$

$$\epsilon = \frac{B-p}{1-2p}. \quad (8)$$

Proof. Solution of the quadratic in part (4) of Lemma 1 is given by equation (6). A necessary and sufficient condition for p to lie in the unit interval is that $Q \geq 0$. When this holds, the two resulting solutions (p, ϵ, δ) are easily seen to be complementary so that validity need be examined for only one of them. We study the solution that results from choosing the negative sign in equation (6). For this choice, $p < \frac{1}{2}$ and the denominator, $1 - 2p$, in the expressions for δ and ϵ is positive. Thus, $\delta \geq 0$ iff $A \geq p \equiv \frac{1}{2} - \frac{1}{2} \sqrt{\frac{1}{1+Q}}$ iff $1 - 2A \leq \sqrt{\frac{1}{1+Q}}$. Since an analogous requirement must hold for B , we obtain condition (2) of the theorem. Finally, $\delta \leq 1$ iff $A - p \leq 1 - 2p$ iff $A \leq 1 - p \equiv \frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{1+Q}}$ iff $2A - 1 \leq \sqrt{\frac{1}{1+Q}}$. Combining this with the analogous requirement on B gives condition (3) of the theorem.

There are two complementary solutions when the conditions of Theorem 2 hold. Is there any way of deciding between these two solutions. The next result gives some guidance.

Lemma 2. Suppose the conditions of Theorem 2 holds so there are two complementary solutions, one with $0 \leq p < \frac{1}{2}$ and the other with $\frac{1}{2} < p \leq 1$. For the solution with $p < \frac{1}{2}$, we have

1. $\delta < \frac{1}{2} \iff A < \frac{1}{2}$
2. $\epsilon < \frac{1}{2} \iff B < \frac{1}{2}$
3. $\epsilon + \delta < 1 \iff A + B < 1$

For the solution with $p > \frac{1}{2}$, we have

1. $\delta < \frac{1}{2} \iff A > \frac{1}{2}$
2. $\epsilon < \frac{1}{2} \iff B > \frac{1}{2}$
3. $\epsilon + \delta < 1 \iff A + B > 1$

Proof. Suppose $p < \frac{1}{2}$ so that $1 - 2p$ is positive. Then, $\delta < \frac{1}{2}$ iff $\frac{A-p}{1-2p} < \frac{1}{2}$ iff $A - p < \frac{1}{2} - p$ iff $A < \frac{1}{2}$. The other parts of the lemma are obtained in a similar fashion.

Now we can decide which of the two solutions to use. Recall that A and B are the relative frequencies assigned to land cover category 1 by the two classification methods while p is the (estimated) true relative frequency of land cover category 1.

Case 1 (A and B both less than $\frac{1}{2}$). The two classifications consistently rank the categories, i.e., they agree that category 1 is the non-dominant category. Then it is natural to take the solution with $p < \frac{1}{2}$. According to Lemma 2, this is also the solution which minimizes the two estimated error rates ϵ and δ . The other solution is mathematically possible but would imply large ϵ and δ so that both classifications are consistently “flipping” the two categories.

Case 2 (A and B both greater than $\frac{1}{2}$). The two classifications agree that category 1 is the dominant category. It is natural to take the solution with $p > \frac{1}{2}$. According to Lemma 2, this solution minimizes the two estimated error rates ϵ and δ .

Case 3 ($A < \frac{1}{2} < B$ or $B < \frac{1}{2} < A$) The two classifications disagree as to which is the dominant category and it is not intuitively obvious whether to take $p < \frac{1}{2}$ or $p > \frac{1}{2}$. According to Lemma 2, we can minimize the total inaccuracy, $\epsilon + \delta$ by taking $p < \frac{1}{2}$ if $A + B < \frac{1}{2}$ and $p > \frac{1}{2}$ if $A + B > \frac{1}{2}$

4 Solution in the Degenerate Case

Here we suppose that one or both of A and B equal $\frac{1}{2}$. We obtain conditions for a valid solution and find that there are usually uncountably many valid solutions whenever a valid solution exists. First, we state the results.

Case 1 ($A = \frac{1}{2} = B$ and $C = \frac{1}{4}$). There are uncountably many solutions obtained by setting any two of the parameters p, δ, ϵ equal to $\frac{1}{2}$ and choosing the value for the third parameter arbitrarily from the unit interval.

Case 2 ($A = \frac{1}{2} = B$ and $C \neq \frac{1}{4}$). Note that C must be less than or equal to $\frac{1}{2}$ since $C \leq \min(A, B)$. There are uncountably many solutions obtained by setting $p = \frac{1}{2}$, choosing $\delta \in [0, 1] - (\min(2C, 1 - 2C), \max(2C, 1 - 2C))$, and putting

$$\epsilon = \frac{1}{2} + \frac{C - \frac{1}{4}}{\delta - \frac{1}{2}}.$$

Exception: The above prescription gives only two solutions when $C = 0$ or $C = \frac{1}{2}$.

Case 3 ($A = \frac{1}{2}$ and $B \neq \frac{1}{2}$). There is no valid solution unless $C = B/2$ in which case there are uncountably many solutions obtained by setting $\delta = \frac{1}{2}$, choosing $p \in [0, 1] - (\min(B, 1 - B), \max(B, 1 - B))$, and putting

$$\epsilon = \frac{B - p}{1 - 2p}.$$

Exception: The above prescription gives only two solutions when $B = 0, 1$.

Case 4 ($A \neq \frac{1}{2}$ and $B = \frac{1}{2}$). There is no valid solution unless $C = A/2$ in which case there are uncountably many solutions obtained by setting $\epsilon = \frac{1}{2}$, choosing $p \in [0, 1] - (\min(A, 1 - A), \max(A, 1 - A))$, and putting

$$\delta = \frac{A - p}{1 - 2p}.$$

Exception: The above prescription gives only two solutions when $A = 0, 1$.

Before giving the proof, we note that the system of equations in Theorem 1 is equivalent to the following system:

$$1 - 2A = (1 - 2\delta)(1 - 2p) \quad (9)$$

$$1 - 2B = (1 - 2\epsilon)(1 - 2p) \quad (10)$$

$$C - AB = 4p(1 - p)\left(\delta - \frac{1}{2}\right)\left(\epsilon - \frac{1}{2}\right) \quad (11)$$

Equations (9) and (10) are simple rearrangements of the first two equations of Theorem 1. When equations (9) and (10) hold, equation (11) is equivalent to the last equation of Theorem 1 (this was essentially proved in the derivation of equation (5)).

Case 1 is now obvious because the left hand sides of equations 9–11 all vanish. This implies that either $p = \frac{1}{2}$ or $\epsilon = \delta = \frac{1}{2}$. In the former case, equation (11) implies that one of ϵ or δ must equal $\frac{1}{2}$.

In Case 2, the left hand sides of equations (9) and (10) vanish, but that of equation (11) does not. This implies that $p = \frac{1}{2}$ so that equation (11) becomes

$$C - \frac{1}{4} = \left(\delta - \frac{1}{2}\right)\left(\epsilon - \frac{1}{2}\right).$$

This is the equation of a rectangular hyperbola, whose (valid) solutions are as described in Case 2 above.

For Case 3, the left hand side of equation (9) vanishes but that of equation (10) does not. This implies that $\delta = \frac{1}{2}$ and $p \neq \frac{1}{2}$. Now, $\delta = \frac{1}{2}$ requires the right hand side of equation (11) to vanish so that $C = AB \equiv B/2$ is necessary for the existence of a solution. Assuming that $C = B/2$, a solution (not necessarily valid) is obtained by rewriting equation (10) as

$$1 - 2\epsilon = \frac{1 - 2B}{1 - 2p}$$

or

$$\epsilon = \frac{B - p}{1 - 2p}.$$

We have to check for validity of this solution. But, $\epsilon \geq 0$ is equivalent to either (i) $p \leq B$ and $p < \frac{1}{2}$ or (ii) $p \geq B$ and $p > \frac{1}{2}$. Similarly, $\epsilon \leq 1$ is equivalent to either (i) $p \leq 1 - B$ and $p < \frac{1}{2}$ or (ii) $p \geq 1 - B$ and $p > \frac{1}{2}$. Collectively, this is the same as

$$p \leq \min(B, 1 - B) \quad \text{or} \quad \max(B, 1 - B) \leq p,$$

which is the same as the restrictions given in Case 3 above.

Case 4 is analogous to Case 3.

5 Interpretation of $Q > 0$: Relation to Kappa

We have seen that $Q > 0$ is a necessary condition for a solution p to exist where

$$Q = 4 \frac{C - AB}{(1 - 2A)(1 - 2B)}.$$

The interesting case occurs when the two classification methods consistently rank the two categories. Then, $1 - 2A$ and $1 - 2B$ have the same sign so the denominator of Q is positive. Now, AB is the expected relative frequency for cell (1,1) that would occur by chance, i.e., if the table were independent with the same marginals A and B . Since C is the actual frequency for cell (1,1), $C - AB$ is positive when the joint agreement between the two tables exceeds chance. In other words, the condition $Q > 0$ means that the two classification methods tend to agree jointly when they are consistent marginally.

It is easy to give an example (for π_{ij}) where the condition $Q > 0$ fails:

$$\begin{array}{cc|c} 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{3}{4} \\ \hline \frac{1}{4} & \frac{3}{4} & 1 \end{array}$$

Here, $C - AB = 0 - \frac{1}{4}(\frac{1}{4}) = -\frac{1}{16} < 0$. There is perfect agreement marginally but the two methods disagree on 50 percent of the pixels and there is not even a single pixel which they agree to label as category 1.

The above suggests that Q is related to the kappa coefficient κ , particularly in the numerator. Let D_c be the sum of the two diagonal entries in the actual array. Let D_o be the sum of the diagonal entries in the independent table having the same marginals. Recall that $\kappa = \frac{D_c - D_o}{1 - D_o}$. We have

1. $D_o = 1 - A - B + 2AB$.
2. $D_c = 1 - A - B + 2C$.
3. $4(C - AB) = 2(D_c - D_o)$.
4. $(1 - 2A)(1 - 2B) = 2(D_o - \frac{1}{2})$
5. $Q \equiv \frac{4(C - AB)}{(1 - 2A)(1 - 2B)} = \frac{D_c - D_o}{D_o - \frac{1}{2}}$.

Corollary. The marginals in a 2×2 table consistently order the two categories if and only if $D_o \geq \frac{1}{2}$.

References

- Congalton, R. G. and Green, K. (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Lewis Publishers, Boca Raton, FL.
- Patil, G. P. and Taillie, C. (2003). *Modeling and Interpreting the Accuracy Assessment Error Matrix for a Doubly Classified Map*. *Environmental and Ecological Statistics* (to appear). Also Technical Report Number 2000-0502, Center for Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA 16802.