Vol 2, No 2, pp. 78–85, 2010
International Journal of
Mathematical and Computational
Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: May 5, 2010
Accepted: Jun. 29, 2010
Published: Aug. 28, 2010
Last Correction: Aug. 5, 2010

THE ROLE OF MISCLASSIFICATION IN ESTIMATING PROPORTIONS AND AN ESTIMATOR OF MISCLASSIFICATION PROBABILITY

Patrick L Zimmerman[1,2], Greg C Liknes[2],
[1] *School of Statistics, University of Minnesota - Twin Cities, Minneapolis, MN*
[2] *USDA Forest Service, Northern Research Station, St. Paul, MN*

Abstract. Dot grids are often used to estimate the proportion of land cover belonging to some class in an aerial photograph. Interpreter misclassification is an often-ignored source of error in dot-grid sampling that has the potential to significantly bias proportion estimates. For the case when the true class of items is unknown, we present a maximum-likelihood estimator of misclassification probability based on agreement between two interpreters. Two of the assumptions underlying the estimator are: (i) the probability that an interpreter makes a misclassification is constant, (ii) both interpreters have the same probability of misclassification. Simulation results suggest the estimator has acceptable performance when (ii) does not hold. This estimator can be used to investigate whether bias due to misclassification has exceeded a threshold, or to correct bias due misclassification.

Keywords: Dot grid, remote sensing, image interpretation, misclassification

1 Introduction

Aerial photography has been used extensively as a tool in natural resource assessments in a variety of ways, particularly in forestry. For example, aerial photography has been used to assess the damage from pests such as mountain pine beetle (White et al. 1983) and eastern spruce budworm (Munson et al. 1985). Hansen (1985) used linear transect sampling with airphotos to inventory wooded strips in Kansas. More recently, Frescino et al. (2006) used aerial imagery to assess forest resources in Nevada. A dataset of forest canopy density across the United States was developed by modeling the relationship between interpreted airphotos and Landsat satellite imagery (Huang et al. 2001).

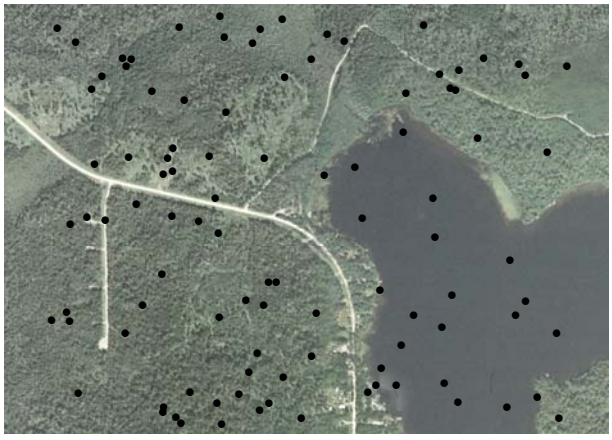

Figure 1: a sample dot grid with randomly located dots.

For decades, airphotos played a vital role in forest inventory programs (Loetsch and Haller 1964). The USDA Forest Service Forest Inventory and Analysis program photo-classified nearly 200,000 1-acre photo points into forest land, unproductive forest, nonforest, and water categories in support of the 1983 Wisconsin (USA) inventory (Spencer et al. 1988). Forest inventory applications of airphotos include area estimation, stratification, photo sampling, and map creation. With regard to photo sampling, aerial photographs may be used in conjunction with dot grids to estimate proportions of a feature of interest. Historically, an interpreter would place a transparent overlay with a regular or systematic dot grid on top of an airphoto and assign dots to a category of interest (e.g., tree/no tree or damaged/not damaged). Proportions are simply the relative counts in each of the categories divided by the total dot count. Dot grid methods have moved forward into the digital age, and tools have been developed for use with digital aerial imagery

(e.g., Clark et al. 2004, Lister et al. 2009) (Figure 1). In addition to digital tools, natural resource practitioners now have access to an ever-increasing collection of high-resolution imagery from the United States. The USDA's National Agriculture Imagery Program (NAIP) has been collecting aerial imagery since 2003. A few states are flown over and photographed each year, with an approximate return interval of three years. Many states have also acquired their own resource photography, such as the New York State Digital Orthophoto Program (NYSDOP) imagery, Pennsylvania's PAMAP imagery, and the New Jersey Department of Environmental Protection (NJDEP) imagery.

While considerable work is being done in the area of image segmentation and automated feature extraction for natural resource inventory and monitoring (e.g., Chubey et al. 2006, Smith et al. 2008, Laliberte et al. 2004), dot grids still represent an efficient method for estimating proportions over large areas. The topics of sample size (Gering and Bailey 1984) and sampling error (Bonnor 1974) for dot grids have been given considerable attention. However, the error associated with proportions estimated from dot grids arises from two sources: sampling error and misclassification of individual dots. Considering only the binary case (tree/no tree, damage/no damage), we discuss a mathematical model of proportion estimates that includes misclassification (Section 2), and then derive and discuss a possible estimator of misclassification probability as a function of interpreter agreement (Section 3). Along with sampling error, this estimate should provide a more realistic estimate of total error in proportions as derived from interpreted data. While application to dot grids is intended, the model applies more generically to any data an interpreter assigns to one of two classes.

2 The impact of misclassification

2.1 A mathematical model for misclassification The effects of misclassification on a sample estimate of a proportion have been described by Bross (1954). Suppose that an estimate of the proportion of items in a population that fall into some class C is desired, and that p is the proportion of items in C in the population with the rest being in class N. Suppose furthermore that there is a θ probability of misclassification for each item (in either class). Now, a sample of size n is drawn, and the number of items, X, classified as C are counted. A typical estimator of p, X/n (denoted $\bar{X}_n$), will be biased. Its expected value is changed due to misclassification such that

$$E(\bar{X}_n) = p - 2p\theta + \theta$$

This implies that the bias of the estimate is

$$E(\bar{X}_n) - p = -2p\theta + \theta \quad (1)$$

Despite its bias, $\bar{X}_n$ still follows a binomial distribution.

To give an idea of the magnitude of this bias in a concrete scenario, suppose that a photo interpreter classifies dots in a dot grid as either falling on tree canopy (class C) or not (class N), and misclassifies dots in both classes 5% of the time. In the notation of the previous section, $\theta = 0.05$. If this were the case, a population with a true proportion in C of 20% would have $E(\bar{X}_n) = 23\%$ - i.e., a 3% bias - and a population with 10% proportion would have $E(\bar{X}_n) = 14\%$. As these examples suggest, the bias is drawing the estimator towards 50% because there are more dots truly in class N, and therefore more opportunities to misclassify class N dots as class C. This example implies that if many proportions are estimated with some misclassification each estimate would be biased toward 50% (from either direction). Importantly, this is not a problem that would be "washed out" across many populations. Whereas sampling error decreases as a sample size increases, bias due to misclassification will remain.

It should be noted that this model assumes that the misclassification rate is identical for items in both classes. In fact, Bross (1954) presents this model for the more general case where the probability of misclassifying items in class C is not necessarily the same as that of misclassifying items in class N. The possibility of developing the current work in this more general framework is discussed later. Also, note that misclassification is not the only potential source of bias. For example, systematic sampling can result in a bias.

2.2 What can be done about bias due to misclassification? There is often no way to collect the information necessary to directly estimate misclassification probabilities. We may never know the correctness of an interpreter's classifications, only his/her agreement with those of another (fallible) interpreter. For example, in the case of photo interpretation studies, it is likely infeasible to verify the correctness of classified dot grids with *in situ* observations. To summarize the problem with proportion estimates where misclassification probabilities are non-zero but unknown, the estimates are biased to an unknown and possibly large extent.

For the case when an estimate of the misclassification probability θ is desired to correct the bias of $\bar{X}_n$, but the true classes of the items classified by an interpreter are unknown, we show there is information about misclassification probabilities contained in the agreement between interpreters on an individual item basis. In fact, the following relationship holds:

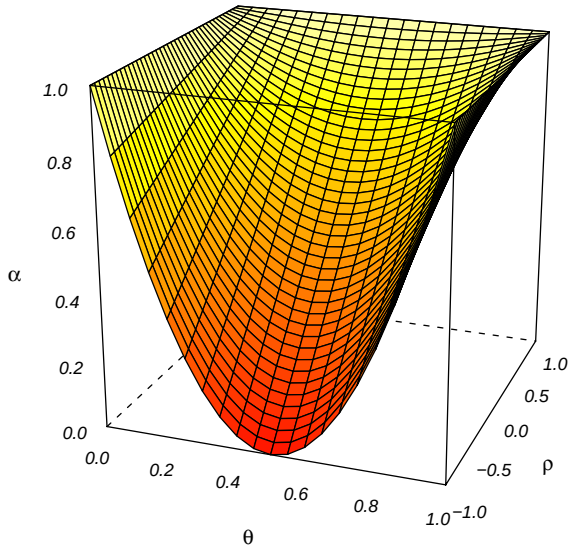

Figure 2: Interpreter agreement (α) as a function of misclassification probability (θ) and the correctness correlation (ρ).

Result. *Let α be the probability of agreement between two interpreters, let θ denote the misclassification probability, and let ρ be the correlation between the correctness of classifications made by the two interpreters (this value is discussed below). Then,*

$$\alpha = \theta^2 + 2\rho\theta(1-\theta) + (1-\theta)^2 \tag{2}$$

This result is fully developed in the appendix, but for now, a little explaining is in order, and a graphical depiction of the relationship is presented in Figure 2. This relationship states that agreement is a quadratic function of misclassification probability. If the two interpreters are either always right or always wrong, they will always agree. What happens in every other case depends on the correctness correlation, ρ. If the interpreters tend to misclassify the same items, ρ will be positive. If one interpreter is more likely to correctly classify the items misclassified by the other interpreter, ρ will be negative. In a case where all interpreters have received the same training and successfully follow the same procedure in classifying items, it is expected that ρ will be non-negative. The two interpreters may have a difficult time classifying certain types of items, and the interpreters should be using the same strategies to classify items that are not immediately obvious. Note that, if ρ is zero, knowing that the first interpreter misclassified an item offers no information about whether or not the second interpreter misclassified that item. Using the above relationship, we can use observed agreement between interpreters to say something useful about misclassification, and, subsequently, the bias of estimated proportions.

3 Estimating a misclassification probability

3.1 Statement of the estimator Suppose that two fallible interpreters classify items as belonging to one of two classes, but that the true class of the items is unknown. With a few assumptions, it is possible to get an estimate of their misclassification probability based on paired classifications made by the interpreters, and to calculate the asymptotic variance of this estimate. Note that the individual paired classifications made by the interpreters must be known for each item, and not just the total number of items in the population classified as belonging to each class by the interpreters. A formal statement of the following theorem and a proof are given in the appendix. Also, the more general case where $\rho \in (-1, 1)$ is considered in the appendix.

Theorem. *Suppose that two interpreters classify n different items, and that $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n$ are observed where $\boldsymbol{A}_i = 1$ if the interpreters agree on item i, and $\boldsymbol{A}_i = 0$ otherwise, for $i = 1, \ldots, n$. Suppose also that the following assumptions hold,*

(A1) *the classifications for each item made by an interpreter are independent and have the same misclassification probability.*

(A2) *the classifications made by both interpreters have a θ probability of misclassification.*

(A3) $\theta < \frac{1}{2}$.

(A4) $\rho = 0$.

then, if we define $\bar{\boldsymbol{A}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i$,

$$\hat{\theta} = \frac{1}{2} - \sqrt{\frac{2\bar{\boldsymbol{A}}_n - 1}{4}} \tag{3}$$

is the maximum likelihood estimator of θ, and is normally distributed with mean θ and variance $\frac{\alpha(1-\alpha)}{4n(2\alpha-1)}$ as $n \to \infty$. If an estimate of the asymptotic variance of $\hat{\theta}$ is desired, it is recommended that the maximum likelihood estimate of α, $\bar{\boldsymbol{A}}_n$, be used in place of α.

3.2 Discussion of assumptions There is one problematic assumption underlying this estimator. **(A1)** is fulfilled if the correctness of an interpreter's classifications are like a series of coin flips: different classifications do not influence each other and the probability of misclassification is the same each time. The first part of this assumption (independence) is realistic. If an interpreter

uses the same procedure to classify each dot, it seems reasonable to believe that the misclassification probability for an item will depend only on characteristics of that item, and not, for example, on how difficult to classify the previous item was or how many items the interpreter has classified. The assumption of a constant probability could be unrealistic, though. For example, if interpreters are classifying dots from a dot grid as landing on tree cover or not, misclassification probabilities could differ depending on where a dot is located in the image. It may be more likely that misclassification will occur in sparse forest than in dense forest or an open area. **(A2)** is not very realistic, but simulation suggests that, if the two interpreters have different misclassification probabilities, $\hat{\theta}$ is an estimator of their average misclassification probability. These results are presented shortly. **(A3)** only requires that the interpreters are classifying at least half of the items correctly. **(A4)** seems problematic at first. However, recall the scenario described above under which ρ would be expected to be positive (Section 2.2): the two interpreters are more likely to misclassify some items than others. Given **(A1)**, this cannot occur, and, as long as the interpreters are working separately, it is difficult to imagine another scenario under which ρ would be non-zero.

3.2.1 Simulation study of (A2) Assumption **(A2)** states that both interpreters have the same misclassification probability. This assumption will certainly never be exactly true. Therefore, a simulation study was conducted to determine the behavior of $\hat{\theta}$ under a variety of scenarios.

First, a population was defined to have either 5%, 20%, or 50% of the items in category C. Classifications of these items by two fallible interpreters were simulated with a correctness correlation of zero. The first interpreter always had a 5% misclassification while the second interpreter had either a 3%, 7%, 9%, or 15% misclassification probability. Then, 10000 samples of size 100 were drawn from this population.

95% confidence intervals were calculated in each sample. Figure 3 and Figure 4 present histograms of $\hat{\theta}$ and the estimated $SE(\hat{\theta})$, respectively, from the population with 50% of the items in class C. Note that, in Figure 4, the sample standard deviations of the estimates are depicted. The sample standard deviation is measuring the variation of $\hat{\theta}$ that actually occurred in the simulation. Thus, it can be considered a target value for estimated standard errors. Table 1 shows numerical results from each population.

This simulation study suggests that, when the misclassification probabilities of two interpreters differs, $\hat{\theta}$ behaves approximately as if it were estimating the misclassification probability of two other interpreters who both had the average misclassification of the first two. Estimates tend to be slightly inflated (less than 1%) when the second interpreter has a 15% misclassification probability. Note also that, although the theoretical standard error, $SE(\hat{\theta})$, is asymptotic (its behavior is known only as $n \to \infty$), the observed coverage probabilities suggest that the estimated standard error is performing well at this sample size.

Table 1: Simulation results. p is the true proportion of items in class C; θ_2 is the misclassification probability of the second interpreter ($\theta_1 = 0.05$ always); θ_{AVE} is the average misclassification probability between the two interpreters; $\mu_{\hat{\theta}}$ is the mean observed $\hat{\theta}$; $sd(\hat{\theta})$ is the sample standard deviation of the simulated misclassificaion probability estimates; $\mu_{SE(\hat{\theta})}$ is the mean observed $SE(\hat{\theta})$; *coverage* is the proportion of times that a 95% CI covered θ_{AVE}.

p	θ_2	θ_{AVE}	$\mu_{\hat{\theta}}$	$sd(\hat{\theta})$	$\mu_{SE(\hat{\theta})}$	*coverage*
0.05	0.03	0.04	0.04	0.015	0.014	0.95
	0.07	0.06	0.061	0.018	0.018	0.94
	0.09	0.07	0.071	0.02	0.02	0.95
	0.15	0.1	0.104	0.025	0.025	0.95
0.2	0.03	0.04	0.04	0.015	0.014	0.95
	0.07	0.06	0.06	0.018	0.018	0.94
	0.09	0.07	0.071	0.02	0.02	0.95
	0.15	0.1	0.104	0.025	0.025	0.96
0.5	0.03	0.04	0.04	0.015	0.014	0.95
	0.07	0.06	0.06	0.018	0.018	0.93
	0.09	0.07	0.071	0.02	0.02	0.95
	0.15	0.1	0.104	0.025	0.025	0.96

4 Discussion

4.1 Usefulness of $\hat{\theta}$ Any study that produces proportion estimates based on fallible classifications and states the precision of its estimates should try to account for misclassification. If misclassification probabilities are not known and proportion estimates are reported as if they were unbiased, they have the potential to be very misleading. Such a study could utilize $\hat{\theta}$ in at least two different ways.

First, a tolerance level for bias due to misclassification could be established, and $\hat{\theta}$ used to detect misclassification probabilities which correspond to biases that exceed the tolerance level (see Equation (1)). Second, $\hat{\theta}$ could be calculated for some pair of interpreters, and then any proportion estimate produced by these interpreters in the future could have its bias corrected. Specifically, if the bias corresponding to the observed $\hat{\theta}$ was subtracted

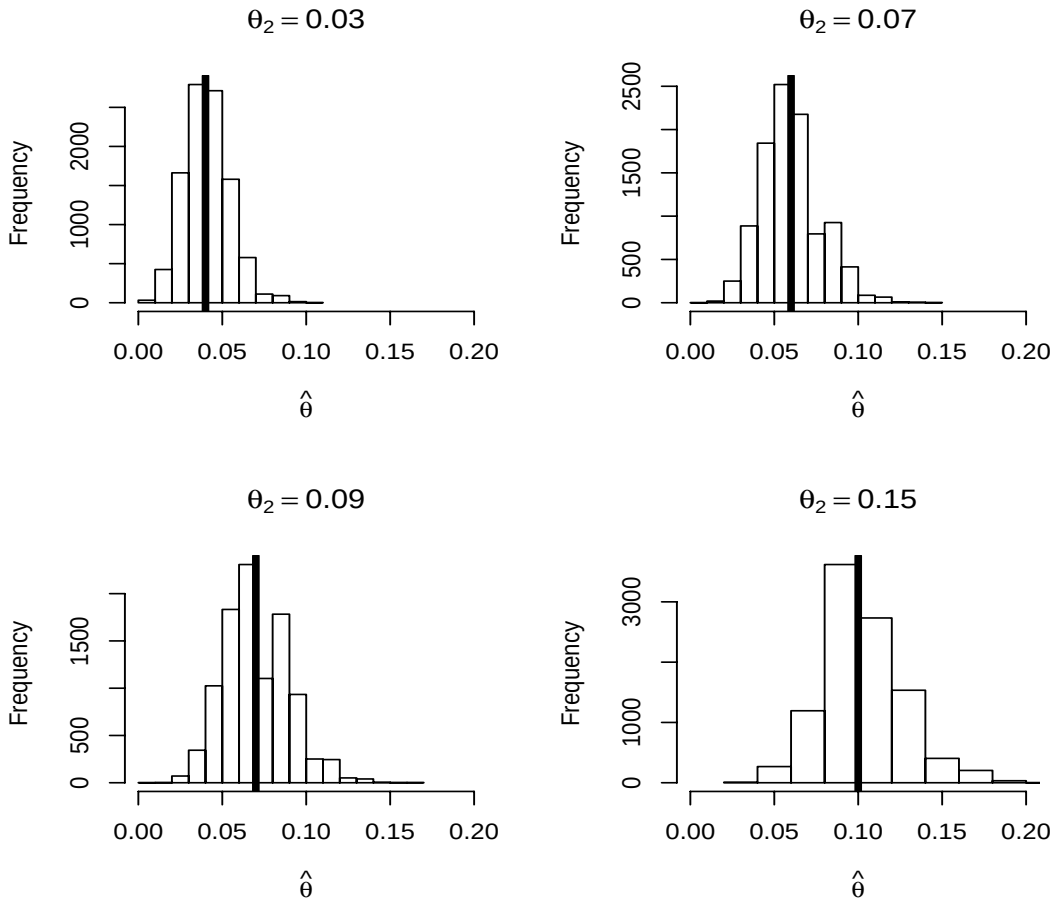

Figure 3: Histograms of misclassification estimates, $\hat{\theta}$. $\theta_1 = 0.05$ and $p = 0.5$ in each histogram. The thick vertical line represents the average misclassification rate.

from a proportion estimate, that estimate's bias due to misclassification would be corrected.

The largest problem with using $\hat{\theta}$ in either of these ways is the requirement that the misclassification probability for each item is identical. A substantial deviation from this assumption would likely be associated with a deviation from the assumption that ρ is equal to zero.

4.2 Future work In this paper, the case was considered where the probability of misclassification was constant across items classified by an interpreter, which may not be a realistic scenario. It may be helpful to consider two ways in which this assumption may be broken: (i) classification may have a difficulty level that varies for each item, and (ii) the probability of misclassification of items in class C may be different from those in class N. For (i), a more realistic model could be developed if a hierarchical model is used. Specifically, the probability of misclassification itself could perhaps be treated as a random variable that is bounded between zero and one. The target value for estimation would then be the mean misclassification probability. The case (ii) may be more difficult if it is necessary to estimate the misclassification probability for both classes. This difficulty arises because, for any given classification, the true class of the item is unknown. Hence, it would be unknown whether agreement on this classification was providing information about the misclassification probability of items in class C or items in class N.

5 Acknowledgments

We thank anonymous reviewers for many helpful comments, as well as M. Hansen, G. Oehlert, P. Patterson, and C.H. Perry for comments on early versions of this manuscript. Finally, thanks to M. Holland for being a helpful conversant in the development of this work.

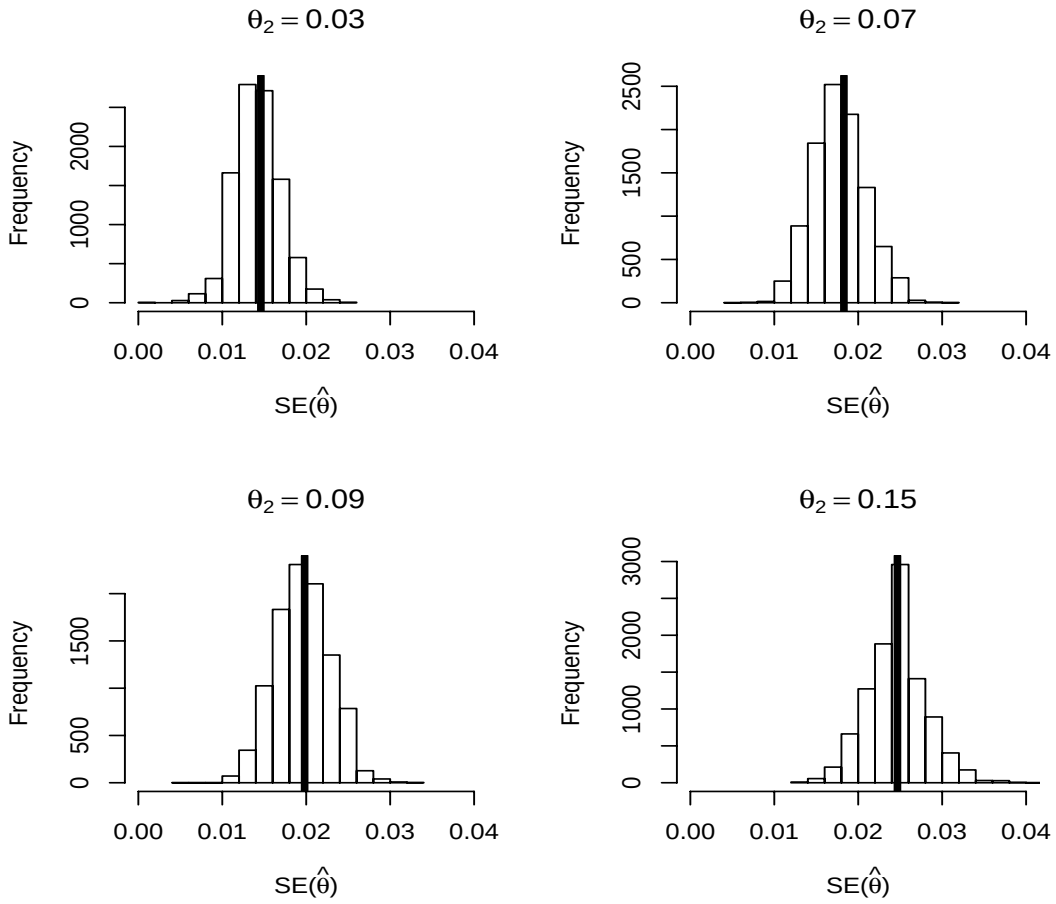

Figure 4: Histograms of the standard error of misclassification estimates, $SE(\hat{\theta})$. $\theta_1 = 0.05$ and $p = 0.5$ in each histogram. The thick vertical line represents the sample standard deviation of the simulated misclassification probability estimates.

References

Bonnor, G.M. 1974. The error of area estimates from dot grids. Can. J. For. Res. 5(10): 10-17.

Bross, I. 1954. Misclassification in 2 X 2 tables. Biometrics 10(4): 478-486.

Casella, G. and R.L. Berger. 2002. Statistical Inference, 2nd ed. Duxbury. 243 p.

Clark, J.T., M.V. Finco, R. Warbington, and B. Schwind. 2004. Digital Mylar: A tool to attribute vegetation polygon features over high-resolution imagery. Available online at http://www.fs.fed.us/r5/rsl/publications/. Last accessed Jan. 20, 2010.

Chubey, M.S., S.E. Franklin, and M.A. Wulder. 2006. Object-based analysis of Ikonos-2 imagery for extraction of forest inventory parameters. Photogrammetric Engineering & Remote Sensing 72(4): 383-394.

Frescino, T.S., G.G. Moisen, K.A. Megown, V.J. Nelson, E.A. Freeman, P.L. Patterson, M. Finco, K. Brewer, and J. Menlove. 2009. Nevada Photo-Based Inventory Pilot (NPIP) photo sampling procedures. USDA For. Serv. Gen. Tech. Rep. RMRS-222. 30 p.

Gering, L.R. and R.L Bailey. 1984. Optimum dot-grid density for area estimation with aerial photographs. Journal of Forestry 82(7): 428-431.

Hansen, M. 1985. Line intersect sampling of wooded strips. Forest Science. 31(2): 282-288.

Huang, C., L. Yang, B. Wylie, and C. Home. 2001. A strategy for estimating tree canopy density using Landsat 7 ETM+ and high resolution images over

large areas. [CD-ROM] Disc 1 in Proc. of the Third International Conference on Geospatial Information in Agriculture and Forestry.

Laliberte, A.S., A. Rango, K.M. Havstad, J.F. Paris, R.F. Beck, R. McNeely, and A.L. Gonzalez. 2004. Object-oriented image analysis for mapping shrub encroachment from 1937 to 2003 in southern New Mexico. Remote Sensing of Environment 93(1-2): 198-210.

Loetsch, F. and K.E. Haller. 1964. Forest Inventory Volume 1: Statistics of Forest Inventory and Information from Aerial Photographs. Bayerischer Landwirtschaftsverlag GmbH. Munich, Germany.

Lister, A., T. Lister, and J.A. Doyle. 2009. Use of a simple photointerpretation method with free, online imagery to assess landscape fragmentation. [CD-ROM] from Proc., 2009 Society of American Foresters national convention, Opportunities in a forested world.

Munson, A.S., W.B. White, R.J. Myhre, and W.H. Hoskins. 1985. Evaluation of three survey methods for determining Spruce-Fir mortality caused by Eastern Spruce Budworm. Forest Pest Management Methods Application Group Report. 85-2. 18 p.

Smith, A.M.S., E.K. Strand, M.S. Caiti, D.B. Hann, S.R. Garrity, M.J. Falkowski, and J.S. Evans. 2008. Production of vegetation spatial-structure maps by per-object analysis of juniper encroachment in multitemporal aerial photographs. Can. J. Remote Sensing 34(Suppl. 2): S1-S18.

Spencer, J.S., W.B. Smith, J.T. Hahn, and G.K. Raile. 1988. Wisconsin's fourth forest inventory, 1983. Resource Bulletin NC-107. St. Paul, MN: U.S. Dept. of Agriculture, Forest Service, North Central Forest Experiment Station.

White, W.B, W.E. Bousfield, and R.W. Young. 1983. A survey procedure to inventory ponderosa and lodgepole pine mortality caused by the mountain pine beetle. Forest Service, Washington, D.C.

APPENDICES

More formal notation will be used in the Appendix in order show results concisely. Also, results will be derived for the more general condition where $\rho \in (-1, 1)$. The theorem from the body of the paper considers the special case where $\rho = 0$.

A Relationship between agreement, misclassification, and conditional correlation

Result. *Suppose that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are binary random variables, such that* $\mathrm{P}(\mathbf{X}_1 = 1) = \mathrm{P}(\mathbf{X}_2 = 1) = 1 - \theta$. *Also, suppose that* $\mathrm{P}(\mathbf{X}_1 = \mathbf{X}_2) = \alpha$ *and that* $\mathrm{Cor}(\mathbf{X}_1, \mathbf{X}_2) = \rho$. *Then,*

$$\alpha = \theta^2 + 2\rho\theta(1-\theta) + (1-\theta)^2 \tag{3}$$

Proof. Note that, for any random event $\mathbf{A}$,

$$\mathrm{P}(\mathbf{A}) = \mathrm{E}(\mathrm{I}(\mathbf{A}))$$

Then,

$$\begin{aligned}\alpha &= \mathrm{P}(\mathbf{X}_1 = 1 \cap \mathbf{X}_2 = 1) + \mathrm{P}(\mathbf{X}_1 = 0 \cap \mathbf{X}_2 = 0) \\ &= \mathrm{E}(\mathrm{I}(\mathbf{X}_1 = 1)\mathrm{I}(\mathbf{X}_2 = 1)) + \mathrm{E}(\mathrm{I}(\mathbf{X}_1 = 0)\mathrm{I}(\mathbf{X}_2 = 0)) \\ &= \mathrm{E}(\mathbf{X}_1\mathbf{X}_2) + \mathrm{E}(\mathrm{I}(\mathbf{X}_1 = 0)\mathrm{I}(\mathbf{X}_2 = 0))\end{aligned}$$

Now,

$$\mathrm{E}(\mathbf{X}_1\mathbf{X}_2) = \mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) + \mathrm{E}(\mathbf{X}_1)\mathrm{E}(\mathbf{X}_2)$$

and

$$\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \rho\sqrt{\mathrm{V}(\mathbf{X}_1)\mathrm{V}(\mathbf{X}_2)}$$

Now, note that $\mathrm{E}(\mathbf{X}_i) = 1 - \theta$ and $\mathrm{V}(\mathbf{X}_i) = \theta(1-\theta)$ for $i = 1, 2$. Thus,

$$\mathrm{E}(\mathbf{X}_1\mathbf{X}_2) = \rho\theta(1-\theta) + (1-\theta)^2$$

Then,

$$\begin{aligned}\mathrm{E}(\mathrm{I}(\mathbf{X}_1 = 0)\mathrm{I}(\mathbf{X}_2 = 0)) &= \mathrm{Cov}(\mathrm{I}(\mathbf{X}_1 = 0), \mathrm{I}(\mathbf{X}_2 = 0)) \\ &\quad + \mathrm{E}(\mathrm{I}(\mathbf{X}_1 = 0))\mathrm{E}(\mathrm{I}(\mathbf{X}_2 = 0))\end{aligned}$$

and since $\mathrm{E}(\mathrm{I}(\mathbf{X}_i = 0)) = \theta$ and $\mathrm{V}(\mathrm{I}(\mathbf{X}_i = 0)) = \theta(1-\theta)$ for $i = 1, 2$, we see that

$$\mathrm{E}(\mathrm{I}(\mathbf{X}_1 = 0)\mathrm{I}(\mathbf{X}_2 = 0)) = \delta\theta(1-\theta) + \theta^2$$

where δ is the correlation between the incorrectness of $\mathbf{X}_1$ and $\mathbf{X}_2$. At this point, all that remains is to show that $\rho = \delta$. Note that, if

$$\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathrm{Cov}(\mathrm{I}(\mathbf{X}_1 = 0), \mathrm{I}(\mathbf{X}_2 = 0))$$

then, $\rho = \delta$. So,

$$\begin{aligned}\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) &= \mathrm{Cov}(1 - \mathrm{I}(\mathbf{X}_1 = 0), 1 - \mathrm{I}(\mathbf{X}_2 = 0)) \\ &= \mathrm{Cov}(-\mathrm{I}(\mathbf{X}_1 = 0), -\mathrm{I}(\mathbf{X}_2 = 0)) \\ &= \mathrm{Cov}(\mathrm{I}(\mathbf{X}_1 = 0), \mathrm{I}(\mathbf{X}_2 = 0))\end{aligned}$$

■

B Derivation of the estimator

Theorem. *Let $\mathbf{X}_{11}, \ldots, \mathbf{X}_{n1}$ and $\mathbf{X}_{12}, \ldots, \mathbf{X}_{n2}$ be random variables such that $\mathbf{X}_{ij}$ represents the correctness interpreter i's classification of the j^{th} item. Denote $\mathrm{P}(\mathbf{X}_{ij} = 0) = \theta_{ij}$ for $i = 1, \ldots, n$ and $j = 1, 2$. Finally, set $\mathbf{A}_i = 1$ where $\mathbf{X}_{i1} = \mathbf{X}_{i2}$ and $\mathbf{A}_i = 0$ otherwise, and denote $\alpha_i = \mathrm{P}(\mathbf{A}_i = 1)$ for $i = 1, \ldots, n$. Now, suppose the following assumptions are met:*

(A1) $(\mathbf{X}_{11}, \mathbf{X}_{12}), \ldots, (\mathbf{X}_{n1}, \mathbf{X}_{n2})$ *are independent and* $\theta_{1j} = \ldots = \theta_{nj} = \theta_j$ *when* $j = 1$ *or* 2.

(A2) $\theta_1 = \theta_2 = \theta$.

(A3) $\theta < \frac{1}{2}$.

Then, if we denote $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i$,

$$\hat{\theta} = \frac{1}{2} - \sqrt{\frac{2\hat{\alpha} - 1 - \rho}{4(1-\rho)}} \tag{4}$$

is the maximum likelihood estimator of θ, and is normally distributed with mean θ and variance $\frac{\alpha(1-\alpha)}{4n(2\alpha-1-\rho)(1-\rho)}$ as $n \to \infty$ where $\mathrm{Cor}(\mathbf{X}_{i1}, \mathbf{X}_{i2}) = \rho$ for $i = 1, \ldots, n$.

Note: if an estimate of the asymptotic variance of $\hat{\theta}$ is desired, it is recommended that the maximum likelihood estimate of α, $\bar{\mathbf{A}}_n$, be used in place of α.

Proof. This proof will proceed in two parts. Part 1 explores the relationship between θ, α, and ρ. In Part 2, the properties of $\hat{\alpha}$ and $\hat{\theta}$ will be derived.

Part 1. Let ρ be defined as above, and apply **(A1)** and **(A2)**. Then, the previous result supplies the following starting point,

$$\begin{aligned} \alpha &= \alpha_i \\ &= \theta^2 + 2\rho\theta(1-\theta) + (1-\theta)^2 \\ &= \theta(1-\theta)(2\rho - 2) + 1 \\ \frac{\alpha - 1}{2(\rho - 1)} &= \theta(1-\theta) \end{aligned}$$

Recognizing a quadratic function of θ, we complete the square and write

$$\begin{aligned} (\theta - \frac{1}{2})^2 &= \frac{1}{4} - \frac{\alpha - 1}{2(\rho - 1)} \\ \theta &= \frac{1}{2} \pm \sqrt{\frac{2\alpha - 1 - \rho}{4(1-\rho)}} \\ \theta &= \frac{1}{2} - \sqrt{\frac{2\alpha - 1 - \rho}{4(1-\rho)}} \end{aligned}$$

where the last step is supplied by assumption **(A3)**. Note that, if θ is required to be a real number between zero and $\frac{1}{2}$, then we must have an α such that

$$0 < \frac{1-\alpha}{2(1-\rho)} < \frac{1}{4}$$

When ρ is zero (the special case considered in estimating θ), this amounts to having α between $\frac{1}{2}$ and one. For increasing values of ρ (larger than zero), α is forced to be increasingly close to one.

Part 2. Note that $\mathbf{A}_1, \ldots, \mathbf{A}_n$, are independent (this follows from **(A1)**), and have the same probability distribution, i.e. $\mathrm{P}(\mathbf{A}_i = 1) = \alpha, \mathrm{P}(\mathbf{A}_i = 0) = 1 - \alpha$ for $i = 1, \ldots, n$. This implies that they are distributed as Bernoulli random variables. Hence, the maximum likelihood estimator of α is $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i$, and, by the Central Limit Theorem, it converges in distribution to a Normal random variable with mean α, and variance $\alpha(1-\alpha)$.

Now, for a fixed ρ, note that θ is a one-to-one function of α when θ is assumed to be less than $\frac{1}{2}$. The invariance property of maximum likelihood estimators states that, if $\hat{\tau}$ is the maximum likelihood estimator of τ and if g is a one-to-one function, then $g(\hat{\tau})$ is the maximum likelihood estimator of $g(\tau)$.

Therefore, the maximum likelihood estimator of θ is

$$\hat{\theta} = \frac{1}{2} - \sqrt{\frac{2\hat{\alpha} - 1 - \rho}{4(1-\rho)}}$$

Furthermore, the delta method can be used to determine the asymptotic distribution of $\hat{\theta}$ (Casella and Berger 2002). First, note that $\hat{\theta}$ can be expressed as a differentiable function h of $\hat{\alpha}$ where

$$h'(\hat{\alpha}) = -\frac{1}{4(1-\rho)}\left(\frac{2\hat{\alpha} - 1 - \rho}{4(1-\rho)}\right)^{-\frac{1}{2}}$$

This derivative will be nonzero when the assumptions of the estimator are fulfilled. Therefore, as $n \to \infty$, $\hat{\theta}$ will converge in distribution to a normal random variable with mean θ, and variance

$$\begin{aligned} \left(h'(\alpha)\sqrt{\frac{\alpha(1-\alpha)}{n}}\right)^2 &= \left[-\frac{1}{4(1-\rho)}\left(\frac{2\alpha - 1 - \rho}{4(1-\rho)}\right)^{-\frac{1}{2}}\right]^2 \\ &\quad \times \left(\frac{\alpha(1-\alpha)}{n}\right) \\ &= \frac{\alpha(1-\alpha)}{4n(2\alpha - 1 - \rho)(1-\rho)} \end{aligned}$$

■