Conference on Applied Statistics in Agriculture 1998 - 10th Annual Conference Proceedings

ASSESSING VARIABILITY OF AGREEMENT MEASURES IN REMOTE SENSING USING A BAYESIAN APPROACH

William J. Price

Bahman Shafii

Lawrence W. Lass

Donald C. Thill

See next page for additional authors

Recommended Citation
Price, William J.; Shafii, Bahman; Lass, Lawrence W.; and Thill, Donald C. (1998). "ASSESSING VARIABILITY OF AGREEMENT MEASURES IN REMOTE SENSING USING A BAYESIAN APPROACH," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1276

Author Information

William J. Price, Bahman Shafii, Lawrence W. Lass, and Donald C. Thill

Assessing Variability of Agreement Measures in Remote Sensing Using a Bayesian Approach

William J. Price and Bahman Shafii
Statistical Programs
Lawrence W. Lass and Donald C. Thill
Division of Plant Sciences

College of Agriculture
University of Idaho
Moscow, Idaho 83844

ABSTRACT

Remote sensing imagery is a popular accessment tool in agriculture, forestry, and rangeland management. Spectral classification of imagery provides a means of estimating production and identifying potential problems, such as weed, insect, and disease infestations. Accuracy of classification is traditionally based on ground truthing and summary statistics such as Cohen's Kappa. Variability assessment and comparison of these quantities have been limited to asymptotic procedures relying on large sample sizes and gaussian distributions. However, asymptotic methods fail to take into account the underlying distribution of the classified data and may produce invalid inferential results. Bayesian methodology is introduced to develop probability distributions for Cohen's Conditional Kappa that can subsequently be used for image assessment and comparison. Techniques are demonstrated on a set of images used in identifying a species of weed, yellow starthistle, at various spatial resolutions and flying times.

I. INTRODUCTION

Images from remote sensing are increasingly becoming useful tools in land management. Using computer interpretation, spectral information in digital and photographic images can be classified into meaningful categories. Common uses might include determination of land use in rangeland management, monitoring and prediction of inventories in forestry, or large scale detection of weed, insect and disease infestations in agriculture. Before using remote sensing for management decisions, however, the quality of the classification should first be assessed. Traditionally, statistics such as Cohen's Kappa (Cohen, 1960) have been used for comparison to known ground truth sites (e.g. Congalton, et al., 1983; Hudson and Ramm, 1987; Rosenfield and Fitzpatrick-Lins, 1986; Congalton, 1991). Kappa provides a relative measure of agreement, ranging from no agreement (random chance) to perfect agreement. Inferences based on Kappa provide a means for assessing and comparing spectral classifications. Variability of the Kappa statistic and related inferential methods have traditionally been computed based on large sample size and asymptotic normality assumptions. While this is a general solution, it may lead to invalid inferential results.

An alternative approach is a Bayesian methodology which directly incorporates the discrete nature of the data. Using numerically derived posterior distributions, point estimation and variability measures can be obtained through most probable values and appropriate moments. Inferential results are then obtained through probability intervals on Kappa estimates, and for the purpose of image comparison, through the distribution of pairwise differences.

These techniques are demonstrated using remotely sensed imagery developed to detect the weedy species yellow starthistle in Northern Idaho. Comparisons are made between different flying times and among various image resolutions.

II. METHODS

The basic unit of measurement within a digital image is the pixel. Pixels represent a point on an image which corresponds to a spatial location on the ground. They are recorded as discrete values based on the spectral response at that location. Computerized classification algorithms place each pixel into C predefined categories. For a fixed number of pixels, N, the true category of that location is field checked resulting in “ground truth”. A cross-classification of ground truth and categorized data results in a C x C error matrix:

Ground Truth

Classification		*1*	*2*	*3*		*C*	
	1	x_{11}	x_{12}	x_{13}	...	x_{1c}	$N_{1.}$
	2	x_{21}	x_{22}	x_{23}	...	x_{2c}	$N_{2.}$
	3	x_{31}	x_{32}	x_{33}	...	x_{3c}	$N_{3.}$
		...	...	...	...	...	...
	C	x_{c1}	x_{c2}	x_{c3}	...	x_{cc}	$N_{c.}$
		$N_{.1}$	$N_{.2}$	$N_{.3}$	...	$N_{.c}$	N

$$\sum_{i=1}^{C}\sum_{j=1}^{C} x_{ij} = \sum_{i=1}^{C} x_{i.} = \sum_{j=1}^{C} x_{.j} = N \quad i=1,2,3...C\ ;\ j=1,2,3...C$$

where
Let $f_{ii} = x_{ii}/N$, $f_{i.} = N_{i.}/N$ and $f_{.i} = N_{.i}/N$, then

$$\hat{K} = \left(\sum_{i=1}^{C} f_{ii} - \sum_{i=1}^{C} f_{i.}\, f_{.i}\right) / \left(1 - \sum_{i=1}^{C} f_{i.}\, f_{.i}\right) \quad ;\ i=1,2,3,...C \qquad (1)$$

is a measure of agreement between rows and columns (Cohen, 1960). $\hat{K}$ (Kappa) has been suggested as a measure of agreement for use in remotely sensed data (Congalton, 1991). The error matrix can also be used to evaluate ommissional ($1 - x_{ii}/N_{.i}$) and commissional ($1 - x_{ii}/N_{i.}$) error rates, where $x_{ii}/N_{.i}$ and $x_{ii}/N_{i.}$ are commonly referred to as producer's and user's accuracy (Congalton, 1991). These alternative measures may provide a more meaningful expression of the degree of agreement between ground truth and classification, justification of which has been addressed elsewhere (Price and Shafii, *In preparation*). For the purpose of this study, we will confine our discussion to Kappa and other related measures.

$\hat{K}$ may be partitioned into conditional agreement (called conditional Kappas) for each of the C classifications as:

$$\hat{K}_i=(f_{ii}-f_{.i}\,f_{i.})/(f_{i.}-f_{.i}\,f_{i.})=(f_{ii}/f_{.i}-f_{i.})/(1-f_{i.})=\theta_{i1}/\theta_{i2} \tag{2}$$

for class i, where $\sum_{i=1}^{C}\theta_{i1} \,/\, \sum_{i=1}^{C}\theta_{i2}=\hat{K}$ (Coleman, 1966; Light, 1969). The estimated large sample variance of $\hat{K}_i$ under the assumption of independence of rows and columns is given by (Bishop, 1975) as:

$$V(\hat{K}_i)=(1/N)(f_{i.})(1-f_{.i})/((f_{.i})(1-f_{i.}) \tag{3}$$

Remote sensing presents a special case for $\hat{K}_i$. For a given image, the ground truth totals, $N_{.i}$, are fixed constants. Rewriting $\hat{K}_i$ to reflect this gives

$$\hat{K}_i=(f_{ii}/f_{.i}-f_{i.})/(1-f_{i.})=(x_{ii}/N_{.i}-N_{i.}\;/N)/(1-N_{i.}\;/N). \tag{4}$$

The number of pixels correctly classified for the ith class, x_{ii}, can be assumed as a binomial random variable, $x_{ii} \sim \text{bin}(P_i, N_{.i})$ where P_i and $N_{.i}$ are the true classified proportion and total for class i, respectively. The quantity, $N_{i.}$, represents the number of pixels in the image indicated as class i. After classification of an image, i.e. from the user's perspective, the $N_{i.}$ are no longer variable and may be considered fixed. Likewise, the total number of pixels in the image, N, is constant. Thus, (4) becomes a monotonic transformation of x_{ii} and the information pertaining to x_{ii} may be used to derive a specified distribution for $\hat{K}_i$.

The Bayesian posterior distribution is proportional to the product of a prior distribution, $\pi(.)$, and a data-based likelihood, $\mathcal{L}(.)$. The likelihood in this case is based on the binomial distribution given by:

$$\mathcal{L}(P_i|x_{ii},N_{.i})=\binom{N_{.i}}{x_{ii}}P_i^{x_{ii}}(1-P_i)^{N_{.i}-x_{ii}} \tag{5}$$

For development of the prior distribution, the principle of maximum entropy may be used. In the discrete case, the entropy of a distribution is defined as:

$$Entropy \propto \sum_{all\ k} p_k \ln(p_k) \tag{6}$$

where p_k is the probability of the k^{th} event (Shannon, 1948). Maximizing (6) constrained to $\Sigma p_k = 1$ and any known information, I, leads to a prior distribution, ψ_I, which has the most uncertainty given I (Jaynes, 1968). When no prior information is assumed, e.g. I=null, then the maximization results in the condition $p_k = p_{k'}$, for $k \neq k'$. This concept can be extended to the continuous case (Jaynes, 1968) and applied to the binomial parameter P_i to give

$$\psi_{I=null} = \pi(P_i) = Constant = A, \tag{7}$$

which is also consistent with the concept of the Least Informative Probability (LIP) as outlined by Loredo, 1990.

The resulting posterior distribution for P_i is now given by:

$$\pi(P_i|x_{ii}, N_{.i}) \propto \pi(P_i)\mathcal{L}(P_i|x_{ii}, N_{.i}) = A\binom{N_{.i}}{x_{ii}} P_i^{x_{ii}}(1-P_i)^{N_{.i}-x_{ii}} \tag{8}$$

Since $E[x_{ii}] = P_i N_{.i}$, the expected value for $\hat{K}_i$ is then

$$E[\hat{K}_i] = (P_i - N_{i.}/N)/(1 - N_{i.}/N) \tag{9}$$

and the posterior distribution of $\hat{K}_i$, $\pi(\hat{K}_i|x_{ii}, N_{i.}, N_{.i})$, can then be derived from

$$P(P_i \le b) = P((P_i - N_{i.}/N)/(1-N_{i.}/N) \le (b - N_{i.}/N)/(1-N_{i.}/N) = P(\hat{K}_i \le b\ ) \tag{10}$$

where b ≠ b are constants.

Based on (10), a $(1\text{-}2\alpha)$ probability interval or credible region for $\hat{K}_i$ can be defined as:

$$P(\hat{K}_i^{\alpha} \le \hat{K}_i \le \hat{K}_i^{1-\alpha}) = 1 - 2\alpha \tag{11}$$

where $\hat{K}_i^{\alpha}$ and $\hat{K}_i^{1-\alpha}$ are the α^{th} and $1\text{-}\alpha^{th}$ percentiles of the estimated posterior distribution for $\hat{K}_i$. This region represents with $1 - 2\alpha$ surety, the most plausible values for $\hat{K}_i$ given the information available, I.

Methods for pairwise comparisons of independent estimates of conditional kappas can also be developed. Let $\omega(\hat{k}_i)$ and $\phi(\hat{k}_j)$ be the posterior distributions for conditional kappas $\hat{K}_i$ and $\hat{K}_j$, respectively. Then the joint distribution of $\hat{K}_i$ and $\hat{K}_j$, assuming independence, is defined as $\tau(\hat{k}_i, \hat{k}_j) = \omega(\hat{k}_i)\phi(\hat{k}_j)$ and the distribution of the difference, $\hat{D}_{ij} = \hat{K}_i - \hat{K}_j$, is given by:

$$m(\hat{d}_{ij}) = \int \tau(\hat{k}_i, \hat{k}_i - \hat{d}_{ij}) \delta_{\hat{k}_j} = \int \omega(\hat{k}_i) \phi(\hat{k}_i - \hat{d}_{ij}) \delta_{\hat{k}_j} \tag{12}$$

Traditionally, Bayesian posterior distributions have been derived analytically. This would require simplifying the necessary computations. While certainly desirable, analytical derivations may not always be practical. Given the recent advances in computational ability and availability, numerical derivations have become a viable alternative. In this study, posterior distributions were computed numerically while probability intervals and contrast distributions were derived through interpolation. All computations were carried out with either the SAS datastep (SAS, 1991) or custom C applications. Programs codes are available from the authors at www.uidaho.edu/ag/statprog/kappa.

III. EMPIRICAL RESULTS

Yellow starthistle is well adapted to the semiarid canyonland topography typical of Northern Idaho. It is toxic to horses and reduces forage quality for other livestock. Due to the steep topography of the region, ground surveys are difficult and remote sensing offers an attractive alternative for the detection and location of yellow starthistle. The data used in this study was collected as a means of assessing the remote sensing potential.

The remote sensing study area was near Lapwai, Idaho (Lass, Carson, and Callihan, 1996). The target area was two by three kilometers (Figure 1). Digital aerial images were obtained on June 21 and July 17, 1994 at resolutions giving 0.5, 1.0, and 2.0 square meters per pixel. A fourth resolution of 4.0 square meters was simulated by averaging the 2.0 square meter data. Within the study area, 386 ground truth sites were established using the Global Positioning System (GPS) and verified as to their true ground cover. Initial computer classification produced 14 categories representing three levels of yellow starthistle infestation as well as other non-target objects such as grass, trees, bare ground, etc. These categories were redefined and collapsed to retain the three yellow starthistle classes (1: 90-100%, 2: 70-89%, and 3: 30-69%) as well as a fourth class encompassing everything else (non-starthistle).

Based on these classes and the ground truth sites, an error matrix for each flight date - resolution combination was developed. The error matrix for the 4.0 m^2 resolution on June 21, for example, is given in Table 1 (other error matrices are not reproduced here). This matrix indicates that there were 519 on the ground true pixels in Class 1 (90-100% yellow starthistle) of which 323 were correctly identified. The classified image resulted in 361 pixels for this class, indicating that in this case, pixels from other categories were misclassified as Class 1. A total of 1414 pixels were available at the ground truth sites for this resolution.

$\hat{K}_i$ values were computed for all classes at each resolution within each date. For simplicity, only a subset is shown here. Table 2 lists the $\hat{K}_i$ values for the 0.5 m^2 and 4.0 m^2 resolutions on the June 21 flight. In general, the 4.0 m^2 resolution had larger $\hat{K}_i$ values than that of 0.5 m^2, suggesting better agreement of 4.0 m^2 resolution with ground truth. Classes 2 and 3, 70-89% and 30-69% yellow starthistle , respectively, showed poor agreement at both resolutions. Class 4 (non-starthistle) had the highest $\hat{K}_i$ values, which is not surprising since this class represented all non-starthistle objects and, therefore, comprised the majority of the pixels and was the most easily detected.

Both Bayesian and asymptotic standard errors for each $\hat{K}_i$ are also given in Table 2. As might be expected, the larger sample sizes of the 0.5 m^2 resolution resulted in standard errors much smaller than those of the 4.0 m^2 resolution. While the results are similar between the two techniques, the Bayesian standard error estimates are slightly larger than their asymptotic counter parts. This is probably due to the approximate nature of the asymptotic technique and its reliance on large sample gaussian theory which assumes more known information about the parameters and the underlying distribution than the Bayesian posterior.

The Bayesian posterior distributions for $\hat{K}_1$ (Class 1) at each resolution on the first flight date are presented in Figure 2. This figure readily illustrates the effect of resolution on both the point estimates (most probable values or mode) as well as their associated variability (spread). Image resolutions 1.0, 2.0, and 4.0 m^2 provided similar $\hat{K}_1$ values with a large degree of overlapping, while clearly out performing the 0.5 m^2 resolution.

Probability intervals or credible regions developed from these distributions are shown in Figure 3. Class 4 (non-starthistle) and Class 1 (90-100% yellow starthistle) gave the best agreement with ground truth at all resolutions. The intermediate yellow starthistle Classes 2 and 3, had poor agreement at all resolutions. The 0.5 m^2 resolution appears to be in the worst agreement with ground truth. In order to quantitatively assess this difference, pair-wise comparisons were conducted among resolutions. The probability intervals for the differences between $\hat{K}_1$ values on June 21 are listed in Table 3. As might be inferred from Figures 2 and 3, only the 0.5 m^2 resolution shows any significant difference (interval does not cover zero).

Based on similar pair-wise comparison procedures, optimum resolutions for each flight time can be established (Table 4a). In cases where no clear optimum existed, larger resolutions were chosen for economic reasons. Smaller resolutions entail more flight time and result in larger image files which require more storage and, thus, are more expensive to acquire and process. Within the first flight date, the 4.0 m^2 resolution was optimum for all classes. The second flight date differed only in the intermediate yellow starthistle, Classes 2 and 3, where the 0.5 m^2 resolution worked best.

Using the results from Table 4a, the best flight date was selected from pair-wise comparisons of the optimum resolutions for each class (Table 4b). These indicated a preference for the later flight date in Classes 1 and 2, and no preference for Classes 3 and 4. Class 1, 90-100% yellow starthistle, is likely to give better results at later flight dates because more plants will be in bloom at that time, increasing their visibility. The lack of date preference in Class 4, all non-starthistle, is also to be expected since this class includes a conglomerate of objects such as roads, water, forest, crops, grass lands, and buildings which show up on the images equally well no matter when the image was aquired. Interpretation of flight time results of Classes 2 and 3 was deemed inappropriate since these classes demonstrated poor ground truth agreement (low accuracy) and, therefore, were unlikely to represent their designated intermediate yellow starthistle categories.

IV. CONCLUSIONS

Remote sensing accuracy is an important consideration in agriculture and land management. Cohen's Kappa, which is traditionally used to measure relative accuracy in remote sensing, is one of several accuracy statistics available. Bayesian techniques provide a means of variability assessment and comparison of the Kappa statistic based on its underlying discrete multinomial distribution. Correspondence between Bayesian results and those of the gaussian approximations were good, however, the similarity is expected to decrease with smaller sample sizes. This is due to the truncated nature of the data under these circumstances, whereby the Bayesian techniques would be more reliable, especially for the purpose of interval estimation. Further refinement of agreement measures may be necessary to fully reflect the correct probability model given ground truth data.

REFERENCES

Bishop, Y. M. M., S. E. Feinberg, and P. W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge MA, 557 pp.

Card, D. H. 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing. Vol. 48 No. 3, pp. 431-39.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. And Psychological Measurement. Vol. 20 No.1 pp. 37-47.

Coleman, J. S. 1966. Measuring concordance in attitudes. Mimeograph, Dept. Social Relations, Johns Hopkins Univ.

Congalton, R. G., and R. A. Mead. 1983. A qualitative method to test for consistency and correctness in photointepretation. Photogrammetric Engineering and Remote Sensing. Vol. 49 No. 1, pp. 69-74.

Congalton, R. G., R. Oderwald, and R. A. Mead. 1983. Assessing landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing. Vol. 49 No. 12, pp. 1671-78.

Congalton, R. G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing Environ. Vol. 37, pp. 35-46.

Hudson, W. D. and C. W. Ramm. 1987. Correct formulation of the Kappa coefficient of agreement. Photogrammetric Engineering and Remote Sensing. Vol. 53 No. 4, pp. 421-22.

Jaynes. E. T. 1968. Prior Probabilities. *Reprinted in* E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics, ed. R. Rosenkrantz. 1983. Dordecht: Reidel.

Lass, L. W., H. W. Carson, and R. H. Callihan. 1996. Detection of yellow starthistle (*Centauria solstitialis*) and common St. Johnswort (*Hypericum perforatum*) with multispectral digital imagery. Weed Tech. 10: 466-474.

Light, R. J. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bull. Vol 76 No.5, pp. 365-377.

Loredo, T. J. 1990. From Laplace to supernova SN1987A: Bayesian Inference in Astrophysics.

In Maximum Entropy and Bayesian Methods, P. F. Fougère (ed.). Kluwer Acad. Pub.,Dordecht, Neth. pp. 81-142.

Price, W. J. and B. Shafii. *In preparation.* A reevaluation of agreement measures in ground truth assessment of remotely sensed images. For submission to Photogrammetric Engineering and Remote Sensing.

Rosenfield, G. H. and K. Fitzpatric-Lins. 1986. A coefficient of agreement as a measure of thematic classification accuracy. Photogrammetric Engineering and Remote Sensing. Vol. 52 No. 2, pp. 223-27.

SAS Institute Inc. 1991. SAS Language: Reference, Version 6, First Edition. SAS Institute Inc., Cary, NC, 1042 pp.

Shannon, C. F. 1948. A mathematical theory of communication. Bell System Tech J. Vol. 27, pp 379-423 and 623-656.

Table 1. Error matrix for yellow starthistle detection on June 21, 1994.

		Ground Truth				
		*1**	*2*	*3*	4	
Classification	*1*	323	18	2	18	*361*
	2	34	6	1	17	*58*
	3	59	14	19	53	*145*
	4	103	29	39	679	*850*
		519	*67*	*61*	*767*	*1414*

*
1 = 90-100% yellow starthistle
2 = 70-89% yellow starthistle
3 = 30-69% yellow starthistle
4 = Non-yellow starthisle

Table 2. Conditional Kappa values and their associated Bayesian and asymptotic standard errors for 0.5 and 4.0 m^2 resolutions on June 21, 1994.

Resolution	Class*	$\widehat{K}_i$	Bayes SE	Asym. SE
0.5 m^2	1	0.3072	0.0034	0.0022
	2	0.0526	0.0048	0.0034
	3	0.0360	0.0068	0.0056
	4	0.6827	0.0043	0.0045
4.0 m^2	1	0.4929	0.0285	0.0204
	2	0.0506	0.0364	0.0247
	3	0.2328	0.0661	0.0423
	4	0.7124	0.0289	0.0299

*
1 = 90-100% yellow starthistle
2 = 70-89% yellow starthistle
3 = 30-69% yellow starthistle
4 = Non-yellow starthisle

Table 3. Pair-wise differences and the associated 95% probability intervals for all resolutions. Differences are based on $\hat{K}_1$ (90-100% yellow starthistle).

Resolution	Lower	Difference	Upper
1m vs 2m	-0.0009	0.0262	0.0519
1m vs 4m	-0.0466	0.0021	0.0499
1m vs .5m	0.1742	0.1878	0.2003
2m vs 4m	-0.0774	-0.0241	0.0275
2m vs .5m	0.1363	0.1616	0.1850
4m vs .5m	0.1363	0.1857	0.2306

Table 4. Optimum resolutions and flight times for the four yellow starthistle classes.

Table 4a.

Flight Time	Class*	Resolution**
Early (June)	1	4m (1m)
	2	4m (2m)
	3	4m
	4	4m
Late (July)	1	4m
	2	.5m
	3	.5m
	4	4m (2m)

Table 4b.

Class*	Flight Time
1	Late
2	Late
3	Early/Late
4	Early/Late

*
1 = 90-100% yellow starthistle
2 = 70-89% yellow starthistle
3 = 30-69% yellow starthistle
4 = Non-yellow starthisle

** Values in parentheses denote actual optima. Larger resolutions were chosen for economic reasons.

Figure 1. Grey scale image of yellow starthistle classification. Lighter areas indicate higher levels of yellow starthistle.

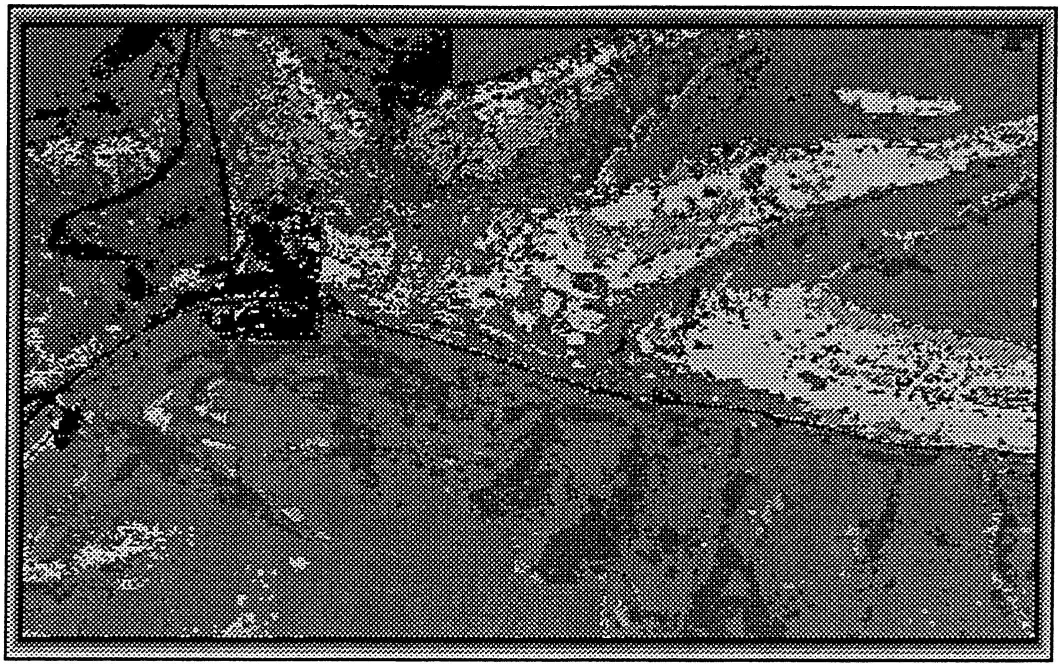


Figure 2. Posterior distributions of $\hat{K}_1$ for the four image resolutions.

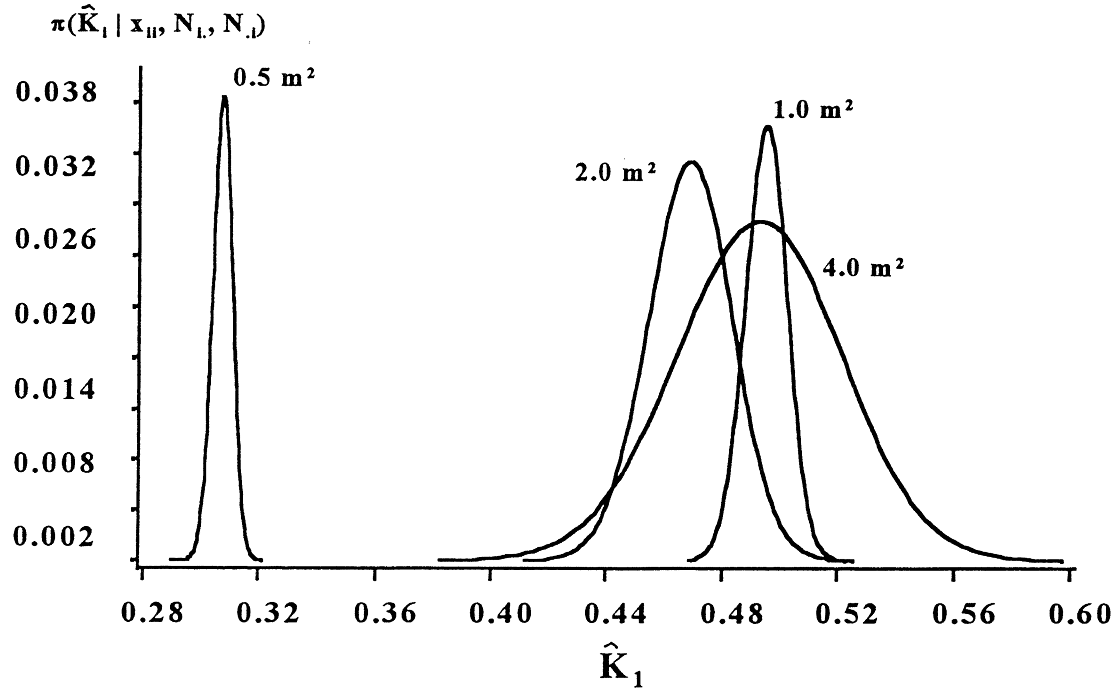
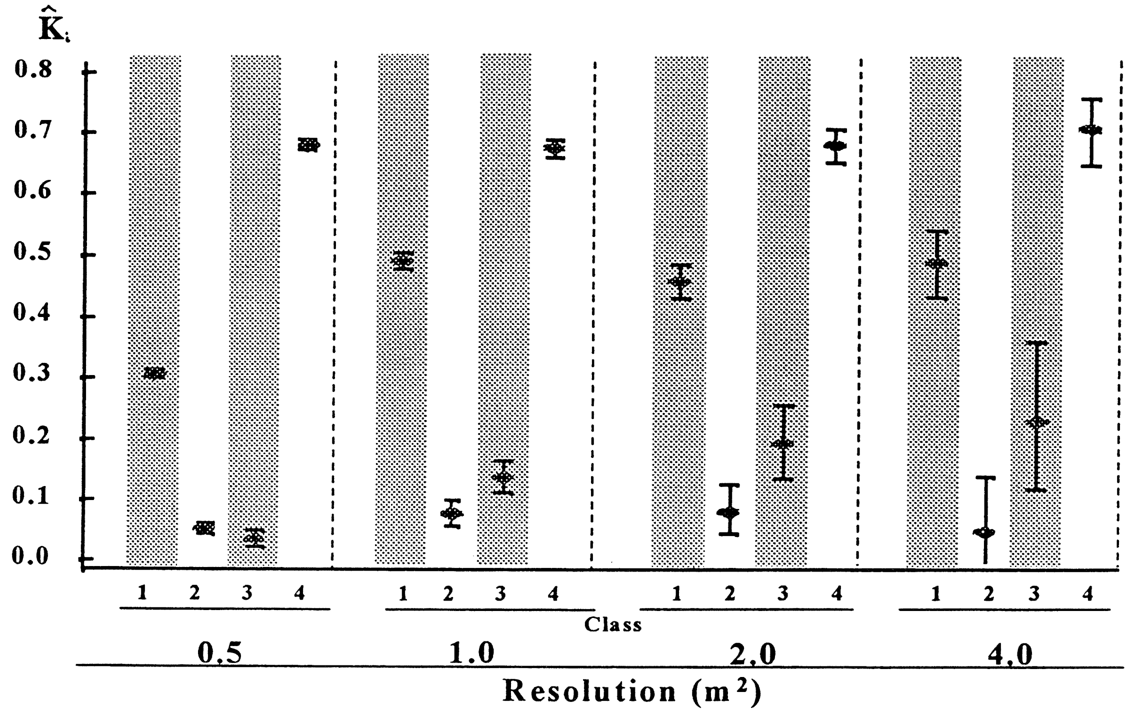


Figure 3. 95% probability intervals for $\hat{K}_i$ at all image resolutions on June 21, 1994.



1 = 90-100% yellow starthistle
2 = 70-89% yellow starthistle
3 = 30-69% yellow starthistle
4 = Non-yellow starthisle