Estimation of *a Priori* Probabilities of Landcover Categories for Bayes' Classifier

Kunihiko Yoshino, Keiji Kushida
Assistant Professor
The Faculty of Agriculture
University of Tokyo, JAPAN

XVIII ISPRS Congress Vienna 1996
Commission III

KEY WORDS: Decomposition of mixed data, A Priori Probability, Landcover Classification, Bayes' Classifier, Landsat TM

ABSTRACT

A priori probabilities of landcover categories in the study area improve the landcover classification accuracy, although the probabilities are very difficult to estimate in advance of the analysis. Algorithms for the decomposition of mixels to pure landcover categories were developed to estimate landcover area ratios in mixels. If the study area is supposed to be a very large mixel which contains several landcover categories, some of the decomposition algorithms of mixed data can be applied to the centroid vector of the study area. The area ratios of landcovers in the study area are equal to the *a priori* probabilities of landcovers.

The algorithm of maximum likelihood estimation was applied to estimate the *a priori* probabilities of landcovers in the study area in this research. As a result of this research, the estimation algorithm worked well and the *a prior* probabilities of landcovers in seven small study sites were estimated very well. Moreover, those estimated *a priori* probabilities of landcovers improved the accuracy of landcover classification in the study sites.

1. Introduction & Objectives

Supervised classification has been used for most landcover classification of remotely sensed images. The maximum likelihood classification algorithm has been generally used since it is simple and reasonable in terms of statistics.

This algorithm has some assumptions in particular. One of the assumptions is that the Bayes' method needs the *a priori* probabilities of landcovers, but estimating the *a priori* probabilities of landcovers in the study area in advance of the analysis is very difficult. Therefore, most of the landcover research studies have used even weights of *a priori* probabilities for the maximum likelihood classification. As Strahler(1980) reported, the use of *a priori* probabilities of landcovers in the study area for the maximum likelihood classifier improves the classification accuracy.

The *a priori* probabilities of landcovers in the study area are equal to the area ratios of landcovers in the study area following the definition of *a priori* probability. The area ratios of landcovers in the study area can be stimated from the mean valu es of multispectral data of the study area, applying the algorithm of the decomposition of mixed data. Then, we can improve the landcover classification accuracy of the maximum likelihood method by introducing the area ratios of landcovers in the study area.

This paper has three objectives. The first objective is to show that the algorithm of decomposition of mixed data can be applied to estimate area ratios of landcovers as *a priori* probabilities in the study area. The second is to show that *a priori* probabilities of landcovers can be estimated well by this algorithm. The last is to show that the introduction of *a priori* probabilities of landcovers to the Bayes' classification results in good accuracy of landcover classification.

2. Method to estimate the area ratios of landcovers

The *a priori* probabilities of landcovers in the study area are supposed to be equal to the area ratios of landcovers in the study area. Equation 1 shows that the mean vector of the study area can be expressed as the summation of the weighted mean vectors of the training data. The weights of the mean vectors of the training data are the *a priori* probabilities of the training categories.

$$
\begin{aligned}
Io[k] &= \sum_{i,j} X[i,j,k] / \sum_{i,j} \\
&= \sum_{i,j} (\sum_{h=1}^{H} a[h,i,j] * It[h,k]) / \sum_{i,j} \\
&= \sum_{h=1}^{H} (\sum_{i,j} a[h,i,j]) / \sum_{i,j} * It[h,k] \\
&\fallingdotseq \sum_{h=1}^{H} A[h] * I[h,k] \text{ ....(eq.1)}
\end{aligned}
$$

where,

$$X[i,j,k] = \sum_{h=1}^{H} a[h,i,j] * It[h,k] \text{...(eq.2)}$$

$$I[h,k] \fallingdotseq It[h,k] \text{ ....(eq.3)}$$

$$a[h,i,j] \geq 0, \quad \sum_{h=1}^{H} a[h,i,j] = 1$$

$$A[h] = \sum_{i,j} a[h,i,j] \ / \sum_{i,j}$$

$$A[h] \geq 0, \quad \sum_{h=1}^{H} A[h] = 1$$

(Mean vectors of landcovers in the study area are very close to those of the training data.)

Note:

- H: Number of landcovers
- K: Number of bands of image data
- h: Landcover ID No. (1 to H)
- i,j: X,Y-coordinates of the image in the study area
- k: Band ID No. (1 to K)
- Io[k]: Mean vectors of the study area
- I[h,k]: Mean vectors of Landcover [h] in the study area
- It[h,k]: Mean vectors of landcover training data [h]
- A[h]: Area ratio of landcover[h] in the study area(unit: IFOV)
- a[h,i,j]: Area ratio of landcover[h] in the IFOV of pixel[i,j] (unit: IFOV)
- X[i,j,k]: Image data vector pixel[i,j]

The area ratios of landcovers in the study area can be estimated from the image by applying the algorithm of the decomposition of mixed data. We can improve the landcover classification accuracy by introducing the area ratios of landcovers into the Bayes' classifier as *a priori* probabilities of landcovers in the study area.

Methods to decompose mixed data, especially mixels, have been presented [Inamura 1987, Ito 1987,Matsumoto 1991]. The algorithm which was presented by Matsumoto (1991) was applied to decompose mixed data into landcover categories in this reseach. This algorithm performs better than other algorithms, because it takes account variances of data distribution.

The method to decompose mixed data which has been presented by Matsumoto [1991] is shown in the Appendix A.

3.Results of the estimation of the area ratios of landcovers

Innba district, an area near Tokyo, was selected as the study area. The ground truth was conducted on March 16, 1994 around this area. The seven bands of TM image data(path 107 row 35,radiometrically calibrated, resampled by cubic convolution) of Landsat-5 acquired on April 22, 1994 were analyzed.

Seven training categories(Paddy field, Forest,Upland field, Waste land, Water body, Residential area, Sports ground) were selected. These training areas were selected from the whole study area, and then their signatures were calculated in this area. Moreover, separability of those training data was repeatedly examined. Seven small study sites, where the area ratios of landcovers were also measured using the ground truth data, were selected from the study area. Each small study site had two to five landcovers (Table 1). After the mean vectors of each small study site were computed, they were decomposed into landcover categories by the method of the maximum likelihood estimation.

Landcovers on the ground truth map were digitized and those areas were computed on the high resolution digital image data. On the other hands, classified pixels to landcovers in the study sites were counted, considering pixels on the boundary of study sites. The pixel on the boundary was assumed that that pixel contained 50% of the landcover category to which that pixel was classified. Then, the area of each landcover was computed by multiplying the area of IFOV on the ground. The estimation errors of the landcovers in each small study site are shown in Table 2. The area ratios of landcovers in each small study site was estimated well.

Table 1. Landcover Categories in the small study sites on Ground Truth Map (Unit:%)

Study Sites	area-1	area-21	area-22	area-31	area-32	area-33	area-5
Paddy	0	20.6	0	1.2	1.0	1.2	0
Forest	32.7	69.1	91.8	37.2	47.7	68.8	44.6
Upland	55.1	3.1	3.8	42.7	30.0	5.5	30.4
Waste	0	3.8	0	5.3	3.3	0	6.3
Water	0	0	0	0	0	0	0
Residence	11.7	3.3	4.4	13.6	12.9	24.5	18.7
Ground	0	0	0	0	5.1	0	0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 2. Estimation Root Mean Square Errors of Landcovers in study sites (Unit:%)

Study Sites	area-1	area-21	area-22	area-31	area-32	area-33	area-5
RMSE	5.4	5.2	3.5	5.7	4.6	1.8	5.0

$$RMSE(k) = \sqrt{\frac{\sum_{i=1}^{N(k)} (A(k,i) - Ar(k,i))^2}{N(k)}}$$

k: Study Site ID
N(k): Number of landcovers in the study site (k)
A(k,i): Estimated area ratio of landcover(i) in the study site(k) by the decomposition of mixed data
Ar(k,i):Actual area ratio of landcover(i) in the study site(k)

Table 3. Effect of introducing the estimated area ratios of landcovers as *a priori* probabilities in study sites (Unit:%)

Study Sites	area-1	area-21	area-22	area-31	area-32	area-33	area-5
RMSE-A	7.75	2.10	2.48	6.93	7.78	7.40	6.22
RMSE-B	2.45	2.24	0.92	3.36	3.57	1.58	2.54

RMSE-A: Maximum Likelihood Classifier with even *a prio*probabilities of landcovers
RMSE-B: Bayes' Classifier with estimated *a priori* probabilities of landcovers

4.Effect of introducing the estimated area ratios of landcovers as *a priori* probabilities

The effect of introducing the estimated area ratios of landcovers as *a priori* probabilities for the Bayes' classifier is shown in Table 3. The digitized ground truth map was compared with the classification results. The classification errors were evaluated by the RMSE as shown in the following equation.

$$RMSE(k) = \sqrt{\frac{\sum_{i=1}^{N(k)} (A(k,i) - Ar(k,i))^2}{N(k)}}$$

k: Study site ID
N(k): Number of landcovers in the study site (k)
A(k,i): Estimated area ratio of landcover(i) in the study site(k) by classification
Ar(k,i):Actual area ratio of landcover(i) in the study site(k)

By introducing the estimated area ratios of landcovers as *a priori* probabilities, the accuracy of landcover classification was significantly improved. The RMS errors of landcover area ratios in each small study site by Bayes' classification were less than those estimated only by maximum likelihood estimation.

5. Discussion & Conclusion

The results of the experiment to estimate the landcover area ratios as *a priori* probabilities of landcovers showed that the quality of training data affects the accuracy of landcover area ratios. So, a new algorithm like an experiment[Yoshino(1995)] is needed to be developed in order to determine proper training data sets.

As a result of this research, the following are concluded.

1) An algorithm to decompose mixed data can be used to estimate the area ratios of landcovers in the study area as *a priori* probabilities for the Bayes' classifier.
2) The area ratios of landcovers were estimated fairly well by the algorithm of the decomposition of mixed data.
3) Introducing estimated area ratios of landcovers as *a priori* probabilities for the Bayes' classifier resulted in better accuracy of landcover classification by maximum likelihood classification.
4) It was found that this new algorithm to estimate area ratios of landcovers in the study area strongly requires very good training data in the study area.
5) Estimating the number of proper landcovers for the study area and selecting good training signatures are very important.
6) A new procedure to determine proper training datasets for the estimation of landcover area ratios must be developed.
7) The results of this research will be very useful for hyper multispectral images like AVIRIS data.

This research was supported by the Sasakawa Scientific Research Grant from the Japan Science Society in 1995. We are greatly indebted to many students in the department of agricultural engineering of the University of Tokyo for their assistance at the time of the ground truth.

Reference

1)Inamura,Minoru: Analysis of Remotely Sensed Image Data by means of Category Decomposition, Trans. of the Institute of Electronics, Information and Communication Engineers, Vol.J70-C(2),pp.241-250, 1987
2)Ito,Tadashi and Sadao Fujimura: Estimation of Cover Area of Each Category in a Pixel by Pixel Decomposition into Categories, Trans. of the Society of Instrument and Control Engineers Vol.23(8),p.20-25,1987
3)Matsumoto,Masao, Kouki Fujiku, Kiyoshi Tsuchiya,Kohei Arai: Category Decomposition based on Maximum Likelihood Estimation, Journal of the Japan Society of Photogrammetry, Vol.30(2), p.25-34,1991
4)Strahler, Alan H.: The Use of Prior Probabilities in Maximum Likelihood Classification of Remotely Sensed Data, Remote Sensing of Environment 10, pp.135-163, 1980
5)Yoshino, Kunihiko: One procedure to determine landcover categories for landcover supervised classification in remotely sensed image, Proc. of the annual conference of Japan Society of Irrigation and Drainage, Reclamation Engineering, p.82-83, 1995

Appendix A. The method of Matsumoto's[1991] to decompose mixed data; The Maximum Likelihood Estimation

The spectral reflectance(**I**) of K bands from the mixel which contains of H landcover categories is expressed like the Eq. 1; it is the multiplication of the mean vector(**M**) of the spectral reflectance from the pure pixels of each landcover category and the mixture ratios(**A**) of landcovers in the mixel.

$$I = M * A \qquad \text{-----Eq.1}$$

where,

$$I = (I_1, I_2, \ldots\ldots, I_K)$$

$$A = (a_1, a_2, \ldots\ldots, a_H)^t$$

$$M = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1H} \\ & & \cdots & \\ & & \cdots & \\ & & \cdots & \\ m_{K1} & m_{K2} & \cdots & m_{KH} \end{pmatrix}$$

Now, suppose the reflectance(**I**ij)of the band (i) on the landcover category (j) has a normal distribution with mean m^*ij and variance σ^2. Then, the mean vector (**M**i) of the spectral reflectance of the band (i) is expressed like Eq.3. The variance of the spectral reflectance is expressed like Eq.4.

$$M_i = M_i^* * A \qquad \text{---------Eq.3}$$

$$Mi^* = (m^*_{i1}, m^*_{i2}, \ldots, m^*_{iH})$$

$$\sigma^2_i = A^t * S_i * A \qquad \text{--------Eq.4}$$

$$S_i = \mathrm{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \ldots\ldots, \sigma_{iH}^2)$$

Then, the likelihood;P(Ii), which the reflectance data of the band i may be observed, is expressed like the Eq. 5. Supposing that the spectral reflectance is independent among bands, the total likelihood; **P(I)**, which the spectral reflectance I1,I2,....,IK may be observed, is expressed like Eq. 6 and 7.

$$P(I_i) = \frac{\exp\{-(I_i - M_i)^2/(2*\sigma_i^2)\}}{\sigma_i * \sqrt{(2\pi)}} \qquad \text{----Eq.5}$$

$$P(I) = \prod_{i=1}^{N} P(I_i) \qquad \text{----Eq.6}$$

$$Q = -\ln\{P(I)\} \qquad \text{----Eq.7}$$

$$\sum_{j=1}^{k} a_j = 1, 1 \geq a_j \geq 0, (j = 1, 2, \ldots, k) \qquad \text{---- Eq.8}$$

Note:

H: Number of landcovers
K: Number of bands of image data
i: Band No. (1 to H)
j: Landcover ID.
ai: Area ratio of landcover[h] in the IFOV of pixel[i,j] (unit: IFOV)

The mixture ratios of landcovers in a mixel can be computed by obtaining **A**, which minimizes **Q** under the conditon which expressed in the Eq.8. This is the method to decompose mixed data using the Maximum Likelihood Estimation developed by Matsumoto in 1990.