Journal of the Korean Society of Remote Sensing, Vol.15, No.3, 1999, pp.195~208

유전자 알고리즘을 이용한 트레이닝 최적화 기법 연구
- 정규분포를 고려한 통계적 영상분류의 경우 -

어양담* · 조봉환** · 이용웅*** · 김용일****

서울대학교 공학연구소 특별연구원*, 국방과학연구소 위성영상처리팀 책임연구원**

국방과학연구소 위성영상처리팀 선임연구원***, 서울대학교 지구환경시스템 공학부 조교수****

A Study on the Training Optimization Using Genetic Algorithm
-In case of Statistical Classification considering Normal Distribution-

Yang-Dam Eo*, Bong-Whan Cho, Yong-Woong Lee***, and Yong-Il Kim****

Researcher, Research Institute of Engineering Science, Seoul National University*

Principal Researcher, Satellite Image Processing and Analysis Team, Agency for Defense Development**

Senior Researcher, Satellite Image Processing and Analysis Team, Agency for Defense Development***

Assistant Professor, Department of Urban Engineering, School of Civil, Urban & Geosystem Engineering, Seoul National University****

Abstract : In the classification of satellite images, the representative of training of classes is very important factor that affects the classification accuracy. Hence, in order to improve the classification accuracy, it is required to optimize pre-classification stage which determines classification parameters rather than to develop classifiers alone. In this study, the normality of training are calculated at the pre-classification stage using SPOT XS and LANDSAT TM. A correlation coefficient of multivariate Q-Q plot with 5% significance level and a variance of initial training are considered as an object function of genetic algorithm in the training normalization process. As a result of normalization of training using the genetic algorithm, it was proved that, for the study area, the mean and variance of each class shifted to the population, and the result showed the possibility of prediction of the distribution of each class.

Key Words : class, sample pixel, training normalization, classification accuracy

요 약 : 위성영상 분류작업에서 분류클래스에 대한 샘플화소의 대표성은 분류 정확도에 많은 영향을 미친다. 따라서, 통계적 영상분류방법에서는 분류 기법 자체보다 분류 확률을 결정하는 트레이닝 단계, 즉 샘플화소의 최적화가 필요하다. 본 연구에서는 SPOT XS, LANDSAT TM을 이용한 위성영상 화소분류작업에서 분류 이전단계, 즉 샘플화소의 정규성을 계산하여, 정규성에 악영향을 미치는 화소를 객관적 기준으로 조정하였다. 정규화과정을 위한 유전자 알고리즘 적용의 생존확률 평가함수로 다변량 Q-Q plot의 상관계수와 트레이닝의 분산값을 고려하였으며, 5% 유의수준을 적용

1999년 6월 2일 접수 ; 1999년 7월 10일 수리.

하였다. 연구결과, 실험대상지역의 경우, 유전자 알고리즘을 이용한 트레이닝 정규화 결과가 대부분의 클래스에 대하여 그 평균과 분산을 모집단에 근사시키고 있다는 것을 입증하였고, 해당 클래스의 모집단 분포를 예측할 수 있는 가능성을 제시하였다.

1. 서 론

위성영상에 대한 분석은 화소특성을 부여하여 영상 내 지형정보를 분류(classification)하고 그 정확도(classification accuracy)를 향상시키는 과정이 수반되어야 하는데, 이에 대한 기준과 방법이 일련의 체계화된 공정으로 이루어지지 않아 클래스 설정, 트레이닝 작업 등에 사용자 자신의 주관적 요소가 많이 개입되게 마련이다. 특히 영상화소를 통계적 방법으로 분류할 경우, 분류에 필요한 매개변수(parameter)는 트레이닝(training) 작업으로부터 얻어지는데, 트레이닝 자체가 작업 주관이 개입되어, 동일영상이라 하더라도 작업자에 따라 분류작업결과가 상이하게 나올 수 있다(Swan and Davis., 1978). 결국 이것은 인공위성영상을 통해 지형정보를 얻기까지 일련의 작업을 표준화시키는데 문제시되어 왔고, 대규모 지역을 여러 개의 영상으로 나누어 획득했을 경우, 객관적인 작업공정과 정확도 평가가 이루어지지 않아 위성영상활용범위를 제한하고 있다.

특히, 통계적 분류방법에서는 샘플화소의 밝기 분포와 그에 따른 매개변수가 클래스의 대표성으로 가정되지만, 샘플링 자체에 주관적 요인이 많이 포함되어 실제 분류확률이 분포의 가정에 적합하지 않을 수 있으므로, 정확도 향상과는 별도로 그 대표성에 문제가 될 수 있고 이것은 분류이전 단계의 최적화 문제로 된다. 따라서 분류 알고리즘의 개발보다 대표성을 가진 트레이닝 자료선정 기법의 설계와 개발이 원격탐사 연구분야발전에 필요하며, 영상분류정확도 향상에 기여를 할 것이다(Hixon et al., 1980).

트레이닝 샘플에 관한 연구로는 Hixon이 정확한 분류결과를 얻기 위해서는 분류 알고리즘 선정보다 트레이닝 샘플의 대표성에 관한 연구가 더 중요하다고 주장한 바 있고(Hixon et al., 1980), Van Deusen은 트레이닝 데이터의 대표성을 갖추기 위한 각 클래스별 샘플화소비율을 확률이론에 근거하여 연구한 바 있다(Van, 1996). 또한 新井康平은 간략화 β분포에 의한 확률밀도함수적용형 화상분류를 연구한 바 있다(新井康平, 1998).

영상화소에 대한 클래스별 분류확률을 결정하기 위한 매개변수 결정은 트레이닝 샘플화소로부터 결정되는데, 항상 그 대표성이 문제시되어 왔다. 그러나 현재 위성영상에 대한 통계적 분류기법과 관련한 연구에 의하면, 정확도 평가에서 트레이닝의 정규성(normality)을 검사하여 분류에 어떤 영향을 미쳤는가를 서술할 뿐이었다. 본 연구는 트레이닝 샘플화소의 대표성을 분류자에서 가정하는 정규분포(normal distribution)에 근사시키고, 트레이닝된 화소에 대한 정규성 검증(normality check)을 실행하여 분류자에 가장 적합한 매개변수(parameter)를 제공하려고 하였다. 이를 위해 먼저 정규성 검증, 이중에서도 다중밴드 영상(multispectral image)에 대한 적용을 위해 다변량 정규성 검증방법을 고찰하였고, 화소의 정규분포를 구축하기 위한 도구로는 최근에 최적화 프로그램으로 많이 활용되고 있는 유전자 알고리즘을 채용하였다.

2. 통계적 영상화소분류에서의 트레이닝

통계적 분류방법은 일반적으로 클래스와 관련된 확률함수를 이용한다. 그런데 확률함수는 항상 미지(unknown)이고, 그것은 트레이닝으로부터 측정된다. 만약 확률분포를 기지(known)의 값으로 놓는다면 트레이닝 데이터로부터 필요한 몇 개 정보(샘플화소의 클래스별 평균과 분산값)만을 사용하게 된다. 대부분의 판별함수(discriminant function)는 트레이닝 패턴(training pattern), 즉 대

상이 되는 클래스(class)를 대표한다고 고려되는 기지(known)의 identity 측정벡터(measurement vector)세트의 정보로부터 유추된다. 따라서 트레이닝의 목적은 전체 영상 데이터 세트에서 각 화소의 분류를 위한 결정 규칙(decision rule)을 결정하는데 쓰일 수 있는 스펙트럴 데이터 세트를 얻는 것이다(Swan and Davis, 1978).

트레이닝이 갖춰야 할 특성으로는 해당 클래스의 대표값을 가지고 있어야 하고, 각 클래스에 대한 분류결정규칙에 근거한 분포가정에 적합할 만큼 근접해 있어야 한다는 것이다. 주로 클래스를 설정하고 트레이닝을 하는 전문가는 그의 경험과 전문지식(tone, shape, size, texture, pattern, association 등)으로 실행한다. 이때 얻어지는 매개변수를 이용한 통계적 분류는 현실적으로 각기 다른 토지피복은 각기 다른 확률 모델을 가지고 있으며, 각 클래스의 스펙트럴 특성도 상황에 따라 불규칙적으로 변하기 때문에 문제가 되어 왔다(Kartikeyan et al., 1995). 또한 트레이닝은 전문가의 모든 전문지식을 동원해서 이루어지나 정작 분류는 분류자(classifier)가 채용하고 있는 분포의 가정에 의해서만 실행되는 것이 통계적 분류방법이었다. 따라서 분류시 가정한 화소값 분포 모델과 영상 내 클래스별 화소값 분포가 얼마나 근사하고 있는가의 문제와 이와 관련한 클래스 설정의 문제 그리고 분류에 이용할 밴드 조합의 문제가 서로 유기적으로 연결되어 영상분류 정확도에 영향을 미치게 된다.

3. 다변량 정규성

원격탐사 데이터는 다중파장대 분석을 근간으로 하므로 정규확률밀도분포함수에 대한 다차원 분석을 해야 한다. 이것은 실제 데이터가 정확히 다차원 정규분포가 아니더라도, 정규밀도는 실제 모집단분포로 유추하는데 유용하기 때문이다(Johnson and Wichern, 1992).

다변량 표본의 정규성을 검정하는 첫 단계는 각 차원별 분포(marginal distribution)을 검정해보는 것이다. 다변량의 비정규성은 marginal distribution이나 scatterplot에 반영되기 때문에 실질적인 작업에 있어서는 일차원이나 이차원의 검정만으로 충분한 경우가 대부분이다. 그러나 일변량에서 정규성이 검정되었다고 하여, 다변량에서 정규성을 나타낸다는 보장은 없다. 따라서 다변량에서의 정규성 검정을 해주어야 한다(Johnson and Wichern, 1992). 다변량에서의 정규성 검증방법은 여러 가지가 있으나 본 연구에서 채택하려고 하는 Q-Q plot을 이용하는 것과 왜도와 첨도를 이용하는 방법에 대해서만 서술하기로 한다(Johnson and Wichern, 1992).

Q-Q plot을 이용하는 방법은 다음과 같은 순서로 이루어진다.

단계1 : 식 (1)과 같이 정의된 d^2_j을 계산하고 $d^2_{(1)} \le d^2_{(2)} \le \cdots \le d^2_{(n)}$ 이와 같이 그 크기 순으로 정렬시킨다.

$$d^2_j = (x_i - \bar{x})^T S^{-1}(x_i - \bar{x}),\ i=1,2,\ldots,n \quad \ldots(1)$$

단계2 : $\chi^2_p((i-\frac{1}{2})/n)$ 를 자유도 p를 가지는 chi-square 분포의 $100(i-\frac{1}{2})/n$ percentile이라 할 때, $(d^2_{(i)}, \chi^2_p((i-\frac{1}{2})/n))$의 쌍을 plotting한다.

단계3 : 일변량에서의 분석방법과 같이 시각적 직선성 분석이나 상관도 분석을 한다.

다음으로는 왜도와 첨도를 이용하는 것이다(Rencher, 1995).

관측값을 다음의 행렬$(n \times p)$ matrix

$$Y = \begin{bmatrix} y_{11} & y_{12} \cdots y_{1p} \\ y_{21} & y_{22} \cdots y_{2p} \\ \vdots & \vdots \quad \vdots \\ y_{n1} & y_{n2} \cdots y_{np} \end{bmatrix},\ y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})^T,$$

$$\bar{y} = (\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_p,)^T,\ S = \frac{1}{n-1}\sum_i (y_i - \bar{y})(y_i - \bar{y})^T$$

이라 둔다. 다변량에서의 왜도는

$$b_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(y_i - \bar{y})^T S^{-1}(y_i - \bar{y})]^3 \qquad (2)$$

이 되고, 첨도는

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \bar{y})^T S^{-1}(y_i - \bar{y})]^2 \qquad (3)$$

이 된다. b_1에 대하여, 아래의 추정량은 자유도 $\frac{1}{6}p(p+1)(p+2)$를 갖는 카이제곱분포에 근사한다.

$$\chi^2 = \frac{nb_1}{6}$$

또한 b_2에 대하여는 다음의 두 추정량을 사용한다.

$$z_1 = \frac{b_2 - p(p+2)}{\sqrt{8p(p+2)/n}} \quad (4)$$

$$z_2 = \frac{b_2 - p(p+2)(n+p+1)/n}{\sqrt{8p(p+2)/(n-1)}} \quad (5)$$

upper 2.5% points에 대해서는 식4를 사용하며 이 추정량은 N(0, 1)에 근사한다. lower 2.5% points z_2에 대해서는 그 관측량에 따라 두 가지의 경우로 나뉘는데 $50 \leq n \leq 400$인 경우에는 식 (5)의 를 사용하며 $n \geq 400$인 경우에는 식 (4)의 z_1을 사용한다. 식 (4), 식 (5)의 추정량들도 $N(0, 1)$에 근사한다(Basilevsky, 1994).

4. 유전자 알고리즘의 도입

본 연구에서 유전자 알고리즘을 이용한 최적화 과정을 고려하고 있는 이유로는 샘플화소에 대한 정규성 검증시 조합의 수에 따른 문제와 트레이닝 정규화 이후 특성 값의 변질위험을 들 수 있다. 임의 클래스의 n개 샘플화소에 대한 정규성 검증 조합의 수는 2^n-1개로 된다. 이것을 현재 보급되어 있는 PC로 계산한다면 너무나 많은 시간이 소요되어 그 실용성에 문제가 된다. 더구나 영상분류 결과 대부분이 3개 클래스 이상인 것을 감안다면, 이를 일일이 계산하여 최적의 조합을 찾는 것은 무리이다. 따라서 실질적인 방법으로 최적화 프로그램을 고려하고 있다. 또한 트레이닝된 샘플화소 중 정규성에 악영향을 미치는 화소를 소거하는 과정에서 다량의 소거로 인한 국소해(local optima)로 수렴되어 전문가가 고려한 트레이닝 특성을 잃어버릴 가능성도 배제할 수 없다. 따라서, 분산의 현격한 축소를 방지하면서 샘플화소수를 최대한 유지하고, 정규성을 높여야 하는 문제를 해결해야 한다. 이를 위해서는 두 개이상의 수렴조건을 설계변수에 적용시킬 수 있어야 하는 것이다.

이러한 측면에서 볼 때, 유전자 알고리즘은 사용자의 목적함수 구성에 무리가 없고, 특히 잠재적인 해의 모집단을 유지하면서 다중방향으로 탐색을 한다. 즉, 집단에서 집단으로의 접근은 지역해로 정착되는 것을 방지해주는 역할을 한다.

유전자 알고리즘에 의한 최적화 방법은 자연선택과 유전체계에 입각한 것으로서, 설계변수를 유전인자로 설계결과를 개체로 생각하고 유전인자의 조작에 의해 개체를 설계목적에 맞게 진화시키는 방법이다. 이때 진화는 한 개체만으로 이루어지는 것이 아니라, 여러 개체가 모여 이루어진 집단을 단위로 이루어진다. 다시 말해서 집단내의 개체들이 세대(generation)를 거치면서, 교배(crossover), 변이(mutation)와 유사한 과정으로 번식하여 점점 더 우수한 특성을 지닌 개체들이 만들어지게 되는 원리이다.(Fig. 2 참조)

본 연구에서는 초기 모집단의 발생을 최대화소로부터 순차적으로 발생시켜 화소수의 극단적인 축소로 인한 초기 트레이닝 속성의 소멸을 막고, 샘플화소조합으로부터의 정규성 정도와 분류에 가정한 분포함수와의 적합성을 생존확률 기준으로 채택하였으며, 이와 함께 국소수렴으로 인한 분산값의 현격한 감소를 막기 위해 생존확률에 분산값에 대한 고려도 하고 있다.

다변량 정규성을 검정하는 기법 중 왜도 및 첨도의 경우, Fig. 1과 같이 정규분포성이 적은 확률

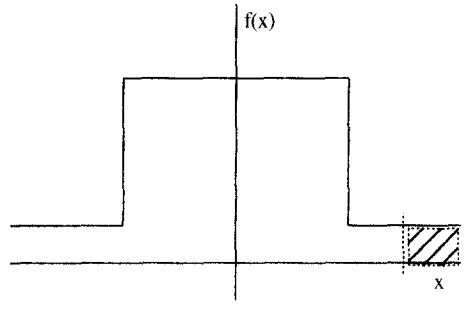

Fig. 1. The example of a probability distribution problem

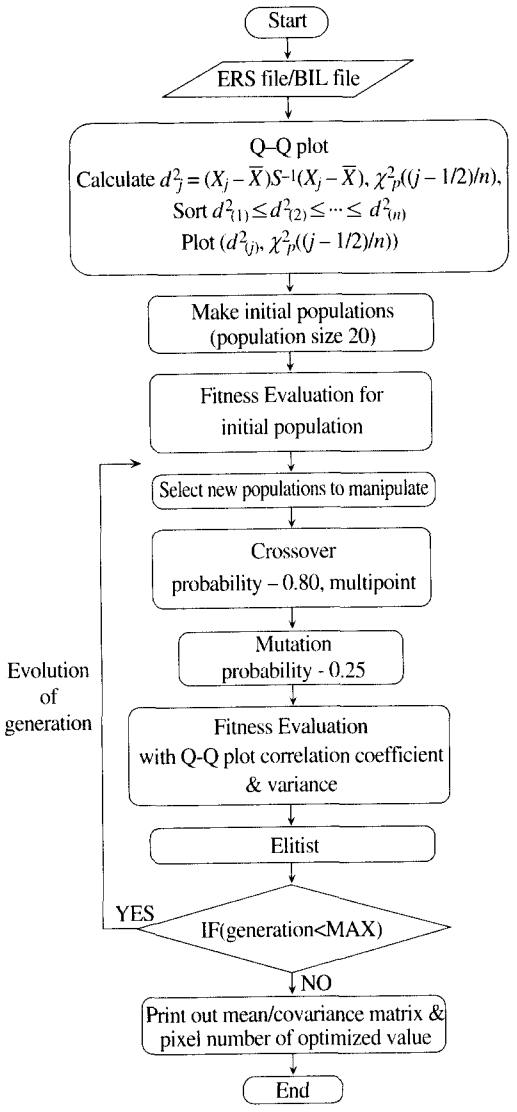

Fig. 2. The training optimization flow using genetic algorithm

분포함수의 경우에도 그 추정량이 정규분포에 근사한 것으로 판단될 수 있다. 이것은 왜도가 대칭성만을, 첨도가 꼬리의 두터운 정도만을 고려하고 있기 때문이다.

이에 대하여 Q-Q plot은 표준정규분포 확률 함수형태와의 상관성을 지표로 삼고 있기 때문에 Fig. 1의 문제는 충분한 샘플수만 획득할 수 있다면 해결될 수 있다.

또한 유전자 알고리즘 적용시 왜도와 첨도를 동시에 고려한다면, 평가함수(Evaluate Fuction)구성에서 양쪽에 대한 가중치(weight) 부여 조정의 문제가 생긴다. 이에 대하여 Q-Q plot은 수렴조건으로 상관계수만을 고려하기 때문에 평가함수 구성에 문제가 없다. 따라서 본 연구에서는 트레이닝 샘플화소에 대한 다변량 정규성 기준으로 Q-Q plot을 채택하였다. 그런데 샘플화소조합마다 식(1)에서 평균값이 이동되어 화소를 조합시킬 때마다 Q-Q plot실행을 해야 하는 문제가 발생하였다.

이러한 문제는 결국 화소조합의 최적화문제로 되고 이것을 해결하기 위한 수단으로 본 연구에서는 유전자 알고리즘을 적용하였다. 각 클래스별 트레이닝 화소값들을 인식하고, 개개화소들을 조합시켜 유의수준 내 정규화 한계 내로 들어오는 화소조합 경우를 채택하기로 한다. 이 작업의 역할은 트레이닝 샘플화소값들의 분포에 대하여 그 정규성을 향상시키는 것으로, 특정 유의 수준의 의미는 없다.

최적화 과정에서 특히 수렴에 이르게 하는 평가함수가 생존확률을 만들게 하는데, 본 연구에서는 목적함수(object function)로 트레이닝 화소값의 분산과 상관계수를 동시에 고려하였다.(식 6 참조) 분산값과 Q-Q plot 상관성을 연결시켜 최적화를 구성하려고 하였다. 또한 교배율과 변이율의 적용을 위해 교배율 0.6~0.9, 변이율 0.01~0.3 사이에서 반복실험한 결과 클래스에 따라 교배율 0.8~0.85, 변이율 0.2~0.25사이에서 200세대안에 수렴하는 것을 알 수 있었다. 본 연구의 목적이 유전자 알고리즘에 의한 화소조작과 처리 결과의 추이에 중점을 두고 있으므로 200세대까지 반복연산 후 그 결과를 채택하였고, 결과적으로 유전자 알고리즘을 통한 트레이닝 화소 조합 작업이 화소조합의 평균, 분산을 모집단분포에 근사시키고 있다는 것을 알 수 있었다.

$$object\ function = r + \frac{1}{2}\ln\sqrt{\sum_{i=1}^{m} cov_i} \qquad (6)$$

r : Q-Q plot의 상관계수, n : 공분산값에 대한 가중치, m : 영상 밴드수

5. 분석 및 평가

연구 대상지는 충청남도 공주, 청양지역이며, 실험영상으로는 1995년 5월에 획득한 SPOT XS(row 670×column 540)영상과 1994년 5월에 획득한 LANDSAT TM(row 471×column 380)영상으로 하였다. 평가를 위한 현지 자료는 항공사진 데이터로부터 클래스별로 폴리곤을 구축시켰으며, ARC/INFO에서 클래스 조합별로 grid를 생성하였다. 연구진행의 효율성 측면에서 MLC 분류 프로그램, 트레이닝 정규화 프로그램, reference grid 생성 프로그램 등을 작성하여, 기존 상용 소프트웨어 사용을 최대한 배제하였다.

1) 영상 적용

본 연구는 트레이닝 데이터의 대표성을 정규화로 구현하고, 이를 공정의 기준으로 채택하여 영상처리공정의 객관화와 화소분포예측을 시도하는 것이다. 따라서 영상분류에 적용할 클래스 선정은 미국지질조사국(U.S.G.S : United States Geological Survey)의 토지이용/토지피복 분류체계(Landuse/Landcover classification)를 기본으로 하여 이중에서 실험대상 영상 내 화소중 정규분포가정에 근거한 샘플화소추출이 가능하고, 현지자료(reference data)로부터 오차행렬을 구성할 수 있는 클래스를 검토한 결과 삼림, 논, 밭, 과수원, 도시, 건천, 물 등 7개 클래스를 적용하였다.

기하보정은 분류 정확도 평가의 reference자료인 수치지도를 기준으로 하였다. 변환식으로는 선형 다항식변환(linear polynomial transformation)식을 적용하였으며, 원래의 화소값을 최대한 유지시키기 위해 최근린내삽법(nearest neighborhood interpolation)으로 화소재배열을 한 결과 SPOT XS, LANDSAT TM영상에 대하여 RMSE가 각각 0.570화소, 0.456화소였다.

2) 트레이닝 정규화 결과분석

실험영상에 대하여 8개 클래스로 트레이닝하였고(산의 경우, 밝은 부분과 어두운 부분을 각각의 클래스로 트레이닝하여 분류결과 작성시 이들을 통합하였다.), 트레이닝으로 선정된 화소들 중 다변량 정규성에 악영향을 미치는 화소를 제거하는 방법으로 객관성을 유지하기 위해 앞에서 서술한 바와 같이 정규화 알고리즘을 도입하였다. 본 실험결과의 의의는 어느 정도 숙련되지 않은 영상분류 작업자라도 충분한 수의 샘플화소를 취한다면 트레이닝 매개변수 설정 범위가 일정수준안으로 수렴할 것으로 기대하는 것이다.

(1) SPOT XS영상의 트레이닝에 대한 적용 결과 분석

SPOT XS영상의 각 클래스별로 트레이닝된 화소들에 대하여 정규화 알고리즘을 적용하였고, 그 결과 모든 클래스가 유의수준 5%를 만족하거나 정규성이 향상되었음을 Fig. 1에서 알 수 있다. 단,

Table 1. Comparison of Q-Q plot correlation coefficient between Before and After training normalization in case of SPOT XS image

sampling / class	Before training normalization		After training normalization	
	significance level 5%	Q-Q plot correlation coefficient	significance level 5%	Q-Q plot correlation coefficient
forest	0.9914	0.9894	0.9818	0.9958
watr	0.9813	0.9826	0.9628	0.9847
field	0.9900	0.9528	0.9801	0.9856
paddy	0.9924	0.9721	0.9858	0.9847
orchard	0.9735	0.9796	0.9460	0.9926
dry-brook	0.9869	0.9965	0.9749	0.9901
city	0.9852	0.9785	0.9700	0.9865
forest(shadow)	0.9899	0.9974	0.9768	0.9969

본 논문은 샘플화소의 통계치가 정규성에 가깝거나 먼 정도에 대한 분류 클래스 분리도 측정 및 분류정확도 향상에 대한 연구이므로 유의 수준값에 큰 의미를 부여하고 있지 않고, 여기서 적용한 수치는 일반적 통계적 검정에 많이 사용되는 경우에 의한다.

정규화 이후 밭을 제외한 클래스들에 대하여 밴드별 평균과 분산 대부분이 초기 트레이닝에 비해 모집단에 더 가까워졌다는 것을 Table 2에서 알 수 있다. 여기서, 모집단의 기준데이터로 사용한 자료는 항공사진 도화자료의 경계를 영상과 접합시킨 뒤 해당 클래스별 화소값을 추출하여 구한 것이다. 또한 트레이닝으로 추출된 화소의 최대·최소값이 크게 변하고 있지 않다. 이것은 정규화

Table 2. Comparison of mean · variance among Reference, Before and After training normalization in case of SPOT XS image

class	statistic		band 1	band 2	band 3
water	mean	reference	25.202	41.939	33.694
		before training normalization	22.031	41.246	31.861
		after training normalization	22.857	41.357	31.929
	variance	reference	96.432	26.133	53.275
		before training normalization	34.593	3.876	8.621
		after training normalization	50.497	5.571	13.254
field	mean	reference	40.381	41.801	36.138
		before training normalization	44.067	42.425	36.985
		after training normalization	44.850	42.800	37.300
	variance	reference	30.019	11.000	20.485
		before training normalization	80.650	5.254	12.526
		after training normalization	127.045	7.044	19.909
paddy	mean	reference	40.004	43.630	38.291
		before training normalization	40.704	45.816	40.447
		after training normalization	40.528	45.832	40.337
	variance	reference	19.135	7.374	12.906
		before training normalization	14.475	7.039	6.567
		after training normalization	14.661	9.483	7.931
orchard	mean	reference	38.456	39.766	33.573
		before training normalization	40.714	39.786	33.976
		after training normalization	40.556	39.778	34.167
	variance	reference	26.177	5.493	9.302
		before training normalization	4.551	1.343	2.609
		after training normalization	5.909	1.948	3.324
dry-brook	mean	reference	46.763	54.264	51.853
		before training normalization	50.253	58.072	56.867
		after training normalization	49.844	57.778	56.378
	variance	reference	154.132	65.774	125.217
		before training normalization	44.386	21.458	41.994
		after training normalization	60.998	25.404	53.377
city	mean	reference	38.731	45.842	39.889
		before training normalization	38.235	49.153	42.659
		after training normalization	38.243	48.838	42.324
	variance	reference	16.868	15.897	19.218
		before training normalization	3.944	5.084	4.751
		after training normalization	5.911	7.084	7.170

가 트레이닝된 전체 화소값 범위를 크게 왜곡시키지 않고 있다는 것이며, 그 평균과 분산값이 국소 최적해(Local Optima)로 수렴되지 않았다는 것을 보여준다.

그런데 Table 2에서 볼 수 있듯이 밭의 경우는 모집단의 평균과 분산값에 대하여 반대방향으로 수렴하고 있다는 것을 알 수 있다. 이것은 밴드 1의 트레이닝 최대값이 모집단의 최대값과 같은 것으로 미루어 볼 때, out-lier가 포함되어 있음을 알 수 있다.

Fig. 3은 밭의 모집단 화소분포를 표시한 것인데, 밴드 2와 3에서 시각적으로 약간 왜곡(skew)되어 있고, 특히 밴드 3은 multi-modal한 형태를 보이고 있다.

그러나 이러한 경우에 작업자는 모집단 화소분포를 알 수 없으므로 자신의 트레이닝이 밴드 1에서와 같이 한쪽으로 치우쳐 있거나 모집단 분포에 왜도가 있는지에 대한 정보가 주어지지 않는다. 이러한 경우의 문제를 해결하기 위해서 Q-Q plot에서 화소값 중심에 가까운 화소 집단(본 연구에서는 총화소수의 50%를 적용)을 고정화소로 하여 정규화조작을 실시한다. 이것은 트레이닝이 대표성이 있고 모집단 분포가 정규성에 가깝다면, 고정화소가 중심으로부터 가깝게 위치하고 있으므로 평균이 모집단쪽으로 수렴하고 그 분산이 증가하지만, 트레이닝이 밴드 1에서와 같이 한쪽으로 치우쳐 있거나 모집단 분포가 정규성에 가깝지 않으면 그 수렴방향이 고정화소가 없을 때와 반대로 수렴하게 된다는 의미이다.

Table 3은 고정화소를 취했을 때의 결과값을 정리해 놓은 것인데, 밭의 경우를 제외하고는 모두 모집단으로 수렴하고 있는 것을 볼 수 있다. 이때 고정화소 영향 때문에 전체 화소를 모두 고려하는 경우에 비해 그 수렴폭이 적은 것도 알 수 있다.

결과적으로 트레이닝시 충분한 화소를 취하고, 영상상태가 양호할수록 정규성을 기준으로 한 정규화 작업결과 그 수렴방향이 상기실험과 같이 일정하다고 할 때, 객관화된 트레이닝 공정으로 도입 가능성을 입증하고 있다. 또한 그 수렴방향에 의해 모집단의 분포 예측도 가능하다는 것을 알 수 있었다.

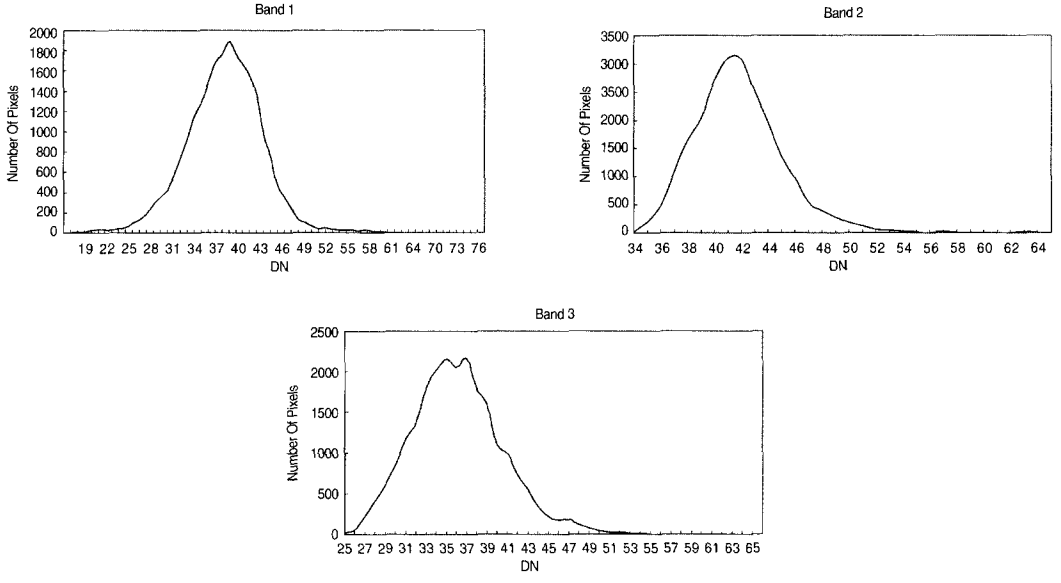

Fig. 3. DN distribution of field class in SPOT XS imagery

(2) LANDSAT TM영상의 트레이닝에 대한 적용 결과분석

LANDSAT TM영상에 대하여 정규화 알고리즘을 적용한 결과를 Table 4~6에 작성하였다. Table 5에서 몇몇 밴드의 평균과 분산값이 모집단으로 수렴하지 않는 것은 정규화 알고리즘이 다변량에서의 정규성을 평가함수로 설정하고 있기 때문이며, 다차원에서의 모집단 평균 및 분산으로 수렴경향을 판단하여야 한다. 그럼에도 불구하고 도시와 밭 클래스의 경우에는 평균값은 모집단에 대하여 수렴하고 있으나 분산이 축소되고 있고, 특히 도시의 경우 분산과 최소·최대값 범위가 급

Table 3. Comparison of mean · variance among Reference, Before and After training normalization in case of SPOT XS image considering fixed sample pixels

class	statistic		band 1	band 2	band 3
water	mean	reference	25.202	41.939	33.694
		before training normalization	22.031	41.246	31.861
		after training normalization	22.143	41.163	31.898
	variance	reference	96.432	26.133	53.275
		before training normalization	34.593	3.876	8.621
		after training normalization	37.458	4.056	9.552
field	mean	reference	40.381	41.801	36.138
		before training normalization	44.067	42.425	36.985
		after training normalization	44.119	42.458	36.975
	variance	reference	30.019	11.000	20.485
		before training normalization	80.650	5.254	12.526
		after training normalization	84.123	4.968	11.957
paddy	mean	reference	40.004	43.630	38.291
		before training normalization	40.704	45.816	40.447
		after training normalization	40.717	45.761	40.389
	variance	reference	19.135	7.374	12.906
		before training normalization	14.475	7.039	6.567
		after training normalization	15.062	6.898	6.865
orchard	mean	reference	38.456	39.766	33.573
		before training normalization	40.714	39.786	33.976
		after training normalization	40.379	39.793	33.793
	variance	reference	26.177	5.493	9.302
		before training normalization	4.551	1.343	2.609
		after training normalization	5.030	1.384	2.813
dry-brook	mean	reference	46.763	54.264	51.853
		before training normalization	50.253	58.072	56.867
		after training normalization	49.739	57.652	56.261
	variance	reference	154.132	65.774	125.217
		before training normalization	44.386	21.458	41.994
		after training normalization	45.402	22.701	43.960
city	mean	reference	38.731	45.842	39.889
		before training normalization	38.235	49.153	42.659
		after training normalization	38.083	49.067	42.583
	variance	reference	16.868	15.897	19.218
		before training normalization	3.944	5.084	4.751
		after training normalization	4.247	5.283	5.264

Table 4. Comparison of Q-Q plot correlation coefficient between Before and After training normalization in case of LANDSAT TM

sampling / class	Before training normalization		After training normalization	
	significance level 5%	Q-Q plot correlation coefficient	significance level 5%	Q-Q plot correlation coefficient
forest	0.999	0.9993	0.9936	0.9939
watr	0.9772	0.9275	0.9413	0.9805
field	0.9920	0.9936	0.9913	0.9988
paddy	0.9823	0.9653	0.9603	0.9922
orchard	0.9664	0.9699	0.9437	0.9978
dry-brook	0.9740	0.9893	0.9652	0.9981
city	0.9884	0.8359	0.9863	0.9932
forest(shadow)	0.9932	0.9948	0.9867	0.9943

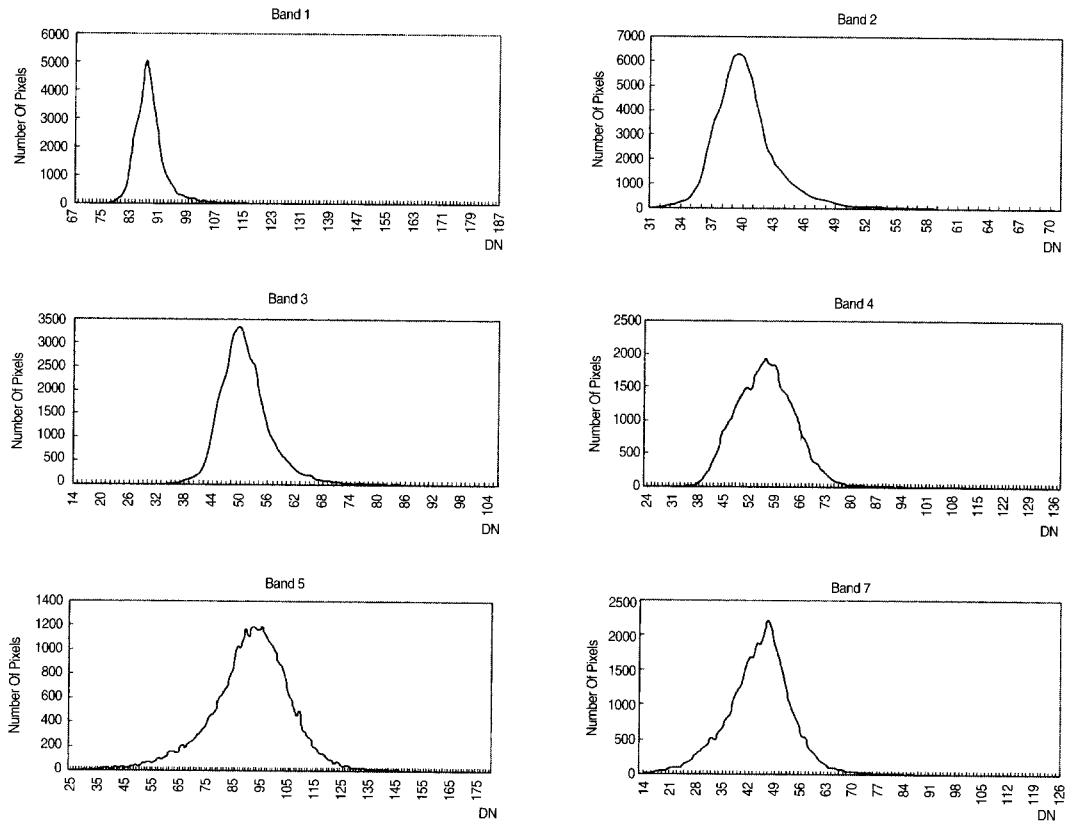

Fig. 4. DN distribution of paddy class in LANDSAT TM imagery

격히 축소되는 것을 볼 수 있다.

이들을 분석하기 위해 비교적 수렴이 되고 있다고 판단되는 논 클래스와 다른 클래스에 비해 현저히 다른 결과를 작성하고 있는 도시 클래스의 모집단 화소를 추출하여 각각에 대한 밴드별 화소값 분포를 Fig. 4, 5에 나타내었다. Fig. 4, 5에서 논의 경우 밴드 5를 제외한 대부분의 밴드에서 화소값의 분포가 정규성에 가깝게 분포되어 있는 반면, 도시의 경우 분포자체가 multi-modal하게 형성되어 있음을 알 수 보여주고 있다. 이것은 정규화

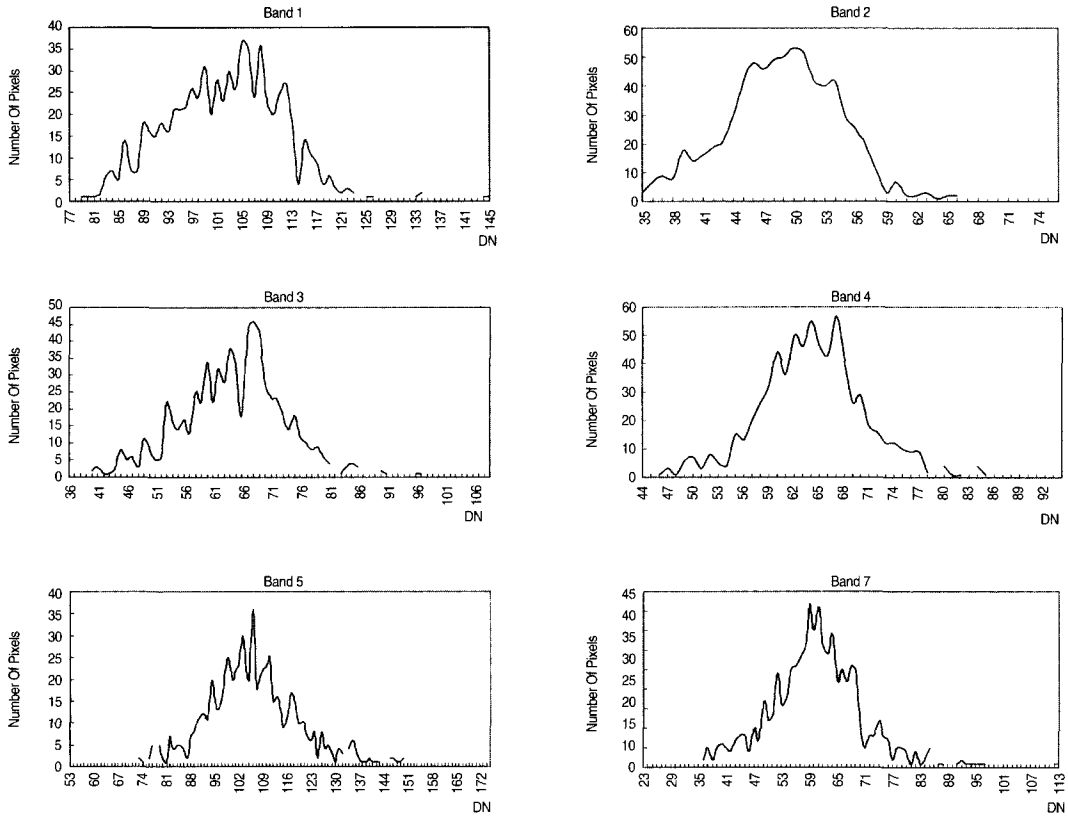

Fig. 5. DN distribution of city class in LANDSAT TM imagery

알고리즘의 기본 가정, 즉 모집단의 화소분포가 정규성이라는 것에 위배되고 있고, 모집단 화소들에 대한 Q-Q plot 상관계수도 0.805로 유의수준 1%에 훨씬 못 미치고 있는 것으로도 입증된다.

또한, Table 6에 의하면 트레이닝 화소의 최소·최대값 범위가 대부분의 밴드에서 지나치게 한쪽으로 편의되어 있음을 알 수 있다. 결국 도시 클래스의 트레이닝 화소들에 대한 정규화 결과는 국소영역에 대하여 수렴되며, 이것이 트레이닝 정규화 이후 화소조합의 최대·최소값의 범위가 정규화이전보다 급격히 축소되는 이유이다.

SPOT XS, LANDSAT TM 영상에 대한 정규화 알고리즘을 적용을 통해서, 트레이닝에 대한 정규화 결과가 (a) 정상적으로 분산이 커진다고 하더라도 고정화소를 지정하고 정규화를 시켜 그 결과가 초기 결과와 상반된 결과 즉, 평균과 분산의 수렴방향이 반대로 된다면 트레이닝이 모집단에 대하여 편의되어 있거나 모집단 자체가 왜곡되어 있을 가능성이 높고 (b) 분산이 줄어들고, 최소·최대값 범위도 현저히 축소되면 모집단의 분포가 multi-modal형태를 포함하여 정규성이 적은 분포형태를 가질 가능성이 매우 높으며, (c) 분산이 줄어들고, 최소·최대값은 그대로 있으면 (a)와 (b)의 결과를 모두 예상할 수 있다.

6. 결 론

일반적으로 인공위성 영상을 통계적 기법으로 분류를 하는 경우, 영상 화소값은 여러 종류의 원인이 결합되어 다각적으로 영향을 미친다. 이러한 통계적 기법들로 원인별 모집단 특성을 규명할 경

Table 5. Comparison of mean · variance among Reference, Before and After training normalization in case of LANDSAT TM image

class	Statistic		band 1	band 2	band 3	band 4	band 5	band 7
water	mean	reference	83.513	37.044	40.412	27.013	31.954	17.052
		before	82.627	36.314	38.706	24.941	26.549	13.843
		after	83.000	36.625	38.875	25.938	28.563	14.875
	variance	reference	12.475	7.941	26.071	47.900	175.006	59.877
		before	14.278	14.860	34.132	13.856	25.573	7.575
		after	23.600	22.117	49.317	33.129	72.529	18.917
field	mean	reference	88.842	41.378	53.990	63.059	103.314	52.386
		before	91.633	43.953	59.503	66.166	114.213	59.485
		after	91.407	43.733	59.007	65.773	113.527	59.053
	variance	reference	34.246	20.976	69.806	70.997	286.591	109.768
		before	20.174	13.010	42.799	42.151	299.276	111.287
		after	17.760	11.150	37.121	41.009	291.983	106.548
paddy	mean	reference	88.093	40.363	51.194	57.163	92.002	45.914
		before	86.174	39.435	50.667	52.014	97.594	50.739
		after	86.654	40.231	53.000	54.885	101.192	51.500
	variance	reference	23.464	11.451	35.904	70.829	224.370	78.252
		before	10.881	7.191	52.431	59.897	216.803	32.725
		after	12.395	11.225	87.760	98.586	393.362	55.220
orchard	mean	reference	87.922	40.813	53.684	62.604	106.165	53.064
		before	87.688	40.094	51.750	63.469	105.875	52.844
		after	87.471	39.882	51.588	62.177	104.765	52.882
	variance	reference	18.148	9.967	36.373	63.362	206.066	61.717
		before	6.480	3.184	11.290	34.967	71.790	25.620
		after	6.890	2.985	9.757	23.404	59.066	26.360
dry-brook	mean	reference	114.995	61.260	87.280	80.954	150.397	91.839
		before	118.302	63.581	93.209	85.744	158.674	96.605
		after	117.367	63.000	92.267	84.767	157.033	95.500
	variance	reference	110.586	68.807	208.052	217.356	1346.156	616.827
		before	75.597	32.440	84.979	96.862	721.749	392.197
		after	85.137	38.000	97.926	115.771	850.516	454.535
city	mean	reference	102.388	49.051	64.637	64.383	106.383	60.542
		before	110.088	53.053	70.851	66.474	107.412	65.009
		after	108.936	52.505	69.817	65.065	104.914	63.226
	variance	reference	93.606	34.716	88.059	43.480	221.950	118.788
		before	49.444	15.696	37.544	26.429	135.377	70.593
		after	20.713	7.035	12.173	8.452	47.732	21.959

※ Here, before : before training normalization
after : after training normalization

우 다변량 통계분석기법이 주로 활용된다.

본 연구에서는 다변량 통계이론을 근간으로 한 트레이닝 화소 정규화의 객관적 공정으로의 도입 가능성을 연구하여 트레이닝 화소에 대한 정규화 알고리즘을 개발하였다. 실험 결과 유전자 알고리즘을 이용한 트레이닝 정규화 결과가 대부분의 클래스에 대하여 그 평균과 분산을 모집단에 근사시키고 있다는 것을 입증하였고, 고정화소를 고려한 트레이닝 정규화 알고리즘으로 해당 클래스의 모집단 분포를 예측할 수 있는 가능성을 제시하였다.

Table 6. Comparison of maximum · minimum DN value among reference, Before and After training normalization in case of LANDSAT TM image

class	Statistic		band 1	band 2	band 3	band 4	band 5	band 7
forest	Min.	reference	61	11	10	20	19	2
		before	74	31	32	36	52	21
		after	75	31	33	36	52	21
	Max.	reference	206	119	100	137	224	198
		before	89	40	54	66	130	65
		after	89	40	54	66	128	65
water	Min.	reference	74	29	28	18	13	7
		before	75	30	30	21	22	10
		after	76	30	30	21	22	10
	Max.	reference	129	180	108	99	189	117
		before	93	45	53	40	46	25
		after	93	45	53	40	46	25
field	Min.	reference	35	21	32	24	28	13
		before	85	38	49	50	66	33
		after	85	38	49	50	66	33
	Max.	reference	217	124	97	109	180	108
		before	107	58	85	78	159	96
		after	106	57	82	78	159	96
paddy	Min.	reference	67	31	14	24	25	14
		before	81	36	42	42	82	41
		after	83	36	43	43	82	41
	Max.	reference	204	71	106	138	181	126
		before	99	47	75	78	159	74
		after	98	47	75	78	159	74
orchard	Min.	reference	77	31	37	39	54	26
		before	82	37	45	56	95	45
		after	82	37	46	56	95	45
	Max.	reference	115	59	80	88	149	75
		before	93	44	58	78	124	62
		after	93	44	58	74	119	62
dry-brook	Min.	reference	82	31	38	18	17	7
		before	99	53	74	62	77	43
		after	99	53	74	62	77	43
	Max.	reference	132	180	111	111	202	126
		before	129	71	104	96	188	120
		after	129	71	104	96	188	119
city	Min.	reference	77	35	36	44	53	23
		before	95	45	57	57	84	51
		after	99	46	63	57	84	51
	Max.	reference	145	76	109	94	175	113
		before	144	76	107	94	175	113
		after	117	59	79	71	112	74

본 연구의 트레이닝 정규화결과로써 분류정확도를 향상시키는 구체적인 방법에서는 아직 한계를 갖는다. 이것은 본 연구의 초점이 위성영상처리의 객관적 공정확립에 맞추어져 있기 때문이며, 특히 위성영상분류에서 클래스 설정의 문제 및 분류결과의 사용자 주관적 요인을 배제시키려는데 의의가 있다.

참고문헌

新井康平, 1998, 簡略化 β分布による確率密度函數適用型 畵像分類, 寫眞測量とリモ-トセンシング, 社團法人日本寫眞測量學會, 37(1):40~44.

Basilevsky, A., 1994, *Statistical Factor Analysis and Related Methods-Theory and Applications*, Wiley Series in Probability and Mathematical Statistics, pp. 235~241.

Hixon, M., Schoez, D., and Fuhs, N., 1980, Evaluation of several schemes for Classification of remotely sensed data, *Photogrammetric Engineering and Remote Sensing*, 46(12):1547~1553.

Johnson, R.A., and Wichern, D.W., 1992, *Applied Multivariate Statistical Analysis*, third edition, Prentice Hall, pp. 126~133, pp. 152~153, pp. 158~164.

Kartikeyan, B., and Majumder, K.L. and Dasgupta, A.R., 1995, An Expert System for Landcover Classification, *IEEE Transactions on Geoscience and Remote Sensing*, 33(1):58~66.

Rencher, A.C., 1995, *Methods of Multivariate Analysis*, Wiley Series in Probability and Mathematical Statistics, pp. 112~113.

Swain, P.H. and Davis, S.M., 1978, *Remote Sensing: The Quantitative Approach*, McGrow-Hill, p. 142, pp. 159~164.

Van Deusen, P.C., 1996, Unbiased Estimates of Class Proportions from Thematic Maps, Photogrammetric Engineering and Remote Sensing, 62(4):409~412.