Journal of Korean Society of Remote Sensing, Vol. 1, No. 1, 1985

BASIC 언어를 사용한 Hill-Sliding 무감독 분류법 Algorithm 개발*

鄭 夢 炫, 崔 圭 弘
연세대학교 천문기상학과

and

朴 景 允
한국건설기술연구원

(1985년 5월 17일 받음)

Development of the Hill-Sliding Clustering Algorithm Using BASIC Language

Mong Hyun Chung, Kyu Hong Choi
Department of Astronomy & Meteorology, Yonsei University

and

J. Kyoungyoon, Park
Korea Institute of Construction Technology, Inchon, Korea

(Received May 17, 1985)

ABSTRACT

An algorithm for the Hill-Sliding Clustering (HSC) method was developed using the BASIC language for Apple II personal computer. It was designed for initialization of clusters from multivariate multimodal Gaussian data. Landsat multispectral imagery data of a Korean coastal area were used for its performance test. The test showed encouraging results.

요 약

Hill-Sliding Clustering이라는 다변량 자료의 무감독 분류 방법을 Apple Ⅱ personal computer의 BASIC 언어를 사용한 Algorithm으로 개발하였다. 이 Algorithm으로 다변량

* Yonse: University Observatory Contribution No. 27.

multimode를 갖는 정규 분포 자료에서 사전 지식없이 자료를 집단화하여 구분해 낼 수 있게되었다. 한국 연안 지역의 Landsat (지구 자원 탐사 위성)의 다중 Spectrum 영상 자료에 적용한 시험 결과, 매우 고무적 결론을 얻었다.

I. 서 론

원격 탐사는 물체와 물리적인 접촉없이 떨어진 거리에서 행하여진 측정으로 부터 물체에 대한 정보를 얻는 과학이다 (Landgrebe 1978). 원격탐사의 표적이 되는 자연은 끊임없는 변화의 요인을 갖고 복잡한 현상을 나타낸다. 이러한 자연으로부터 정보를 유출하는 것은 그리 쉬운 일은 아니며 더우기 이러한 정보를 분류 (classification)하는 것은 어려운 작업일 수 밖에 없다. 아뭏든 자연은 전자파를 발생시킴으로서 우리가 원격탐사를 수행할 수 있게 하나 이러한 전자파의 양은 지구 대기조건, 주변 물체의 유무, 태양과 지구 표면과의 사이각, 지구표면과 관측기기와의 사이각을 비롯하여 물체 자체의 변화로 인해 항상 일정하지 않고 상대적인 관측치를 갖는다. 이러한 복잡성이 원격탐사를 더욱 어렵게 만든다.

관측된 물체가 발생하는 복사량은 물체 고유의 특성을 갖고 있으므로 타물체와 구별될 수있다. 이 성질을 이용하여 우리는 관측된 자료를 군집화 (grouping) 하여 적절히 분류를 할 수 있다. 이러한 분류는 방대한 자료를 취급하므로 전산기를 이용하고 또한 계속적인 전산기의 발달은 더욱 좋은 결과를 얻고 있으나 적절한 수학적 표현의 결핍과 과중한 다중해로 인하여, 혹은 광대한 지식의 요구로 인해 쉽게 언급될 수 없는 문제이다 (Park *et al*. 1979, Park *et al*. 1985).

군집화에 의한 분류 기법은 크게 감독 분석 (supervised analysis)과 무감독 분석 (unsupervised analysis)의 두 가지로 나눠진다. 감독 분석은 분류할 당시 그 지역의 필요한 정보 (information)를 알고 있는 경우 이를 training sample로 이용하는 방법이다. 이 방법은 정확한 결과를 얻을 수 있으나 정보를 추출하는 작업에서 어려운 점이 있고, 또한 정보가 빈약한 지역에서는 이용할 수 없는 약점이 있다.

이에 반해 무감독 분석은 다변량 자료 (multivariate data)를 분석하는 데 있어서 전산기를 이용한 algorithm을 이용하여 정보가 빈약한 상태에서 분류하는 방법인데 감독 분석과는 역행하는 과정을 보인다. Maxwell (1977)은 training sample을 선택하여 감독 분석을 수행한다 하여도 목표물에 대한 대표적인 표현은 여러면에서 힘든 작업임을 표현했고 Nagy (1974)는 무감독 분석이 자료의 다중 (multimodal) 분포에 따르는 난점을 많이 경감시키는 이점을 갖고 있다고 하여, 감독 분석이 해결하기 어려운 여건에서 무감독 분석이 이용됨을 보인다.

Park 등 (1979)은 무감독 분석 Hill-Sliding clustering algorithm을 개발했다. 이는 다변량 자료의 Gauss 분포를 이용하여 분석하는 방법인데, 이 논문에서는 Hill-Sliding clustering 방법을 microcomputer에 이용할 수 있도록 BASIC 언어의 개발과 분석 작업을 보여 주려고 한다.

Ⅱ. Hill-Sliding Clustering 방법

Park 등(1979, 1985)은 다음과 같은 기본 원리를 제시하였다.

1. 한 집단의 탐사 자료는 unimodal 분포를 가정한다. 여기서 우리는 다변량 정규 분포의 확률 밀도 함수(probability density function)를 얻는다.
2. 원격탐사 자료의 등방성(isotropic) 정규 분포를 가정한다. 집단 i를 얻기 위해 자료의 mode를 집단의 중심이라 생각하고 이 중심에서부터 거리에 대한 단위 체적 당의 갯수에 대한 함수로부터 타 집단과의 경계치(valley)를 구한다. 우리는 이 거리의 제곱을 threshold value r_t^2이라 명명한다.
3. 중심으로 부터 경계치안에 있는 d차원으로 표시된 자료는 일단 하나의 집단으로 생각하여 집단화 함수(clustering function)에 적용하기 위해 집단의 사전 확률(apriori probability), 공분산 행렬(covariance matrix)과 평균값을 얻는다. 여기서 집단화 함수는 판별함수인 maximum likelihood function(Tou and Gonzalez 1974)에 기초한 것으로 자료의 집단 귀속 여부를 결정해 준다.

끝으로 집단에 속하지 않은 자료는 Mahalanobis 거리 $D_i(\vec{x})$를 이용하여 분류하는데 이는 확률적인 의미를 갖는 거리로서 다음과 같이 표시된다.

$$D_i(\vec{x}) \equiv (\vec{x} - \vec{\mu_i})^T C_i^{-1} (\vec{x} - \vec{\mu_i}) \quad \cdots\cdots (1)$$

여기서 C_i는 집단 i의 d×d 공분산 행렬, $\vec{x}$는 자료 vector, $\vec{\mu_i}$는 집단 i의 평균 vector이고 d는 특징 공간(feature space)을 나타낸다.

이 algorithm에 사용된 다변량 정규 분포의 확률 밀도 함수(Duda and Hart 1973)는

$$p_i(\vec{x}) = \frac{P_i}{(2\pi)^{d/2}|C_i|^{1/2}} exp\left(-\frac{D_i}{2}\right) \quad \cdots\cdots (2)$$

이며 여기서 P_i는 집단 i에 있어서의 사전 확률(apriori probability)이다.

집단화 함수(clustering function)를 다음과 같이 정의하면

$$G_i(\vec{x}) = ln\frac{p(\vec{x})}{p_i(\vec{x})} \quad \cdots\cdots (3)$$

이고 이는 곧

$$G_i(\vec{x}) = \frac{D_i(\vec{x})}{2} - ln P_i + \frac{1}{2} ln|C_i| + \frac{d}{2} ln(2\pi) + ln\, p(\vec{x}) \quad \cdots\cdots (4)$$

로 표현되는데 $p(\vec{x})$는 $\vec{x}$에서의 확률 밀도로서 자료에서 계산된 값이다. 이론적으로 집단 i의 영역안에서는 $G_i(\vec{x}) < ln\ 2$의 조건을 만족한다.

특징 공간에서 구분이 가능한 수 개의 집단이 섞여있는 자료군에서의 확률 밀도 $p(\vec{x})$는 다음과 같이 계산된다.

$$p(\vec{x}) = \sum_{\text{all } i} p_i(\vec{x}) \quad \cdots\cdots (5)$$

그리고 집단 i의 영역과 집단 j의 영역의 경계지역에서는

$$0 < G_i(\vec{x}) = G_j(\vec{x}) \lesssim ln\ 2 \quad \cdots\cdots (6)$$

가 된다. 이때

$$G_i\ (\vec{x} \in \text{집단 } i) = \min_{\text{all } k} \{ G_k(\vec{x}) \} \quad \cdots\cdots (7)$$

가 성립하므로 이를 만족하는 집단 i의 pattern 벡터 $\vec{x}$를 찾는 작업이 집단화가 된다.

Ⅲ. Hill-Sliding Clustering Algorithm

우리는 Park 등 (1979, 1985)의 Hill-Sliding 방법을 이용하여 microcomputer에 이용할 수 있는 computer program을 개발했다. 사용한 기종은 CPU 6502의 Apple Ⅱ (64K) computer로서 BASIC 언어로 작성된 program이다. 약 400개의 자료를 2차원 특징 공간에서 분류하는데 15분 정도의 시간을 필요로 한다. 기계어(machine language)로 program을 작성하면 소요시간이 훨씬 단축되리라 생각되는데 이것은 다음 기회에 시도해 보려고 한다.

그림 1에 우리가 개발한 program의 흐름도를 실었다. 2차원 특징 공간의 자료가 입력되면 우선 가장 큰 확률 밀도 값을 sorting program에 의해 찾는다. 다음은 이 최대값을 집단의 중심으로 하여 다른 집단들과의 경계치를 계산한다. 그후 집단중심으로부터 이 경계치내의 자료를 한 종류의 특성을 가진 집단이라 가정하고 집단화 함수를 이용하기 위해 이에 필요한 사전 확률, 공분산 행렬, 중심 밀도를 계산하고 이를 집단화 함수에 대입한다. 여기서 조건식을 이용하여 집단화가 확률적으로 가능한 구역을 새로 정한다. 이렇게 하여 하나의 집단을 형성한다. 이제 program 사용자가 제시한 최대 집단수가 넘지 않았거나 집단화 되지 않은 자료가 있을 경우에는 이 program은 다시 처음 단계로 거슬러 올라가 남은 자료들을 sorting하게 되고 계속 다음 단계로 반복 진행한다. 결국 최대 집단수가 넘게 되면 집단화 되지 않은 남은 자료들은 Mahalanobis 거리를 이용해 각 집단들에 귀속시키게 되어 자료의 분류가 완성된다. 집단화한 결과는

자료가 위치한 원래의 지역에 symbol 이나 색으로 표시하여 그 지역의 집단 분포를 알 수 있게 한다.

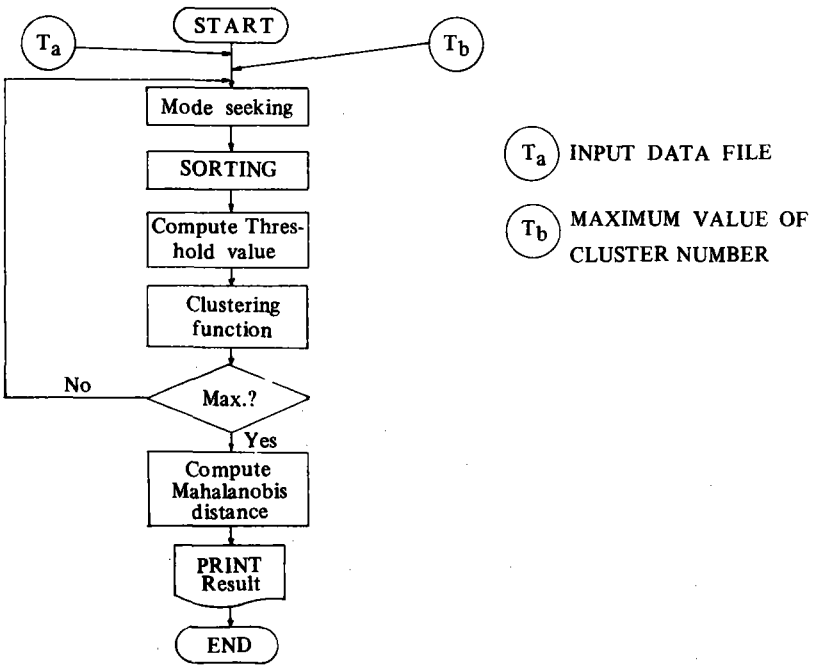

그림 1. Microcomputer 를 이용한 algorithm의 흐름도

Ⅳ. 원격 탐사 자료에의 응용

이 장에서는 우리가 개발한 microcomputer용 Hill-Sliding Clustering algorithm을 원격탐사 자료에 응용하였다. 원격탐사자료는 일반적으로 Landsat (Land satellite)의 MSS (Mutispectral Scanner) 자료가 쓰인다. 여기서 Landsat 은 미국 NASA의 자원탐사위성이고 MSS 자료는 한 지역을 여러개 (Landsat 1,2 호는 4 개 3 호는 5 개) 의 분광대로 동시에 읽은 자료이다. 이러한 MSS 자료의 분해능은 기본단위인 한 pixel (picture element) 이 지상 $79m \times 79\,m$ 를 나타낸다. Landsat 안에서의 검출기의 출력은 디지탈화 되고 변조계에서 코오드화 되어서 지상 수신소로 원격 수신된다. 수신소를 통한 신호는 증폭되고, 변조를 풀어서 디지탈 형식으로된 전산기용 테이프 (CCT) 를 만든다. 이렇게 만들어진 자료는 전산기를 이용하여 분석하게 된다.

그림 2 는 해양연구소 (유홍룡 1984) 로부터 얻은 자료를 2 차원 특징 공간분포로 만든 것이다. 이 자료는 1979 년 10 월 4 일 한국 서해안 지역을 주사한 Landsat 2 호의 MSS 자료로 line 2001 부터 2400, column 2401 부터 2880 의 CCT 자료이다. 이는 microcomputer 로

분류하기에는 너무 많은 양이므로 sampling을 하지 않은 일부분을 추출하였다. 분광대는 5 ($0.6 \sim 0.7 \mu m$) 와 7 ($0.8 \sim 1.1 \mu m$) 을 이용하였다. 그림 2 는 육안으로 보기에는 쉽게 집단을 나누기 어려우나 세 집단으로 나누어 볼 수 있을 것이다.

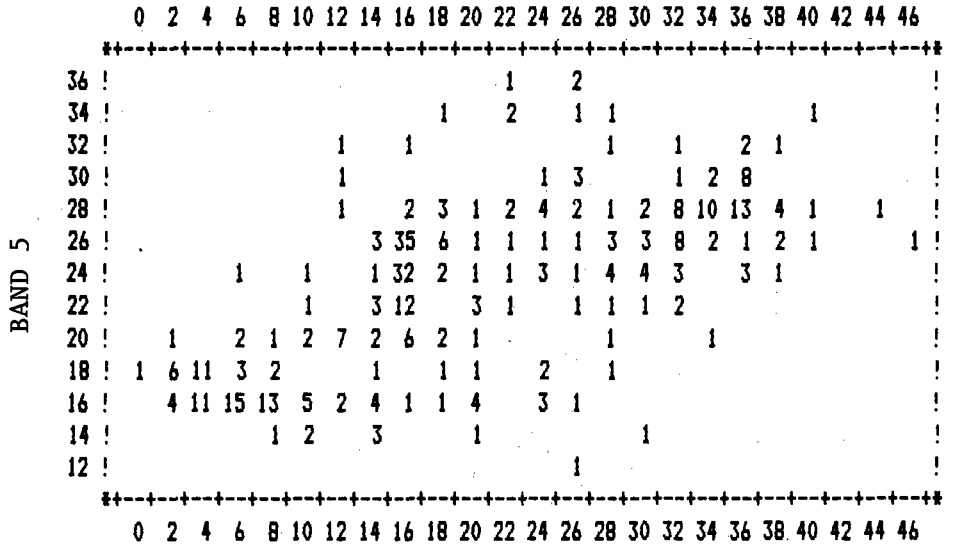

그림 2. 374개의 LANDSAT 자료 분포 (유홍룡 1984).
line 2212 - 2222, pixel 2713-2746의 지역으로 1979년 10월 4일의 LANDSAT 자료이다.

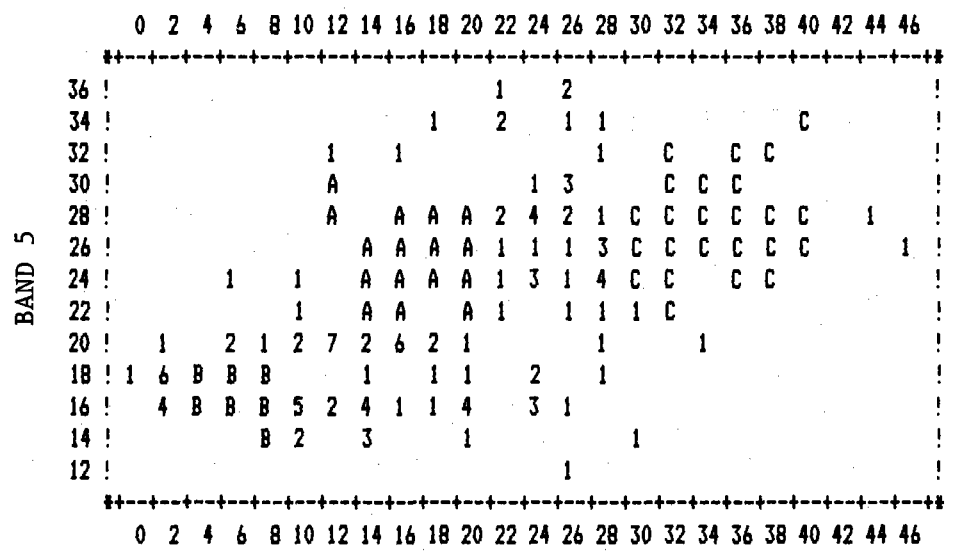

그림 3. 집단화 함수를 이용한 집단 분석.

그림 3은 HSC algorithm에 의해 세 개의 집단으로 나눈 최초의 분석 결과를 보여준다. 이는 집단화 함수를 이용한 것으로서 확률적 의미를 최대한으로 이용한 분석 방법이다. 이어 집단화된 주위에 분류되지 않은 많은 자료를 보게 된다. 이 자료들 역시 우리가 추출한 지역의 확률밀도 만큼의 pixel 수, 즉 해당 pixel에 해당하는 지역을 표시하므로 어느 집단에 귀속시키느냐의 문제는 매우 중요하다. 잘못 처리되면 실제와는 다른 성질의 지역으로 오판되기 때문이다. 이러한 오판율을 최소로 하기 위해 확률적인 거리를 나타내는 Mahalanobis 거리를 이용한다. 그림 4는 Mahalanobis 거리를 이용한 최종 분석을 보여주고 있다. 이는 Mahalanobis 거리가 오판될 확률을 최소로 한다는 점에서 믿을 만한 분석 결과가 된다.

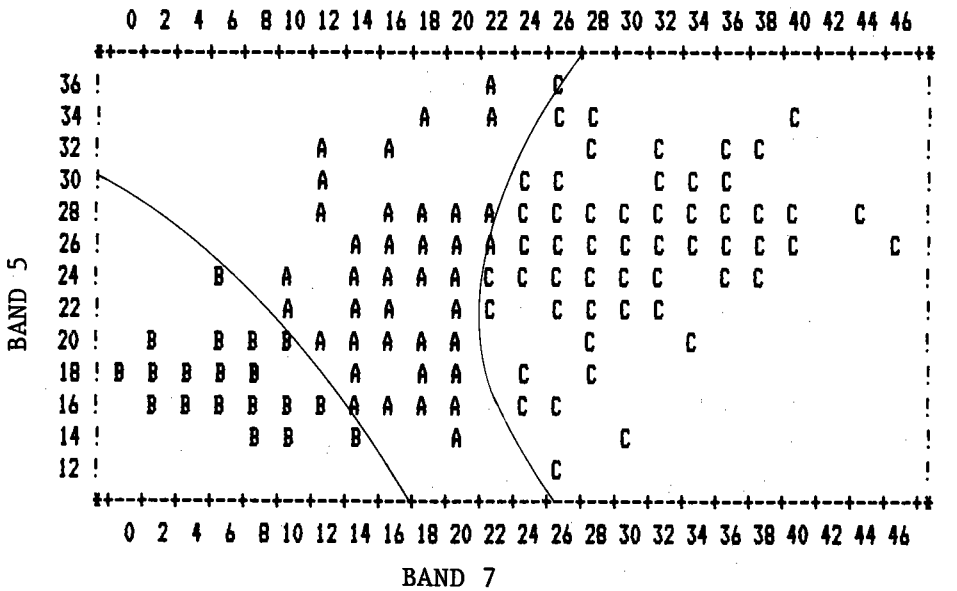

그림 4. Mahalanobis 거리를 이용한 최종 집단화 (clustering).

이 분석 결과를 원래의 위치에 symbol로 표시하였다. 그림 5가 그 결과이다. "*"는 갯벌을 나타내고 "="와 "+"는 각각 해수와 육지를 나타낸다. 세 부류의 지역이 나타남을 선명히 볼 수 있다. 또 갯벌과 해수사이에 육지 기호 "+"가 간혹 보이는데 이는 조사 결과 방조제임이 밝혀졌다.

결국 우리의 program은 비록 Apple Ⅱ computer에 의존하였으므로 약 500개의 자료 이상은 분석하기 어려운 단점이 있으나 협소한 지역을 조사할 목적이면 충분한 활용성이 있고 자료만 입력되면 무감독 분석 분류를 할 수 있음을 입증했다.

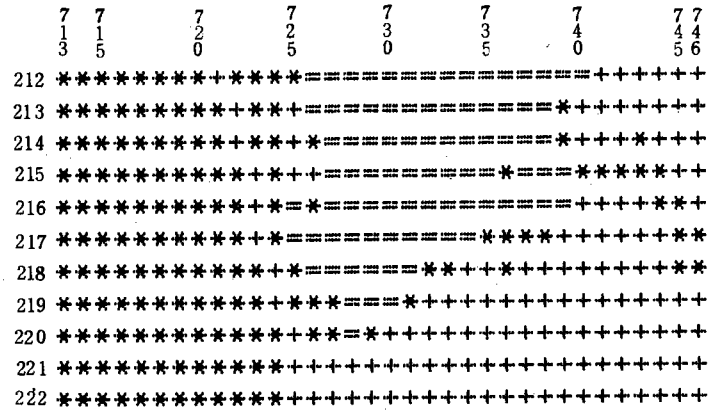

그림 5. 아산만 일부 지역의 LANDSTA 자료 374개
1979년 10월 4일 CCT 자료로 2212-2222 line과 2713-2746 Column의 지역을 집단화 (clustering)한 결과이다.

V. 결 론

원격 탐사에 있어서 분류(classification) 기법의 하나인 Hill-Sliding Clustering method를 microcomputer용 BASIC 언어로 code화 하는 algorithm을 개발하였다. 이를 이용하여 training sample이 없는 곳의 자료를 분석하여 분류함으로서 자료안의 군집화 현상을 찾아낼 수 있었다. 이 algorithm은 원격 탐사 자료 이외의 어떠한 digital 자료에도 적용할 수가 있는데 성격상 평균치가 다른 정규 분포를 갖는 여러개의 집단이 혼합되어 있는 자료를 분류하는 데 적합하다. 원격탐사 자료의 이용에 있어서는 실제 가보지 않은 지역을 조사할 수 있으므로 경제적 이익을 볼 수 있어 댐을 건설하는 경우 넓은 지역을 한눈에 보고 조사할 수 있게 하고 지도 작성에도 유용하게 쓸 수 있다. 또한 농산물 수확량을 추정하는데에도, 광물을 조사하는 자원 탐사들에도 유익하고 기상예보에도 기여하는 등 광범위한 실용성을 내포하고 있다.

원격 탐사 기법을 이용한 분류(classification)는 주로 대형 전산기에 의존하던 것에서 경제성을 고려한 microcomputer를 이용하는 것이 현재의 추세이고 보면 우리의 연구는 실용성 있는 결과라 할 수 있다. 물론 용량면에서는 대형 전산기에 비교가 안될 정도로 떨어져 제한이 많으나 microcomputer 정도의 용량으로 가능한 분석 작업은 허다하므로 이러한 연구 및 개발은 앞으로도 계속 추진되어야 할 것이다.

참 고 문 헌

유홍룡. 1984, 개인 서신.

Duda, R. O., and Hart, P. E. 1976, *Pattern Classification and Scene Analysis* (N. Y.: John Wiley and Sons), pp. 189-256.

Landgrebe, D. A. 1978, *Remote Sensing: The Quantitative Approach,* ed. by P. H. Swain, and S. M. Davis (N. Y.: McGraw-Hill), pp. 1-21.

Maxwell, E. L. 1977, *Journal of Range Management,* **29**, 66.

Nagy, G. 1968, *Proceedings of the IEEE,* **56**, 836.

Park, J. K., Chen, Y. H., and Simons, O. B. 1979, Ph. D. Thesis of Colorado Univ.

Park, J,K., Chen, Y.H., Simons, O.B., and Miller, L.D. 1985, *Journal of Korean Society of Remote Sensing,* 1, No. 1, pp. 3-25.

Tau, J. T., and Gonzalez, R. C. 1974, *Pattern Recognition Principles* (London: Addison-Wesley) pp. 362-395.