韓國地形空間情報學會論文集 第4卷 第2號 1996年 12月 pp. 181-193

정준상관분석을 이용한 원격탐사 수치화상 분류기법의 개발 : 무감독분류기법과 정준상관분석의 통합 알고리즘

Development of Classification Method for the Remote Sensing Digital Image Using Canonical Correlation Analysis

김 용 일* 김 동 현** 박 민 호***
Kim, Yong-Il Kim, Dong-Hyun Park, Min-Ho

要 旨

본 연구는 원격탐사의 수치화상분류에 적용된 바 없는 정준상관분석(Canonical Correlation Analysis)기법을 무감독분류한 위성화상데이터에 적용하여 토지피복분류하는 새로운 방법을 개발하는 것을 목적으로 한다. 개발된 분류기법은 기존의 분류기법인 최대우도분류기법에 비해 분류기준용 표본데이터 선정이 용이함을 알 수 있었다. 즉, 정준상관분석에 의한 분류결과는 분류기준용 표본데이터의 선정위치에 거의 영향을 받지 않는다. 또한 무감독분류 후 정준상관분석에 의해 결정된 각 군집의 토지피복은 최대우도분류를 위한 사전정보로 활용가능하다. 동일한 분류기준용 표본데이터 사용시, 무감독분류 후 정준상관분석에 의한 분류가 최대우도분류보다 분류정확도가 우수하였다. 이상과 같은 결과로 판단해 볼 때 본 연구에서 시도된 분류기법은 원격탐사의 분류기법 분야에서 실용화 될 수 있으며, 나아가서는 GIS 데이터베이스 구축에 중요한 역할을 할 수 있을 것이다.

ABSTRACT

A new technique for land cover classification which applies digital image pre-classified by unsupervised classification technique, clustering, to Canonical Correlation Analysis(CCA) was proposed in this paper. Compared with maximum likelihood classification, the proposed technique had a good flexibility in selecting training areas. This implies that any selected position of training areas has few effects on classification results. Land cover of each cluster designated by CCA after clustering is able to be used as prior information for maximum likelihoodlassification. In case that the same training areas are used, accuracy of classification using Canonical Correlation Analysis after cluster analysis is better than that of maximum likelihood classification. Therefore, a new technique proposed in this study will be able to be put to practical use. Moreover this will play an important role in the construction of GIS database

1. 서 론

현대 과학기술의 급속한 발전으로 인간의 활동이 광역화, 다양화되고 인류역사상 유례없는 물질적 풍요를 인간이 향유하게 된 반면, 우리가 일상생활을 하고 있는 지구는 자연환경의 파괴, 자원의 고갈 등으로 고통받고 있으며, 더불어 그곳에 속해있는 모든 문명과 생명체의 생존이 위협받고 있다. 이러한 지구의 총체적 위기를 극복하기 위한 방안의 하나로 원격탐사 기술의 개발에 대한 필요성과 중요성이 널리 인식되어 왔다. 즉, 원격탐사 기술을 이용하여 지구의 물리적 성질과 환경적 특성에 관한 제반 정보를 수집하고, 지

*서울대학교 도시공학과 조교수
**서울대학교 자동화시스템공동연구소 특별연구원
***목포대학교 지적학과 전임강사

구상에 존재하는 모든 대상물의 관측, 해석 및 감시기능을 수행함으로써 범인류적인 공동대응을 모색하려는 것이다. 현재 세계 도처에서 나타나는 기상이변과 자연재해가 독립된 개별사건이 아닌 인류에 대한 자연의 위협이라는 인식이 가능하게 된 저변에는 원격탐사 기술의 발전이 자리하고 있다. 이러한 원격탐사 기술은 기상 및 환경정보 수집, 국토개발에 관한 각종 정보의 수집, 각종 주제도(thematic map)작성, 도시환경 및 토지이용정보의 추출, 농수산자원 조사, 지질조사, 군사시설물 및 이동상황 파악 등 광범위한 응용분야에 적용되고 있다.

일반적으로 원격탐사 데이터는 항공기나 인공위성의 센서(MSS, TM, HRV 등)를 사용하여 디지털 형태로 취득되며, 인공위성을 이용하는 경우 항공기에 의한 방법에 비해 광범위한 지역을 주기적으로 관측할 수 있으므로 대규모 프로젝트 수행시에는 인공위성 원격탐사 데이터를 사용하는 것이 경제적이다. 그러나 인공위성의 고도(LANDSAT 평균고도 약 705km)가 너무 높기 때문에 해상력의 한계성이라는 결정적인 단점이 있다. 항공기 탑재용 MSS의 경우 비행고도와 관측기기에 따라 원하는 해상력을 얻을 수 있으나, LANDSAT의 MSS는 79m, TM은 30m 그리고 SPOT의 HRV는 10m(XS는 20m)로서 센서에 따른 해상력이 이미 정해져 있다. 비록 각 위성보유국들이 보다 우수한 해상력을 가지는 센서의 개발에 심혈을 기울이고 있으나, 고고도 위성용 센서의 급속한 공간해상력 향상에는 상당한 기간이 소요될 것으로 예상된다. 이러한 해상력의 한계는 결국 분석정확도에 영향을 미치며, 특히 우리나라와 같이 토지 이용단위가 작고 여러가지 피복상태가 섞여 분포하는 경우, 그 정도는 더욱 심각하다. 그럼에도 불구하고 환경보호에 대한 인식과 국토 및 자원의 효율적인 관리에 대한 필요성이 더욱 요구되고 있으므로, 이에 대한 가장 효율적인 방안으로서의 원격탐사 기술이 보다 고도화되는 추세이다. 원격탐사 기술은 인공위성이나 항공기를 이용하여 취득한 데이터의 분석정확도를 향상시키려는 노력에 의하여 발전되어 왔다. 이러한 노력은 크게 세가지 측면 - 데이타의 상호결합에 의한 해상력 향상, 분광특성을 고려한 데이터의 질적개선, 분류기법의 고도화 - 으로 진행되어 왔으며, 그 결과 초기의 단순한 농작물 수확량 예측에서 현재는 복잡한 도시의 공간분포 해석까지 가능한 수준에 이르고 있다. 국내에서의 원격탐사 기술은 기상 및 해양분야에서 주로 이루어지고 있는 자연과학적 접근방법과 도시관련분야에서 이루어지고 있는 공학적 접근방법으로 나눌 수 있으며, 그 중에서 후자는 주로 토지피복분포에 대한 분류에 관한 것이다. 현재까지 도시관련분야에서 이루어진 성과들은 주로 이미 개발되어 있는 기술에 대한 확인 또는 적용에 그치고 있으며, 이 분야에 대한 연구도 활성화되지 못하여 비교적 미미한 수준에 그치고 있다. 이러한 상황의 주된 요인으로는 분석된 결과의 신뢰도와 실용성에 다소의 한계를 분석자 자신이 느끼고 있기 때문이다. 또한 연구의 결과가 적절히 적용될 수 있는 관련분야보다 비교적 빠른 시기에 이 분야에 대한 연구가 진행됨으로써, 원격탐사 기술의 수용폭이 확대되지 못한 것도 하나의 요인으로 볼 수 있다. 그럼에도 불구하고 최근의 환경, 교통, 도시 그리고 국토정보에 대한 국가적 관심과 투자는 원격탐사 기술의 확대와 고도화에 원동력을 제공하고 있으며, 무엇보다 위성보유국으로서의 위상이 이를 뒷받침하고 있다.

본 연구는 이러한 종합적인 상황 인식을 기반으로, 현재까지 도시관련분야에서 이루어진 원격탐사 기술의 성과에 대한 신뢰도와 실용성을 제고하고, 보다 고도화된 기술의 창출을 목적으로 한다. 즉, 이 분야에 대한 기존의 기술을 면밀히 검토함으로써 응용의 폭을 넓힘과 아울러 새로운 분류기법의 개발을 연구의 목적으로 한다. 구체적으로 말하면, 계량 지리학과 경제학 분야에서 주로 적용되는 다변량통계분석기법 중에서 정준상관분석(CCA: Canonical Correlation Analysis)을 사용하여 수치화상을 분석하는 기법의 개발에 관한 것이다.

2. 연구내용 및 방법

2.1 연구 개요

위성영상 분류에 있어서 무감독(unsupervised) 분류기법이 갖는 의의는 실제적인 토지피복(land cover)

분포에 관한 정보가 없이, 위성에서 관측된 대상물의 분광특성 자료만을 이용해서 통계적으로 토지피복분포의 분류작업을 자동적으로 수행할 수 있다는 것이다. 이러한 무감독 분류기법의 장점은 처리 속도가 매우 빠르다는 것이다. 무감독 분류는 군집화(clustering)라고도 하는데, 이는 분광공간 상에 도화된 수치화상 데이터를 자연스럽게 무리짓는다는 원리에 기초하기 때문이다. 현재 주로 사용되는 무감독분류기법으로 순차적(sequential) 군집화, 통계적(statistical) 군집화, ISODATA(Iterative Self-Organizing Data Analysis Technique) 군집화, RGB 군집화등이 있으며, 이들은 모두 분석자가 프로그램 실행 초기에 군집의 수, 군집간의 최소거리, 군집소거 임계치 등을 지정함으로써 이러한 초기변수에만 의존하여 분광특성분포, 즉 분광거리(spectral distance)에 의한 군집화 분류를 수행한다.2) 일반적으로 군집화 분류를 위해 입력되는 초기변수에 대한 적절한 선택은 분석자의 경험과 주관에 의존하기 때문에, 그 결과 분류된 군집들에 대한 객관성을 얻기가 어려우며, 주로 시각적 판단에 의해 이들 군집들을 재결합시키는 과정을 거치게 된다. 따라서 무감독 분류만으로는 정확한 분류결과를 얻기가 어렵기 때문에, 분류기준용 표본데이터 선정을 위한 사전분류 또는 감독 분류기법등의 전처리 작업에 사용되고 있다.

정준상관분석은 서로 상관관계 또는 인과관계를 갖는 두 변수군중에서 변수의 갯수가 적은 변수군을 예측변수군으로 하고 이에 대응하는 변수군을 기준변수군으로 하여 서로 매우 높게 상관되어 있는 변수들을 밝혀내는 통계적 기법이다.4),6),7) 이러한 정준상관분석은 각 변수군 내의 변수들의 설명되는 변량을 극대화하면서, 하나의 변수군으로부터 다른 변수군의 예측치를 극대화하는 구조를 밝힘으로써 두 변수군 내의 각 변수들의 상관관계를 제공한다. 따라서 정준상관분석에 의해 나타나는 각 변수들의 상관관계를 분류를 위한 판별함수로 사용할 수 있다. 본 연구는 이 기법을 수치화상분석에 적용하는 방법으로 개발하는데 목적이 있다. 수치화상분석으로의 적용은 앞서 설명된 무감독 분류기법의 장점을 활용하고 단점을 보완하는 측면에서의 접근, 즉 무감독 분류기법과 정준상관분석을 결합한 새로운 분류기법의 개발을 말한다.

개발된 분류기법의 효용성 평가를 위해 동일 위성데이타에 대해 기존의 분류기법 중 최대우도분류를 수행하고, 지상실제데이터를 기준치로하여 이를 비교평가 한다.

2.2 무감독분류기법과 정준상관분석을 결합한 분류알고리즘 개발

기준변수(Q)

T/F / band	1	2	.....	q
1	x_{11}	x_{12}	.....	x_{1q}
2	x_{21}	x_{22}	.....	x_{2q}
.		.		
.		.		
n	x_{N1}	x_{N2}	.....	x_{Nq}

예측변수(P)

개별군집 / band	1
1	y_1
2	y_2
.	.
.	.
n	y_N

T/F : Training Field, 분류기준용 표본데이터의 토지피복항목

기준변수군의 q개의 변수는 원래의 수치화상에서 식별이 분명한 대상들에 대한 분류기준용 표본데이터들이며, 자료행렬의 각 요소값(xnq)은 각 밴드별 분류기준용 표본데이터별 화소값들에 대한 평균값 또는 대표값을 Z-score를 사용하여 변환한 값이다. 예측변수군의 1개의 변수는 무감독분류에 의한 분류결과의 군집들 중 한 개 이며, 요소값(yn)은 각 밴드별 군집별 화소값들에 대한 평균값을 Z-score를 사용하여 표준화한 값이다.

이 방법이 갖는 의의는 각 밴드의 분광특성만을 이용하여 매우 빠른 속도로 순수히 통계적으로 분류하여 군집을 만든 후, 최소한의 사전정보를 이용하여 정확하게 지정된 분류기준용 표본데이터들 중 각 군집과 상관관계가 가장 높은 토지피복을 찾아냄으로써 효율적인 분류결과의 통합을 수행할 수 있다는 점이다. 즉, 무감독분류 수행후의 군집들에 대한 재통합에 객관성을 부여할 수 있게 된다. 또 다른 의의는 무감독분류시 초기변수들에 대한 의존성을 줄임으로써, 즉 군집수와 군집간의 최소거리를 고정값으로 하여 군집형성에 있어 작업자의 의도가 포함되는 시행착오과정을 배제한 상태의 군집을 형성한 후, 이들을 정준상관분석에 의해 결합하는 새로운 분류기법을 제시하는데 있다.

2.3 연구수행과정

무감독분류 후 정준상관분석에 의한 분류와 이 결과에 대한 정확도 및 효용성 평가 수행과정을 흐름도로 작성하면 그림 3.1과 같다. 위성데이터의 정준상관분석을 위해 MATLAB을 이용하여 프로그램을 작성하였다. 위성데이터에 대한 정준상관분석 과정을 단계별로 구체적인 수행내용에 대해 정리하면 다음과 같다.

(1) 자료 취득(Read Data)

분류할 원 영상을 군집분석한 후, 밴드별로 독립되고 헤더 bytes가 없는 binary 파일로 구성한다. Thematic Maper 데이타의 경우 밴드수가 7개이므로 7개의 파일로 구성된다. 또한 분류할 토지피복을 대표하는 각 토지피복별 분류기준용 표본데이타를 취득하기 위해 표본데이타로 사용할 영역의 밴드별 평균값을 계산하여 m밴드에 n항목의 텍스트파일을 작성한다. 본 논문의 경우 토지피복항목의 수를 4개로 선정하였다.

(2) 기본 자료행렬의 구성

(1)에서 취득된 무감독 분류 후의 위성데이타로부터 분류할 1군집에 대해 m×1 행렬로 만들고 이를 예측변수군으로 둔다. 또한 분류기준용 표본데이타 파일로 부터 m×n 행렬을 구성하여 기준변수군으로 한다. 다음으로는 기준변수군 행렬과 예측변수군 행렬을 병합하여 m×(1+n) 기본자료행렬을 구성한다.

(3) 상관계수행렬 계산 및 분할

(2)의 과정에서 생성된 자료행렬로 부터 (1+n)×(1+n)의 상관계수행렬을 산출한다. 이 상관계수행렬이 정준상관분석에 있어 고유치와 고유벡터를 생성하기 위한 기본행렬이 된다.

산출된 상관계수행렬은 (1+n)×(1+n) 행렬이므로 이를 1×1, 1×n, n×1, n×n 의 독립행렬로 분할한다. 이렇게 분할된 각 행렬로 부터 정준상관분석이론에서 요구되는 역행렬, 제곱근행렬, 그리고 제곱근 행렬의 역행렬을 개별적으로 계산한다.

(4) 정준방정식 계산

계산된 분할행렬을 이론식에서 제시된 형태로 재구성하여 (3)에서 언급한 기본행렬(M)을 만든다. 이 행렬에 대해 정준방정식을 구성하여 고유벡터와 고유값을 산출한다. 산출된 고유벡터로 부터 두 변수군의 정준벡터(정준가중치)인 열벡터 $(^{a}_{11})$, $(^{b}_{11})$ 를 계산한다. 고유치의 개수는 예측변수군의 수 만큼 생성되므로 본 논문에서는 1개만 존재한다. 열벡터 **b** 의 원소의 개수는 토지피복의 개수를 나타내는 것으로서, 원소의 값은 정준상관관계에 대한 각 토지피복항목의 비중을 나타낸다. 원소 중 가장 큰 값이 해당되는 토지피복항목을 각 군집의 토지피복으로 결정하는 것이 정준상관분석에 의한 위성영상분류과정의 핵심이다.

(5) 고유치의 통계적 유의성 검정

정준상관분석에 의해 토지피복항목을 결정하기 위해서는 두 변수군의 공분산 Σ_{12} 가 zero가 아님이 선행되어야 한다. 즉 정준상관계수(고유치의 양의 제곱근)가 zero가 아니어야 한다. 이를 확인하기 위해 Fisher가 제안한 고유치의 분포함수를 채택하여 유의수준 α(=5%)에서 고유치가 통계적으로 유의한가, 즉 두 변수군간에 상관관계가 밀접한 지를 검정한다.1) 이 때 각 군집은 귀무가설이 기각되면 토지피복항목이 결정되는 것이며, 귀무가설이 채택되면 어느 토지피복항목으로도 소속될 수 없다고 판단되는 것이므로

미분류 군집으로 지정되게 된다. 이와 같은 미분류 군집은 분류대상지역내에 선정된 토지피복항목과 매우 다른 특성을 가진 지역이 존재한다는 것을 짐작하게 해 준다.

3. 위성영상분류에 적용

3.1 연구대상영역

연구대상영역은 일반적으로는 위성영상으로부터 다양한 토지피복이 혼합된 직사각형 영역으로 적당히 절출되나, 이 경우 분류결과에 대한 정확도 평가를 위한 지상실제데이타를 제작하는데 어려움이 따르며 상당한 오차가 포함될 우려가 있다. 따라서 본 논문에서는 명확한 정확도 비교평가를 위해 1/25,000 지형도를 참조하여 위성영상으로부터 각 토지피복항목별로 거의 100% 분명한 지역을 절출한 후, 이를 모자이크하여 하나의 분류할 대상영역으로 삼았다. 인공위성 데이타는 LANDSAT-5 Thematic Mapper 데이타(1992. 6. 2)를 사용하였으며, 위치는 서울을 포함하는 직사각형 영역이다. 토지피복은 산림(Forest), 논(Rice Field), 도시역(Urban), 수역(Water)의 4가지로 구성하였다. 이는 U.S.G.S.의 토지피복분류체계에 의해 TM 데이타의 해상도 한계내에서 분류될 수 있는 분류코드체계를 따른 것이다.3) 선정된 4가지 토지피복항목 외에 초지나 나대지 등도 토지피복항목으로 존재하기는 하나, 이 항목들은 분포영역이 소규모인 관계로 필요한 만큼 절출해 내기가 어려워 제외하였다. 영역의 크기는 토지피복항목별로 40화소×50화소(행×열)로 하여 전체를 80화소×100화소로 만들었다. 산림과 도시역은 넓게 분포하는 항목이므로 위성영상으로부터 40화소×50화소를 한 번에 절출하였고, 논과 수역은 한 번에 직사각형으로 절출될 수 있는 영역이 없었기 때문에 20화소×25화소를 4개씩 절출하여 모자이크함으로써 40화소×50화소로 만들었다. 각 토지피복항목의 선정위치는 다음과 같다. 산림의 경우 관악산에서 절출하였고, 논은 김포평야지역에서, 도시역은 강남구와 서초구의 경계와 역삼동지역을 포함하는 영역에서, 수역은 행주대교, 한강대교, 잠실대교, 강동대교 부근의 한강 4곳에서 각각 절출하였다. 본 연구에서의 토지피복항목은 다음과 같은 특성을 가지고 이루어져 있다.

① 산림 : 약 2m 이상의 나무들로 이루어진 임야지역을 지칭

② 논 : 위성데이타 취득시기가 6월 2일 이므로, 논에 물이 잠겨있는 상태로서 벼와 물이 함께 분포하는 상태임

③ 도시역 : 도시지역내의 대형 아파트 단지, 도로, 소규모 건물, 일반주택, 콘크리트 시설물, 기타 인공 구조물 등을 지칭

④ 수역 : 한강의 물을 선택한 것으로서, 선정된 4개 장소의 수심이나 탁도등의 차이로 인해, 같은 물이라도 각 장소별 분광반사특성이 상당한 차이가 있음

연구대상영역의 위성데이타는 사진 1과 같다.

3.2 분류기준용 표본데이타(Training Areas) 선정

정준상관분석기법과 기존의 가장 널리 사용되는 최대우도분류를 위성영상데이타에 실제 적용하기 위해서는 우선적으로 분류의 기준이 되는 표본데이타가 선정되어야 한다. 본 논문은 정준상관분석의 시험이 연구의 주 목적이므로 표본데이타에 대한 특별한 선정기준은 없다. 즉, 본 연구는 새로운 분류기법의 개발과 그 효용성 확인에 의미가 있다. 그러므로 표본데이타 선정시 시행착오를 거쳐 가능한 한 분류정확도가 절대적으로 높도록 할 필요는 없다. 특히, 본 논문에서 시험적용되는 영역은 의도적으로 확실한 토지피복항목만을 선정하여 구성하였으므로 각 항목별로 무작위로 선택하면 된다. 따라서 화면에 도시된 대상영역으로부터 각 항목별로 적당히 중간정도의 분광수치를 갖는다고 판단되는 위치에서 선형으로 25화소씩 선택하였다. 수역은 분광특성값이 조금씩 차이가 나는 4개의 영역이 복합되어 이루어져 있으므로, 마찬가지로 화면상에서 대략 중간색으로 나타나는 하나의 영역에서 25화소를 선택하였다. 선정된 분류기준용 표본데이타의 위치는 사진 2와 같다. 표본데이타의 통계적 수치는 표 3.1과 같다.

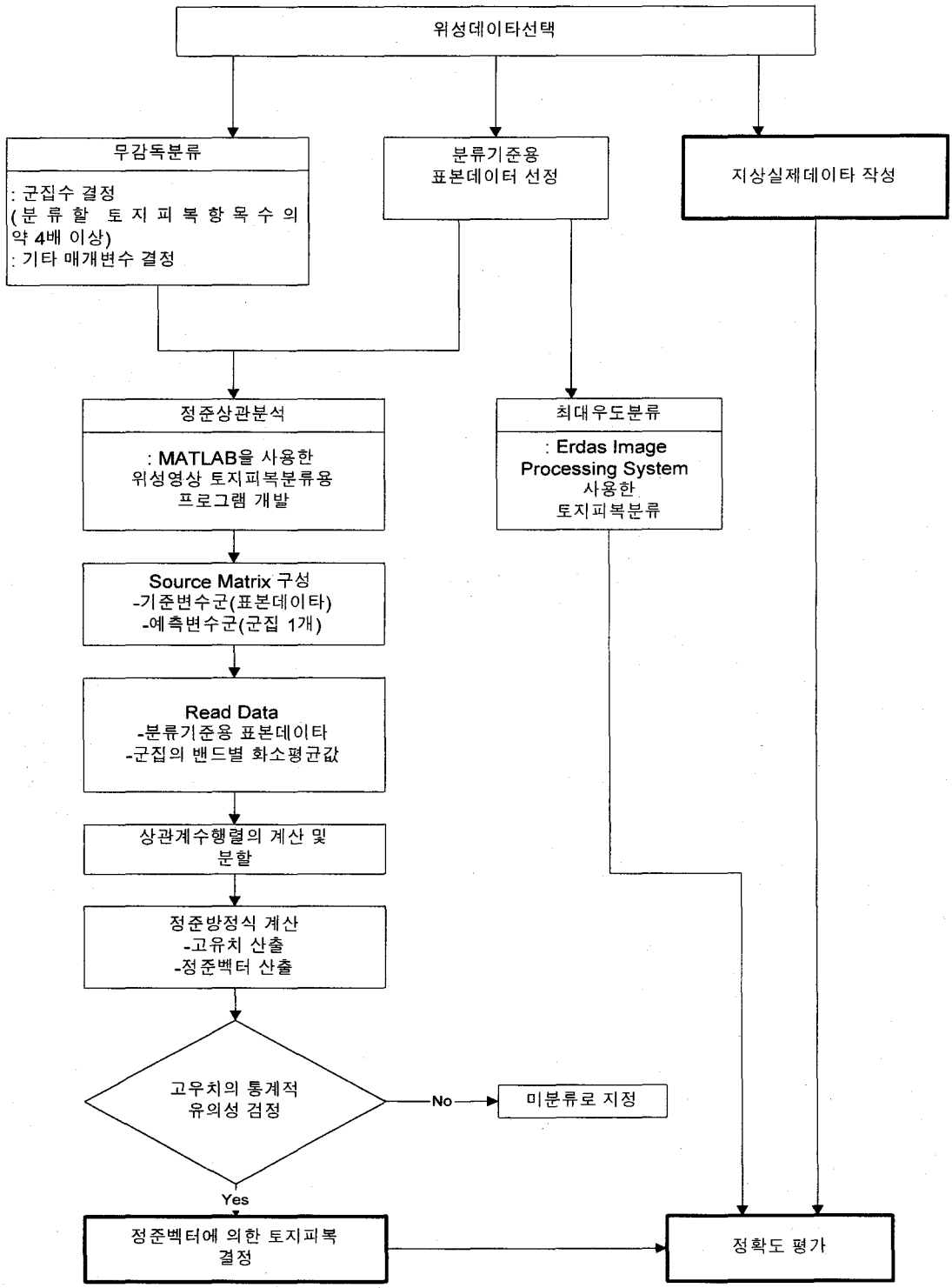
위성데이타선택
무감독분류
: 군집수 결정
(분 류 할 토 지 피 복 항 목 수 의 약 4배 이상)
: 기타 매개변수 결정
분류기준용
표본데이터 선정
지상실제데이타 작성
정준상관분석
: MATLAB을 사용한
위성영상 토지피복분류용
프로그램 개발
최대우도분류
: Erdas Image
Processing System
사용한
토지피복분류
Source Matrix 구성
-기준변수군(표본데이타)
-예측변수군(군집 1개)
Read Data
-분류기준용 표본데이타
-군집의 밴드별 화소평균값
상관계수행렬의 계산 및
분할
정준방정식 계산
-고유치 산출
-정준벡터 산출
고우치의 통계적
유의성 검정
No
미분류로 지정
Yes
정준벡터에 의한 토지피복
결정
정확도 평가

표 3.1 분류기준용 표본데이타의 통계수치

Number of Points = 25	산림						
Band	1	2	3	4	5	6	7
Minimum	89	38	37	67	69	149	23
Mean	91.04	39.44	39.04	87.80	76.32	152.16	25.20
Standard	1.11	0.64	1.61	9.04	3.89	1.51	1.96
Maximum	93	41	42	103	86	156	29

Number of Points = 25	논						
Band	1	2	3	4	5	6	7
Minimum	107	48	53	49	21	156	8
Mean	108.80	49.32	56.96	51.92	29.08	157.40	12.64
Standard	0.94	0.68	1.91	2.02	4.76	1.20	2.71
Maximum	111	51	60	56	37	160	18

Number of Points = 25	도시역						
Band	1	2	3	4	5	6	7
Minimum	106	47	55	51	69	173	41
Mean	109.04	48.80	58.84	54.56	76.40	174.24	47.00
Standard	1.54	1.02	2.03	1.90	4.57	1.03	3.15
Maximum	112	51	62	57	86	176	53

Number of Points = 25	수역						
Band	1	2	3	4	5	6	7
Minimum	101	44	44	27	16	136	5
Mean	102.68	44.76	45.28	28.08	17.48	136.56	7.56
Standard	1.12	0.43	0.53	0.69	0.94	0.98	0.85
Maximum	105	45	46	29	19	140	9

3.3 위성영상분류

3.3.1 정준상관분석에 의한 분류

3.3.1.1 무감독분류

대표적인 4가지 무감독분류방법 중 일반적으로 가장 많이 사용되는 순차적 군집화 방법을 채택하였다.2) 소프트웨어는 현재 널리 활용되는 ERDAS Image Processing System을 사용하였다. 여기서 채택된 고정 매개변수는 다음과 같다.

- 분류하고자 하는 군집수(N) = 15
- 분리될 두 군집 평균(중심)의 최소 분광거리(R) = 8.75
- 군집을 병합할 때 사용되는 거리 매개변수(C) = 12.25
- 군집의 병합전 까지 하나의 군집으로 분석될 화소의 갯수(M) = 100
- 군집을 제거하기 위한 M개 화소에 대한 임계 백분율(T) = 1

위에서 채택된 매개변수 중 군집수 N은 토지피복항목수를 고려하여 10, 15, 20, 30을 각각 적용한 후 분류가 가능한 값 중 가장 작은 값으로 결정하였다. N이 10인 경우는 반드시 서로 구분되어야 할 요소들이 하나의 군집으로 생성되는 결과가 나타났으므로 적절치 않았다. 나머지 매개변수인 R, C, M, T는, 본 연

구가 무감독분류와 정준상관분석을 결합한 분류의 가능성 및 효용성에 있으므로, 가능한 한 정확한 분류가 되도록 의도적인 시행착오를 거치지 않고 소프트웨어에서 기본으로 제공되는 값을 그대로 사용하였다. 이상의 매개변수를 적용하여 무감독분류한 결과를 화면에 도시하면 사진 3과 같다.

3.3.1.2 무감독분류결과의 정준상관분석

2.2의 분류알고리즘을 적용하면 각 군집은 구체적인 토지피복항목으로 결정된다. 무감독분류의 결과인 군집별, 밴드별 화소평균값과 각 군집의 토지피복항목(유의성 검정시 미분류 군집 포함)은 표 3.2와같다. 이는 군집수를 15개로 고정한 결과로서, 군집수를 20개, 30개로 고정한 결과에 비해 거의 차이는 없지만 오히려 더 좋은 분류결과를 보여주고 있다. 일반화하여 결론으로 주장할 수는 없지만, 이는 분류하고자 하는 토지피복항목의 수를 고려하여 군집수를 결정해야 함을 암시하는 것이라 판단된다.

각 군집에 대해 5% 유의성 검정을 실시한 결과, 지상실제데이타가 도시역인 지역에서 2개의 군집이 검정을 통과하지 못하여 미분류로 지정되었다. 그러나 2개의 군집을 화소수로 계산했을 때 전체 8000개의 화소 중 2개 화소만이 미분류로 지정되었으며, 또한 정확도 평가시 최대우도분류결과와 동일한 조건으로 비교하기 위하여, 본 연구에서는 유의성 검정을 수행하지 않은 결과를 사용하였다. 토지피복이 결정된 결과를 화면에 도시하면 사진 4와 같고, 토지피복별 화소수는 표 3.3과 같다.

표 3.2 무감독분류 및 정준상관분석 결과

밴드 / 군집	1	2	3	4	5	6	7	비율(%)	토지피복
1	91.05	39.42	38.80	90.19	78.58	155.43	26.74	12.625	산림
2	90.49	38.61	38.39	74.68	65.86	152.78	23.53	12.375	산림
3	107.48	48.68	54.84	51.00	27.38	155.81	11.56	21.163	논
4	108.67	49.29	55.84	56.29	46.40	156.60	20.91	3.725	논
5	108.78	48.71	58.66	55.26	79.15	174.79	48.34	21.000	도시역
6	106.70	46.67	48.05	30.33	16.67	136.38	7.16	25.000	수역
7	113.36	53.41	67.18	64.68	110.36	174.82	70.27	0.412	도시역
8	117.00	59.20	80.60	78.00	146.20	173.20	94.20	0.062	도시역
9	114.00	57.50	79.00	78.00	166.50	173.50	110.00	0.025	도시역
10	122.00	64.00	92.00	86.00	156.00	172.00	115.00	0.013	미분류(도시역)
11	118.00	62.00	86.00	84.00	185.00	172.00	117.00	0.013	미분류(도시역)
12	113.00	54.57	73.14	71.71	132.14	174.43	80.57	0.112	도시역
13	106.09	45.45	52.00	41.82	49.64	172.00	29.45	0.262	도시역
14	114.00	51.50	64.50	54.50	59.00	172.00	34.00	0.775	도시역
15	109.00	51.00	62.00	67.00	96.00	175.00	58.00	2.437	도시역

표 3.3 정준상관분석 결과

토지피복 / 내용	산림	논	도시역	수역
화소수	2000	1991	2009	2000
GIS value	1	2	3	4
백분율(%)	25.00	24.89	25.11	25.00

3.3.2 최대우도분류

정준상관분석에서 사용한 분류기준용 표본데이타와 동일한 데이타를 사용하여, 기존의 분류기법인 최대우도분류를 적용한 결과는 표 3.4와 같고, 화면에 도시하면 사진 5와 같다.

표 3.4 최대우도분류의 결과

토지피복 내용	산림	논	도시역	수역
화소수	2000	2475	1998	1527
GIS value	1	2	3	4
백분율(%)	25.00	30.94	24.97	19.09

4. 분류정확도 평가

4.1 지상실제데이타 작성

위의 두가지 분류방법에 대해 위성데이타가 얼마나 정확하게 분류되는가를 평가하기 위해서는 비교기준데이타로 지상실제데이타가 필요하다. 본 연구에서는 3.1절에서 언급한 것처럼 토지피복의 내용을 알고 있는 데이타로부터 연구대상영역을 만들었으므로, 각 토지피복별로 GIS 수치를 지정함으로써 지상실제데이타를 작성할 수 있다. 작성된 지상실제데이타는 사진 6과 같다. 지상실제데이타의 내용은 표 4.1과 같다.

표 4.1 지상실제데이타의 내용

토지피복 내용	산림	논	도시역	수역
화소수	2000	2000	2000	2000
GIS value	1	2	3	4
백분율(%)	25.00	25.00	25.00	25.00

4.2 분할행렬표 작성 및 평가

무감독분류후의 정준상관분석에 의한 분류와 최대우도분류에 의한 결과의 정확도를 비교평가하기 위해 표 4.2, 4.3, 4.4, 4.5와 같이 분할행렬표(Contingency Table)를 작성하고 분류정확도를 산출하였다. 대상지역의 총 화소수는 8000 개 이며, 바르게 분류된 화소수는 각 분할 행렬표의 주 대각선 요소를 모두 합한 값이다. 이상의 값을 이용하여 전체정확도(Overall Accuracy)와 각 토지피복항목별 Producer's Accuracy, User's Accuracy를 구하였다. 여기서 전체정확도란 바르게 분류된 화소수를 총화소수로 나눈 값이고, Producer's Accuracy 는 각 토지피복항목별로 지상실제데이타가 바르게 분류된 화소 개수의 비율을 말하는 것이며, User's Accuracy 는 각 토지피복항목으로 분류된 개수 중 정확하게 분류된 화소수의 비율을 말한다.5) 따라서 Producer's Accuracy 에서는 Omission Error 가 포함되며 User's Accuracy 에서는 Commission Error 가 포함된다. 예를 들면 표 4.2의 분할 행렬표에서 논의 지상실제데이타 영역은 화소수로 총 2000개 이다. 이 중 옳게 분류된 화소수는 1991개 이며, 잘못 분류(Omission Error)된 화소수는 표의 세로방향으로 나열된 바와 같이 산림으로 0개, 도시지역으로 9개, 수역으로 0개 이다. 또한 표의 가로방향으로 나타난 오분류(Commission Error) 화소는 어느 토지피복에서도 나타나지 않았다. 즉 원래 타 토지피복항목이 논으로 분류된 화소는 존재치 않는다. 결국 논으로 분류된 총화소수는 표의 오른쪽 열에 나타난 바와 같이 1991개가 된다. 표 4.2의 나머지 토지피복항목들도 이상과 같은 방식으로 정리된 분류 결과이다. 표 4.2의 분류결과를 가지고 분류정확도를 계산한 것이 표 4.3이다. 앞에서 언급하였던 것처럼 논의 Producer's Accuracy 를 구하면 2000개 중 1991개가 바르게 분류되었으므로 99.55% 이며, User's Accuracy 는 논으로 분류된 개수가 1991개 이므로 1991을 1991로 나누면 100% 가 된다. 나머지 토지피복항목들도 동일한 방식으로 계산된다. 전체정확도는, 옳게 분류된 화소수가 총 7991개 이므로 이 값을 대상지역의 총화소수인 8000으로 나누면 99.89% 가 된다.

표 4.2 정준상관분석 결과에 대한 분할 행렬표 (단위 : 화소수)

지상실제 / 분류결과	산림	논	도시역	수역	합계(row)
산림	2000	0	0	0	2000
논	0	1991	0	0	1991
도시역	0	9	2000	0	2009
수역	0	0	0	2000	2000
합계(column)	2000	2000	2000	2000	8000

표 4.3 정준상관분석에 의한 분류정확도

종류 / 분류항목	Producer's Accuracy	User's Accuracy	전체 정확도
산림	2000/2000 = 100%	2000/2000 = 100%	2000+1991+2000+2000 = 7991 7991/8000 = 99.89%
논	1991/2000 = 99.55%	1991/1991 = 100%	
도시역	1999/2000 = 100%	2000/2009 = 99.55%	
수역	2000/2000 = 100%	2000/2000 = 100%	

표 4.4 최대우도분류 결과에 대한 분할 행렬표 (단위 : 화소수)

지상실제 / 분류결과	산림	논	도시역	수역	합계(row)
산림	2000	0	0	0	2000
논	0	2000	2	473	2475
도시역	0	0	1998	0	1998
수역	0	0	0	1527	1527
합계(column)	2000	2000	2000	2000	8000

표 4.5 최대우도분류 정확도

종류 / 분류항목	Producer's Accuracy	User's Accuracy	전체 정확도
산림	2000/2000 = 100%	2000/2000 = 100%	2000+2000+1998+1527 = 7525 7525/8000 = 94.06%
논	2000/2000 = 100%	2000/2475 = 80.81%	
도시역	1998/2000 = 99.90%	1998/1998 = 100%	
수역	1527/2000 = 76.35%	1527/1527 = 100%	

이상의 2가지 분류결과에서 보면 전체 정확도는 무감독분류 후 각 군집에 대해 정준상관분석기법을 적용한 것이 최대우도분류에 의한 결과보다 좋게 나타났다. 일단, 이 결과만을 가지고 생각해 볼 때 본 논문에서 제안하는 정준상관분석에 의한 분류기법이 상당히 유효하다고 판단된다. 위의 2가지 분류방법에 의한 분류정확도가 모두 매우 높게 나타난 것은 연구대상지역의 위성데이타를 인위적으로 토지피복이 분명한 지역에서 추출하여 만들어 냈기 때문이다. 이는 오차가 거의 없는 정확도 평가를 목적으로 하였기 때문이며, 따라서 여기서는 정확도가 절대적으로 높은 것은 의미가 없으며 2가지 분류방법의 상대적인 정확도만이 의미가 있을 뿐이다. 실제로 선택되는 위성데이타에 따라 분류정확도의 변화가 심하리라 예상된다. 특히 본 논문에서 선택된 위성데이타는 최대우도분류 적용시 분류기준용 표본데이타의 변화에 따라 분류정확도가 심한 편차를 보였으나, 무감독분류 후의 정준상관분석에 의한 분류는 거의 변화없이 비슷한 정확도를 보여주었다. 한가지 예를 들면 본 논문에서 사용한 분류기준용 표본데이타가 아닌 다른 2가지 표본데이타로 2가지 분류방법을 각각 적용한 결과 최대우도분류에서는 표본데이타에 따라 분류정확도가 약 100%와 85%의 심한 편차를 보였으나, 정준상관분석을 이용했을 때는 무감독분류과정을 거치므로 90%이상의 동일한 분류정확도를 나타내었다.

토지피복항목별 정확도를 알기 위해서는 Producer's Accuracy 와 User's Accuracy 를 산출한다. 각 분류방법의 Producer's Accuracy 를 서로 비교해 보면 정준상관분석에 의한 분류는 모든 항목이 거의 100%로 나타났는데, 최대우도분류에서는 산림, 논, 도시역은 거의 100%이나 수역은 76.35%로 나타났다. 이는 3.1절에서 언급한 것처럼 수역의 위성데이타를 분광수치가 많은 차이가 나는 여러지역에서 추출하여 모자이크 하였기 때문인 것으로 풀이된다. 즉 최대우도분류는 수역의 분류기준용 표본데이타 선정위치에 따라 분류결과에 심한 변화가 생기게 한다.

이와같이 정준상관분석을 이용한 분류는, 기존의 분류기법 중 대표적인 방법인 최대우도분류가 갖는 단점 중에 하나인 분류기준용 표본데이타의 변화에 따른 분류정확도의 심한 편차를 없애주는 장점을 가지고 있음을 알 수 있다. 이로써 무감독분류 후의 정준상관분석에 의한 분류방법을 사용하면 분류기준용 표본데이타의 선정에 신경을 쓰지 않아도 됨을, 즉 표본데이타의 선정이 매우 용이함을 알 수 있다.

이상과 같은 결과에 의하면 정준상관분석에 의한 분류기법은 이분야에서 가장 널리 알려져 있는 최대우도분류보다 전체적인 면에서도 분류능력이 뒤지지 않는다고 판단되며, 상황에 따라 오히려 유용하게 사용되리라 생각된다. 물론 이와같은 판단은 본 연구에서 나타난 결과에 의한 추정이므로, 일반적으로 모든 경우에 적용된다고 단정지을 수는 없음을 밝힌다.
분류결과를 종합하여 정리하면 표 4.6과 같다.

표 4.6 정준상관분석 분류, 최대우도분류 정확도 비교표 (단위 : %)

항목 / 분류기법	전체 정확도	Producer's Accuracy				User's Accuracy			
		산림	논	도시역	수역	산림	논	도시역	수역
정준상관분석	99.88	100	99.56	99.95	100	100	99.95	99.56	100
최대우도분류	94.06	100	100	99.90	76.35	100	80.81	100	100

5. 결론

본 연구에서 적용된 정준상관분석(CCA)은 토지피복항목들에 대한 변수군과 분류하고자 하는 개개의 관측치들에 대한 변수군간의 상관관계를 최대로 하는 새로운 직교축을 설정하는 선형변환이며, 사용 데이타의 정규성이 일반적으로 요구된다.7) 또한 토지피복항목 변수군의 값들은 표본추출로 얻어진다. 따라서 이 분석방법도 최대우도분류와 마찬가지의 제약조건, 즉 사용 데이타의 정규성과 판별함수 결정에 사용되는 표본선정에 대한 고려가 필요하다. 그러나 정준상관분석에서의 판별함수는 정규확률밀도함수가 아닌 상관관계를 최대로 하는 선형변환식이며, 또한 표본과의 관련성을 가지지만 이들로부터 판별함수가 직접 유도되지는 않는다. 따라서 분류기준용 표본데이타 선정에 어느 정도 유연성을 가지며, 그 결과 일반적인

분류기법들이 분류기준용 표본데이타 선정에 의해 겪게되는 상당한 시행착오를 정준상관분석에서는 피할 수 있었다.

결론적으로 본 연구의 수행결과에 대한 의미와 함께 기존의 분류기법과 다른 특징 및 장점을 정리하면 다음과 같다.

첫째, 정준상관분석 알고리즘의 개념상 분류기준용 표본데이타 선정시 시행착오를 겪지 않고도 정확한 분류를 할 수 있었다. 즉 분류기준용 표본데이타 선정이 용이해 졌다.

둘째, 정준상관분석에 의해 결정된 각 군집의 토지피복항목을 사용하여 분류기준용 표본데이타를 효율적으로 선정하거나, 최대우도분류의 사전정보로 활용한다.

셋째, 동일한 분류기준용 표본데이타 사용시, 무감독분류 후 정준상관분석에 의한 분류가 최대우도분류보다 우월한 결과를 나타내었다. 더구나 지상실제 데이타의 일부지역(예를 들면 논의 잠겨 있는 물, 도시역 내의 식물)이 정준상관분석에서는 바르게 분류하는 지역과 다르게 작성된 경우를 감안하면 정준상관분류의 분류정확도는 더 향상될 수 있다. 총체적으로 말하면 정준상관분석에 의한 분류기법은 최대우도분류기법과 비교하여 전반적으로 분류정확도가 우수하다고 말할 수 있다.

넷째, 이상과 같은 결과로 판단해 볼 때 본 연구에서 시도된 무감독분류 후의 정준상관분석에 의한 분류기법은 원격탐사의 분류기법 분야에서 실용화 될 수 있으며, 나아가서는 GIS 데이타베이스 구축에 중요한 역할을 할 수 있을 것이다.

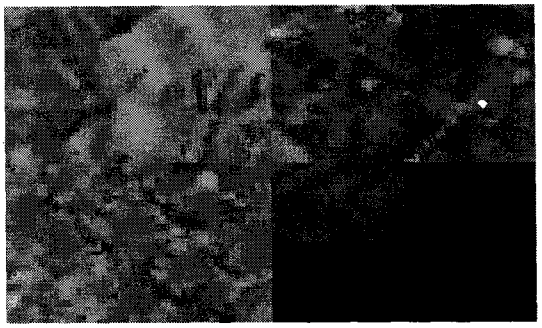

사진 1. 연구대상영역

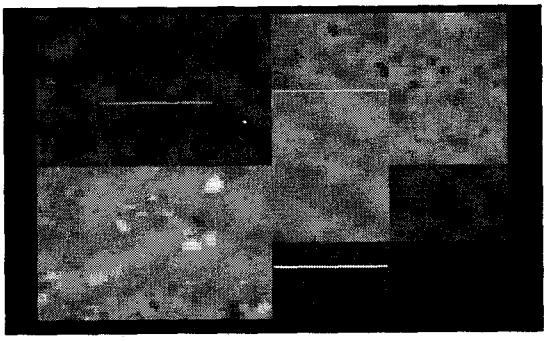

사진 2. 분류기준용 표본데이타의 위치

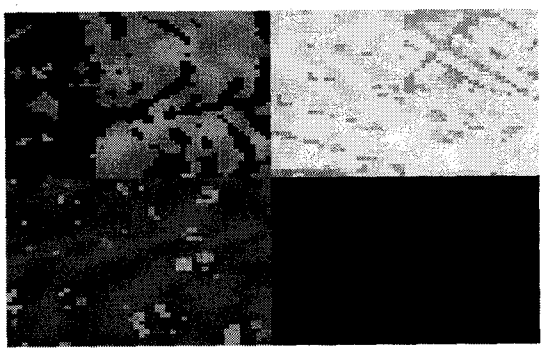

사진 3. 무감독분류 결과

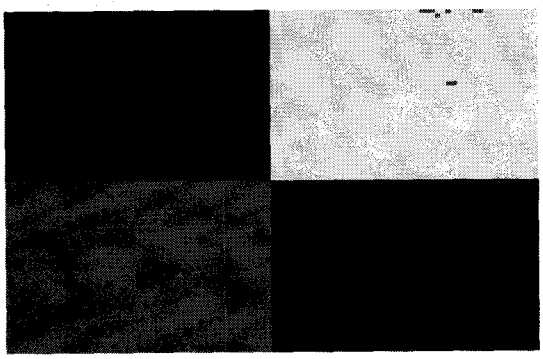

사진 4. 무감독분류 후 정준상관분석 결과

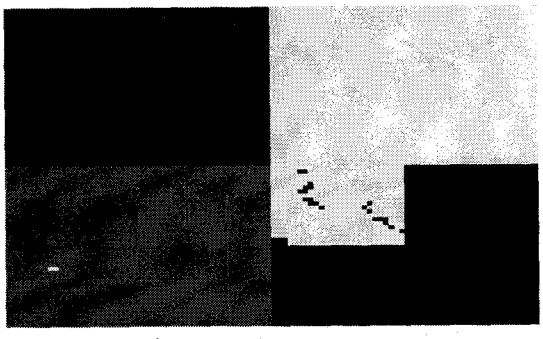

사진 5. 최대우도분류 결과

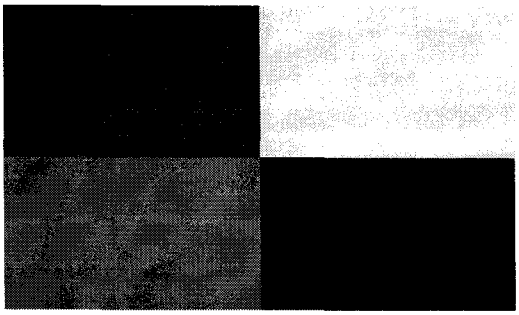

사진 6. 지상실제데이타

산림 논 도시역 수역

감사의 글

이 논문은 1995년도 한국학술진흥재단의 자유공모과제 연구비에 의하여 연구되었습니다. 연구비를 지원해 주신 한국학술진흥재단에 감사를 드립니다.

참 고 문 헌

1. Bartlett, M. S., "The Statistical Significance of Canonical Correlations", Biometrica, Vol. 32, 1941.
2. "ERDAS Field Guide ", Erdas, Version 7.4, 1990.
3. Jensen, John R., "Introductory Digital Image Processing-A Remote Sensing Perspective", Prentice-Hall, 1996.
4. Johnson, Richard A. and Wichern, Dean W., "Applied Multivariate Statistical Analysis", Third Edition, Prentice Hall, 1992.
5. Lillesand, Thomas M. and Kiefer, Ralph W., "Remote Sensing and Image Interpretation", Third Edition, John Wiley & Sons, 1994.
6. 남영우, "계량 지리학", 법문사, 1992.
7. 이희연, "지리 통계학", 법문사, 1991.