국토연구 제103권(2019.12): pp115~129 http://dx.doi.org/10.15793/kspr.2019.103..07

# Iterative Trimming Algorithm을 이용하여 자동추출된 KOMPSAT-3A 훈련자료 신뢰성 평가

## Reliability Evaluation of KOMPSAT-3A Training Data Automatically Selected Using Iterative Trimming Algorithm

조기환 Cho KiHwan*, 정종철 Jeong Jongchul**

**Abstract**

Image classification is one of the key issues of remote sensing technology and selecting training data is an essential process in supervised image classification. Dramatically increasing imagery data require more effective and automated classification techniques. The traditional process of selecting training data requires intensive manpower and, as a result, it has been costly and time-consuming. This study proposed an automatic training data extraction technique using outdated geographic information system (GIS) data and its applicability was tested. We used a high-resolution KOMPSAT-3A satellite image taken on July 7, 2018, and the land cover map in 2015 for the test of automated training data extraction based on the iterative trimming algorithm. First, the training data were extracted based on the polygon of the land cover map. Then, the probability distributions of each land cover class were estimated using kernel density estimation. The outliers were removed in the order of low probability. The bootstrap technique was used to determine the ratio of removing outliers. The ratios were different among the land cover classes. The removing ratio was 0.08 for the urbanized area, 0.16 for agriculture/land, 0.04 for forests, 0.16 for bare soil and 0.04 for water. With the refined training data, image classification was conducted. This approach allows automatic extraction of training data based on GIS data without manual digitizing. It is expected to contribute to an automatic and timely update of the urban land cover map with high-resolution imagery.

Keywords: Training Sample, Land Cover Map, Bootstrap, Image Classification, KOMPSAT-3A

## I. 서론

영상분류는 원격탐사 기술의 핵심 적용 분야 중 하나로, 위성영상에서 나타나는 분광특성을 이용하여 대상물을 유형별로 나누는 기술을 뜻한다. 이를 이용하여 제작되는 대표적인 성과물로 토지피복지도가 있다. 토지피복지도는 국토계획을 수립하는 데 중요한 역할을 한다.

토지피복지도가 효과적으로 이용되기 위해서는 토지피복변화를 충실히 반영할 수 있도록 지속적인 갱신이 필요하다. 현재 제작되고 있는 세분류 토지피복지도는 항공사진과 고해상도 위성영상을 이용하여 제작되어 정확도는 높은(환경공간정보서비스[1]) 반면 비용과 작업 시간 등의 제약으로 인해 자료의 갱신 속도

* 영남대학교 연구원(제1저자) | Researcher, Yeungnam Univ. | Primary Author | khcho@ynu.ac.kr
** 남서울대학교 공간정보공학과 교수(교신저자) | Prof., Dept. of GIS, Namseoul Univ. | Corresponding Author | jjc1017@gmail.com

가 도시지역의 토지피복 변화를 따라가지 못하는 경우가 많다. 신속한 갱신 유무는 크게 두 가지 핵심적인 요인의 영향을 받는다. 하나는 분석에 적합한 최신 원격탐사자료를 얼마나 획득할 수 있는가와 관련된 자료 확보 문제이고 다른 요인은 확보된 자료를 얼마나 신속하게 처리 및 분석하여 신뢰할 수 있는 결과를 만들어내는가와 관련된 자료생성 문제이다. 최근 들어 KOMSAT-3A, GeoEye, QuickBird 등 다양한 센서를 통해 획득된 고해상도 위성영상이 공급됨에 따라 영상 확보는 용이해지고 있는 반면 고해상도 영상을 효과적으로 분석할 수 있는 자동화 기술은 부족한 실정이다.

최근 들어 훈련자료를 수집하는 과정에서 GIS 정보를 이용하여 자동화하는 방법에 대한 연구(조기환, 정종철 2018; Attarzadeh and Momeni 2012; Huang, Weng, Lu and Feng et al. 2015; Zhang, Chen and Qiu 2016)가 시도되고 있지만 아직 초보 단계이다.

영상분류 자동화는 지속적으로 수집되고 있는 고해상도 위성영상정보를 체계적으로 관리하고 효율적으로 활용하는 데 필수적인 과정이다. 그러나 영상분류 자동화를 위한 신뢰할 수 있는 방법은 아직 제시되지 못하고 있으며 연구와 이론적 토대 역시 부족한 실정이다. 특히, 갱신 수요가 많은, 토지피복이 급속히 변화하는 지역을 대상으로 과거 토지피복정보를 활용하여 훈련자료를 자동으로 추출하는 연구는 드물다. 따라서 본 논문은 과거에 생성된 공간정보를 활용하여 효율적으로 훈련자료를 추출하고 이를 활용하여 영상분류를 효율적으로 수행할 수 있는 방법론을 제시하는 데 연구의 목적이 있다.

## II. 연구 방법

### 1. 선행연구 검토

원격탐사를 통해 얻어지는 영상을 활용한 분류기법에 관한 최근의 연구로 Acharya, Yang and Lee(2016)이 고해상도 위성영상인 KOMPSAT-3A를 활용하여 전라북도 부안 지역을 대상으로 Mahalanobis Distance (MahD), 최소거리, 최대우도, Support Vector Machine (SVM) 4가지 감독분류 기법에 대해 검토한 연구가 있다. 분류 결과 MahD와 SVM이 90% 내외의 분류 정확도를 보였으며, NDVI[2]와 NDWI[3]와 같은 정규지수를 추가할 경우 정확도가 개선되는 경향을 보였다.

Huang, Weng, Lu and Feng et al.(2015)은 World View-2, GeoEye-1 등의 고해상도 영상을 이용하여 도시지역의 토지피복을 분류하였다. 이 연구에서는 OpenStreetMap(OSM) 자료를 이용하여 초기 관심영역(Region of Interest: ROI)으로 설정하였으며 서로 다른 분류의 경계나 중첩된 부분은 모두 제외하는 방식으로 관심영역을 선택하였다. 또한, NDVI, NDWI와 같은 정규지수를 사용하여 건물, 그림자, 수역, 식생, 나대지 총 5가지로 분류하였으며 전체 정확도는 80% 초반으로 나타났다.

Attarzadeh and Momeni(2012)는 Quickbird 위성영상을 활용하여 건물을 추출하기 위해 객체기반분류기법을 사용하였다. Green 밴드, 적외선, NDVI, 객체의

1) https://egis.me.go.kr (2019년 5월 21일 검색).

2) 정규식생지수(NDVI: Normalized Difference Vegetation Index)는 다음과 같이 계산됨: (NIR_reflectance − Red_reflectance) / (NIR_reflectance + Red_reflectance)

3) 정규수분지수(NDWI: Normalized Difference Water Index)는 다음과 같이 계산됨: (NIR_reflectance − SWIR_reflectance) / (NIR_reflectance + SWIR_reflectance)

길이, 면적, 직사각형 지수 등을 사용하였으며 80% 이상의 분류정확도를 나타냈다.

Zhang, Chen and Qiu(2016)는 고해상도 원격 영상에서 효과적인 관심 영역을 추출하기 위해 영상의 구조, 가장자리, 형태, 방향, 공간정보를 담아내는 기법을 사용하였다. 영상의 방사값을 히스토그램 기법을 통해 나타낸 정보와 가중 융합하여 정확한 관심 영역을 추출하였다.

조기환, 정종철(2018)은 토지피복지도의 폴리곤을 유형별로 분류하고 각 폴리곤 내부 픽셀들의 영상자료 특성(반사율 평균, 표준편차 등)을 고려하여 피복유형을 대표할 수 있는 폴리곤을 선별하여 관심영역으로 활용하는 방법을 제시하였다. 이를 이용하여 도시지역을 대상으로 Senitnel-2A 영상을 감독 분류한 결과 전체 분류정확도는 80~87%로 나타났다.

Radoux and Defourny(2010)는 ITA(Iterative Trimming Algorithm)을 적용하여 훈련자료를 추출하고 이를 이용하여 기존의 식생지도 중 변화가 발생한 지점을 찾는 연구를 수행하였다. ITA는 이전에 생성된 GIS (Geographic Information System)자료를 이용하여 훈련자료를 수집하지만 지도가 가진 정보의 불확실성과 지도 제작 이후의 변화 등으로 인해 훈련자료 중 일부가 기존 지도상의 유형과 일치하지 않을 가능성을 고려하였다. 분석을 위해 먼저 유형별 훈련자료를 과거 지도를 통해 수집하고 Kernel Density 분석을 실시하여 유형별 확률분포를 추정하였다. 이렇게 추정된 확률분포 중 일정 수준의 이상치(Outlier)를 제거함으로써 실제 토지피복 유형을 더 잘 대표할 수 있는 훈련자료를 추출하여 영상을 분석하였다.

## 2. 이론적 배경

수치지도와 토지피복지도와 같은 공간정보를 활용하여 최신 영상을 분석하기 위한 훈련자료를 자동으로 추출하고자 하는 경우 두 가지 모순되는 조건을 고려할 필요가 있다. 먼저 수치지도와 토지피복지도와 같은 공간정보가 현장 상황을 제대로 반영하지 못한다고 볼 수 있을 정도로 변화가 클 필요가 있다. 그렇지 않다면 영상분류를 통해 새로운 공간정보를 생성할 필요가 없기 때문이다. 동시에 과거 공간정보가 현재 상황을 부분적으로나마 충실히 반영하고 있어야 한다. 이전 영상정보가 현재의 현장 상황을 더 이상 반영하고 있지 못한다면 이전 정보를 이용하여 훈련지역을 추출하는 시도는 성과를 낼 수 없기 때문이다.

따라서 수치지도와 토지피복지도와 같은 공간정보를 이용하여 자동으로 훈련자료를 추출하는 알고리즘의 핵심요소는 과거의 공간자료 중 현재 상황을 충실히 반영하는 부분과 그렇지 못한 부분을 자동으로 구분하는 것이다. 토지피복지도에 숲으로 구분되어 있는 영역 중에는 현재에도 숲으로 존재하는 경우도 있겠지만 일부는 시가지나 나대지로 바뀌어 있는 영역도 있을 것이다. 과거 자료를 이용하여 자동으로 훈련자료를 추출한다는 의미는 과거의 숲 폴리곤 중 다른 유형으로 변한 폴리곤을 제외하고 현재에도 숲으로 남아있는 폴리곤을 자동으로 선별하여 이들 영역의 자료만을 훈련자료로 사용한다는 것을 의미한다.

Radoux and Defourny(2010)는 식생정보를 갱신하기 위해 ITA를 이용하여 식생지도에 나타난 식생 유형별 영상정보의 분포를 분석하고 이를 토대로 이상치를 제거하는 방법을 적용하였으며 결과적으로 식생이 변화한 부분을 효과적으로 구별해 낼 수 있었다. ITA는 자료분포를 유연하게 반영할 수 있어 토지유형별로 특징적인 분광 특성을 가지는 토지피복분류에 적용하기 적합하고 수학적으로 잘 확립된 Kernel Density를 이용함으로써 자료 분석의 신뢰성을 확보할 수 있다는 장점이 있다. 본 논문에서는 식생분류에

적용되었던 ITA를 토지피복분류에 적용하여 자동으로 훈련자료를 추출하고 그 신뢰성을 평가하였다.

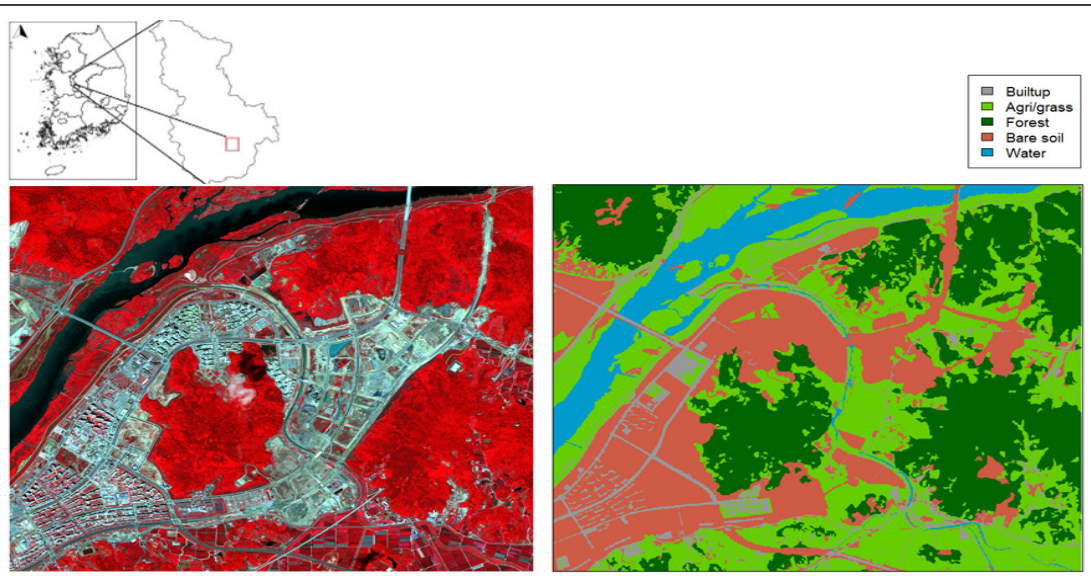
## 3. 연구 대상지 및 자료

연구 대상지역은 세종시 남서쪽 지역(행정중심 복합도시 3-3, 4-1, 4-2 생활권 일대 18.7km$^2$)으로 현재 도시개발이 활발히 진행 중이다(<Figure 1> 참조). 토지피복지도와 영상의 시차가 5년여 차이임에도 불구하고 토지피복지도에는 나대지나 숲으로 구분되어 있는 지역에 건물과 도로가 건설된 곳이 많이 나타나고 있는 것을 확인할 수 있었으며 숲이나 농경지가 나대지로 바뀐 지역도 관찰할 수 있었다. 연구대상지역의 경우 토지피복이 변화하여 과거에 제작된 GIS정보와 현재의 토지피복 유형이 일치하지 않는 지점이 많은 것을 확인할 수 있었다.

본 논문에서 사용된 영상과 공간정보 데이터는 2015년 세종시 세분류 토지피복지도(환경공간정보서비스)[4]와 2018년 7월 7일에 촬영된 KOMPSAT-3A호 위성영상이다(<Figure 1> 참조).

연구대상지 중간에 큰 구름이 있어 영상정보의 불확실성이 존재한다. 영상이 촬영된 시기는 태양고도가 가장 높은 하지와 가까운 시기로 그림자의 크기가 상대적으로 작아 영상분석이 용이한 장점이 있다. KOMPSAT-3A 위성은 태양동기위성궤도(Sun Synchronous Orbit)를 528km 상공에서 돌고 있다. 영상의 다중밴드(<Table 1> 참조) 반사율을 계산하고 이 값을 주성분분석(Principle Component Analysis: PCA)을 통해 주성분 축(Principal Component Axes)으로 변환하였다. 가시광선 영역의 적색, 녹색, 청색 밴드 자료가 높은 상관성을 가지기 때문에 각각의 밴드 정보를 사용하기보다 자료의 변이를 최대로 나타낼

**Figure 1_** Study Area in Sejong(Top)




(a) KOMPSAT-3A False Color (NIR-Red-Green) Image in July 7, 2018

(b) Land Cover Map of Sejong in 2015

4) https://egis.me.go.kr (2019년 5월 21일 검색).

**Table 1_** Characteristics of KOMSAT-3A Image

| Bands | Wavelength(㎛) | Resolution |
|---|---|---|
| Pan | 0.45-0.90 | 0.55m×0.55m |
| Blue | 0.45-0.52 | 2.2m×2.2m |
| Green | 0.52-0.60 | 2.2m×2.2m |
| Red | 0.63-0.69 | 2.2m×2.2m |
| NIR | 0.76-0.90 | 2.2m×2.2m |

수 있는 주성분 축 값으로 변환하여 사용하는 것이 더 효율적이다.

## 4. 연구 방법

토지피복지도는 1m급 고해상도 영상을 사용하여 정밀하게 제작되므로(환경공간정보서비스)[5] 토지피복 현황을 잘 반영할 수 있으나, 본 연구지역과 같이 토지피복이 빠르게 변화하는 지역에서는 토지피복정보 갱신이 변화를 충분히 반영하지 못하는 경우가 발생한다. 이와 같은 문제를 고려하여 본 논문에서는 정보의 불확실성에 유연하게 대처할 수 있는 ITA를 적용하여 훈련자료를 자동추출하고 이를 이용하여 토지피복유형을 분류하였다.

ITA는 다음과 같이 방식으로 진행된다. 1) 기존 GIS자료를 이용하여 유형별 영상정보를 수집한다. 2) 커널밀도추정(Kernel Density Estimation: KDE)을 이용하여 대상영상의 확률분포를 추정한다. 3) 확률이 낮은 자료는 GIS자료에 표시된 유형과 다른 것으로 간주하여 특정한 비율(alpha)로 이상치를 제거한다. 4) 이상치를 제거 후 자료를 훈련자료로 활용한다. KDE는 아래식과 같이 추정하였다.

$$\hat{f}(x) = \frac{1}{nb}\sum_{j=1}^{n} K(\frac{x - x_j}{b}) \qquad <식\ 1>$$

이때, n은 관측자료 수, $x_j$는 관측자료, b는 대역폭(Bandwidth), K는 추정값이 확률밀도함수가 되게 하는 커널함수이다. 본 논문에서는 정규커널(Normal Kernel)을 사용하였다. 대역폭은 KDE(Kernel Density Estimation)의 형태를 결정하는 파라미터이다. 대역폭($\hat{b}$)은 자료의 추정표준편차($\hat{\sigma}$)와 사분위수 범위(Interquartile Range, R)를 1.34로 나눈 값 중, 작은 값을 취하여 다음 식과 같이 추정하였다(Vernables and Ripley 2002).

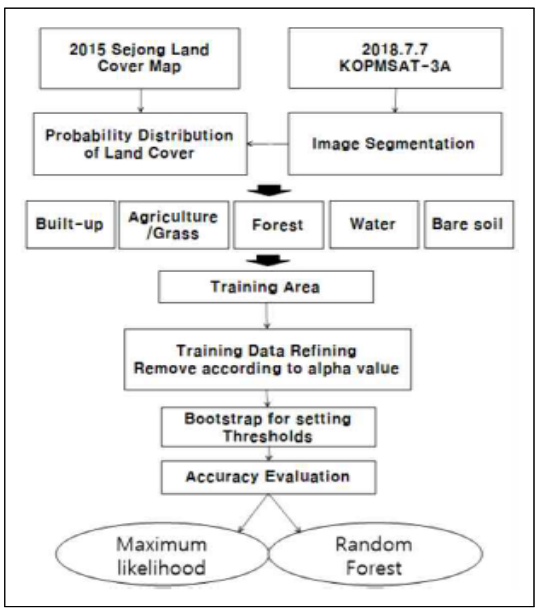
$$\hat{b} = 0.9\min(\hat{\sigma}, \frac{R}{1.34})n^{-1/5} \qquad <식\ 2>$$

ITA는 KDE를 이용하여 확률분포를 추정하므로 분포에 대한 특정한 가정이 없어 적용이 용이하고 자료 특성에 따라 이상치를 제거하는 비율을 다양하게 적용할 수 있다. 이와 같은 유연성으로 인해 GIS자료 제작 이후 발생한 변화뿐 아니라 자료가 가지고 있는 불확실한 정보(잘못 입력된 속성정보, 좌표오류, 구름 등)도 이상치로 처리하여 제거함으로써 훈련자료를 효과적으로 선택할 수 있는 장점이 있다.

본 논문의 흐름은 <Figure 2>와 같으며 다음과 같은 순서로 자료를 분석하였다. 1) 먼저 유사한 분광특성을 가지고 있는 지점들을 객체로 묶기 위해 SAGA GIS에서 자체적으로 제공하고 있는 기능인 객체기반 영상분할기능을 적용하였다. 객체분류는 시행착오 접근법을 적용하여 객체의 군집수를 바꿔가면서 테스트 한후 현장 특성을 가장 잘 반영하는 결과를 선택하였다. 2) 각 객체마다 토지피복유형별 면적 비율을 계산하여 가장 많은 유형의 면적이 80% 이상 차지하는 객체를 선택하여 이후 분석에 사용하였다. 3) 유형별 확률분포를 KDE를 통해 추정하였다. 4) 4가지

5) https://egis.me.go.kr (2019년 5월 21일 검색).

**Figure 2_** Research Flow Chart




alpha(0.04, 0.08, 0.12, 0.16) 값을 적용하여 이상치를 제거한 후 유형별 훈련자료를 확보하였다. 5) Ground Truth Data(GT)를 통해 정확도를 평가하고 가장 높은 정확도를 보이는 alpha 값을 선정하였다.

정확도 평가를 위해 총 400개 지점을 무작위로 선정하여 구글 어스(Google Earth)와 다음 지도 및 KOMPSAT-3A 전정색 밴드 영상을 이용, 각 지점의 토지피복유형을 확인하는 방식으로 GT를 생성하였다. 이 중 9개는 구름 및 그림자 지역에 생성되어 제거하고 391개의 지점을 분석에 활용하였다. 분석에 앞서 이 자료를 두 데이터세트, 즉 250개 지점과 141개 지점 자료세트로 무작위로 나누었다. 250개 자료는 토지피복유형별 alpha 값을 결정하는 데 사용하였고 이 alpha 값들을 적용하여 이상치를 제거한 후 유형별 최종 확률분포를 추정하였다. 이 확률에 근거하여 토지피복분류를 수행하고 141개 지점 자료로 분류정확도를 평가하였다. 분류정확도는 GT와 분류결과를 오차행렬로 만들어 분석하였다(Congalton 1991). 또한 자료 분포에 의해 우연히 일치할 수 있는 경우를 제외한 영상분류 정확도를 Kappa 계수($\hat{K}$)를 이용하여 분석하였다(Cohen 1960).

이상치 제거비율(alpha) 결정은 훈련자료 선택 시 발생할 수 있는 오차를 줄이기 위해 분류된 객체 내의 토지피복 유형 비율을 분석하여 한 유형이 80% 이상 우점하는 객체들만을 선택하였다. 선택된 객체들이 가지고 있는 영상정보(주성분 1, 2축, PC1, PC2)를 바탕으로 2차원 KDE를 통해 유형별 최초 확률분포를 추정하였다. 최초 확률분포 중 확률값이 낮은 지점들을 토지피복이 변한 지점(혹은 토지피복유형을 대표하지 못하는 지점)으로 간주하여 alpha 값 비율대로 제거한 자료를 훈련자료로 사용하였다. 예를 들어 alpha 값이 0.04라면 원 자료 중 확률값이 낮은 순으로 4%까지를 이상치로 간주하여 제거한 경우이고 alpha 값이 0.16이라면 최저 16%까지 자료를 이상치로 간주하여 제거한 것이다. alpha 값에 따라 이상치를 제거한 훈련자료 세트로 총 20개(5가지 토지피복× 4가지 alpha)를 생성하였다. alpha 값과 관련된 사전정보가 없기 때문에 가능한 모든 조합((4단계의 alpha)^(5가지 유형)=1024)을 적용하여 분류를 수행하였다. 정확도 평가는 부트스트랩 방법을 적용하여 250개의 지점 중 200개 지점을 무작위로 선정하여 정확도를 계산하는 것을 1,000회 반복하였다. 각 반복에서 1024개의 alpha 조합을 테스트하였다. 1024개의 조합 중 가장 높은 정확도를 보인 조합(들)을 기록하였다. 1,000회 반복 후 빈도를 분석하여 가장 빈도가 높았던 alpha 값을 최종 제거 비율로 결정하였다.

부트스트랩(Bootstrap)을 통해 산출된 각 토지피복 유형의 alpha 값 비율대로 제거한 자료를 훈련자료로 사용하여 영상분류를 수행하였다. 영상분류는 최대우도법(Maximum Likelihood)과 랜덤포레스트(Random Forest: RF) 두 가지 기법을 사용하였다.

ITA를 통해 선정된 훈련자료를 KDE로 분석하여 확률분포를 추정하였다. 각 GT 지점의 유형별 확률값을 비교하여 최고확률을 가지는 유형을 그 지점의 토지피복유형으로 분류하였다.

랜덤포레스트 기법은 의사결정 기법 중 하나로 사용자가 사전에 정의한 이진 결정 트리 수에 따라 변수들을 분류하는데, 이때 각 변수들에게 할당되는 결정 트리는 랜덤으로 배정되는 형태이다. 모든 이진 결정 트리의 결과들 중 다수결에 따라 최종 예측값으로 결정된다. 랜덤포레스트는 이진 결정 트리들의 결합을 기본으로 하고 있기 때문에 정확하고 빠른 훈련 속도를 가지며 많은 양의 데이터를 실행시키는데 탁월한 기법으로 알려져 있다(Ko, Lee and Nam 2012).

랜덤포레스트 기법 적용 시 유형별 훈련자료의 개수를 맞추기 위해 훈련자료가 가장 적은 수역 자료 개수를 참고하여 유형별로 10,000개 지점을 무작위로 선택하여 분류를 수행하였다.

RF모델에 사용된 변수는 영상의 밴드별 반사율 값을 서로 상관성이 없는 가상축인 주성분(PC1, PC2, PC3, PC4)으로 변환시킨 자료와 객체의 군집 ID 값(1~20), 객체의 면적, 객체의 둘레, 객체의 면적 대 둘레 비율(<Table 2> 참조) 등이었다. 이들 변수를 사용하여 총 6가지 RF모델을 적용하였다.

**Table 2_** Used Information for the Random Forest Models

| Variable | RF1 | RF2 | RF3 | RF4 | RF5 | RF6 |
|---|---|---|---|---|---|---|
| PC1 | O | O | O | O | O | O |
| PC2 | O | O | O | O | O | O |
| PC3 | | | | O | O | O |
| PC4 | | | | O | O | O |
| Segment cluster ID | | O | O | | O | O |
| Area | | | O | | | O |
| Perimeter | | | O | | | O |
| Area/perimeter | | | O | | | O |

### 5. 소프트웨어

자료분석은 R Software(R Core Team 2017)환경에서 수행되었다. KDE에는 'MASS'(ver 7.3-51.1) 패키지(Venables and Ripley 2002)가 사용되었으며 랜덤포레스트를 이용한 영상 분류에는 'RandomForest'(version 4.6-12) 패키지(Liaw and Wiener 2002)가 사용되었다.

## III. 연구결과

### 1. GIS 자료를 이용한 훈련자료 추출

본 논문에서는 영상의 분광특성들을 특징적으로 나타내기 위해 영상 화소들을 비슷한 특성을 지닌 객체들로 분할하였다. <Figure 3>은 객체기반영상분할 매개변수 중 객체의 군집수에 따라 분할된 결과를 나타낸 것이다. 본 논문에서는 객체 군집수를 15개부터 60개까지 바꿔가며 테스트한 결과 군집수가 20일 때 현장특성을 가장 적절하게 반영하고 있다고 판단되어 이를 분석에 이용하였다(<Figure 3> 참조).

<Figure 4>는 주성분 분석을 통해 도출된 4개의 성분 중 설명력이 가장 높은 2개의 성분 축 값을 KDE를 이용하여 추정한 토지피복유형별 확률분포도이다. 가로축에 해당하는 PC1은 주성분 분석 결과 1성분으로 나타난 값이고 PC2는 2성분으로 나타난 값이며, 각 분류의 밀집도에 따라 등고선으로 나타내었다.

농경지/초지 자료의 일부가 나대지 확률이 높은 지점에 위치하고 있는 것을 확인할 수 있다. 이는 토지피복지도상의 농경지/초지가 나대지로 변화한 지역일 가능성이 크다. 한편, 나대지의 상당 부분이 농경지/초지와 겹치는 양상도 나타나는데 이는 토지피복지도상의 나대지 중 일부 지역은 공사가 진행되지 않고 나대지 상태로 유지되면서 초지가 형성되었기 때문으

**Figure 3 _** Image Segmentation with Different Number of Cluster

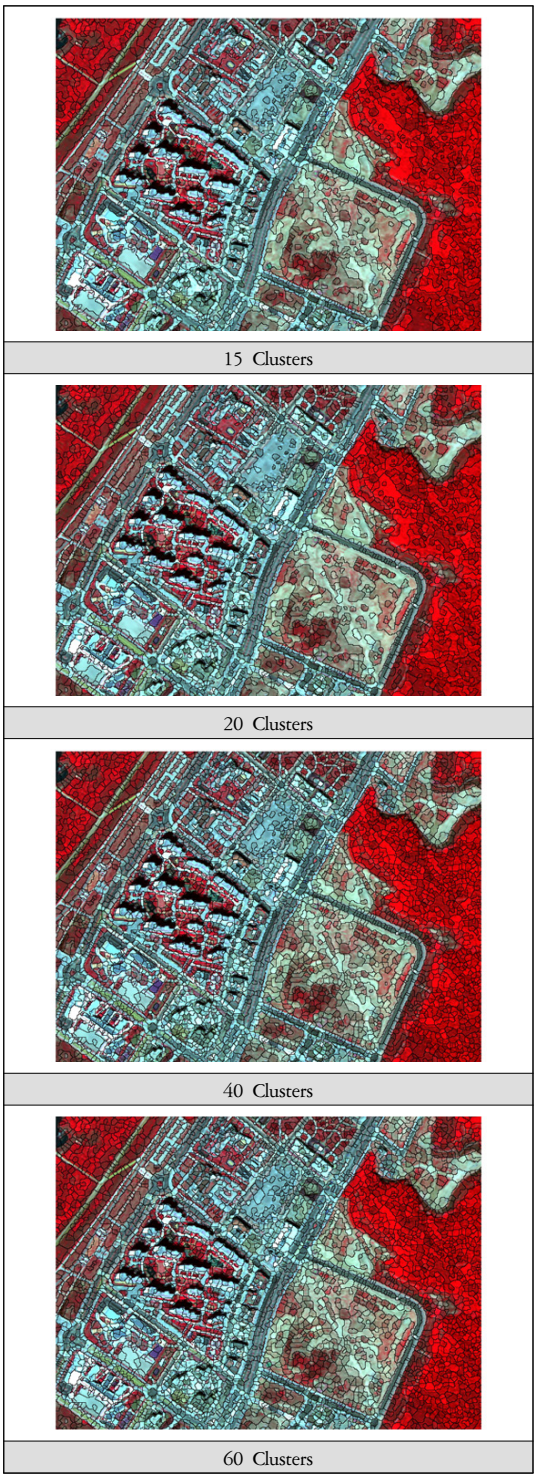



**Figure 4 _** Estimated Probability Distribution of Each Land Cover Class

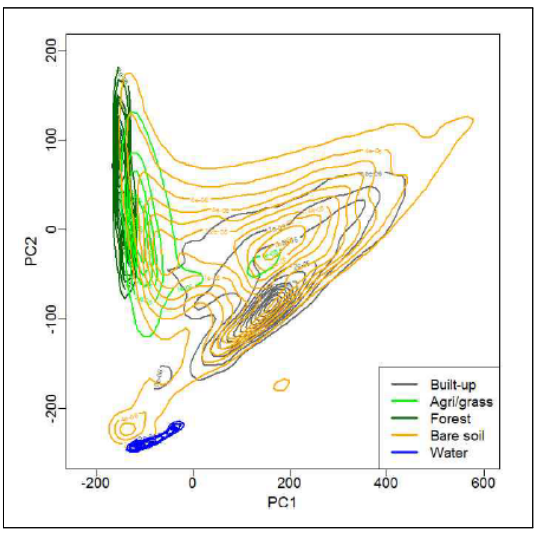
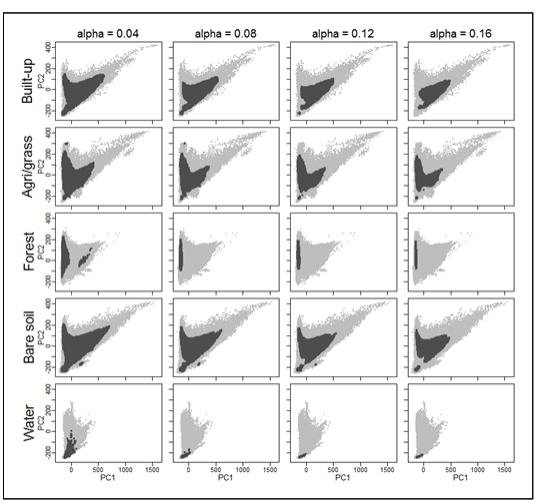


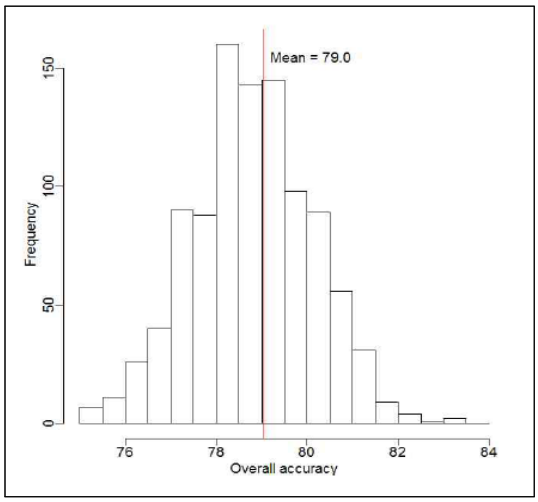

**Figure 5 _** Distribution of Trimmed Data(Black Dots) on All Data(Grey Dot) of Each Land Cover Class




로 판단된다. <Figure 5>는 토지피복유형별 원 자료의 분포(회색)와 alpha 값 비율로 이상치를 제거한 이후 자료 분포를 나타낸다.

훈련자료 추출에 가장 적합한 이상치 제거비율(alpha) 결정방법을 적용한 결과, 전체 정확도 히스토그램은 <Figure 6>과 같았다. 250개의 GT에서 200

**Figure 6 _** Overall Accuracy Histogram of Bootstrap




개를 무작위로 뽑아 정확도를 분석하는 과정을 1,000회 반복한 결과 전체 정확도의 평균은 79.0%로 나타났다. 각 반복에서 가장 높은 정확도를 나타내었을 때의 alpha 값들의 빈도를 분석한 결과 시가화지역은 0.08, 농경지/초지는 0.16, 산림은 0.04, 나대지는 0.16, 수역은 0.04인 경우가 가장 빈번하게 관찰되었다. 시가화지역, 산림 및 수역은 비교적 적은 수의 이상치를 제거(4~8%)하는 경우 좋은 결과를 산출한 반면 농경지/초지 및 나대지는 상대적으로 많은(16%) 이상치를 제거할 필요가 있었다. 산림 및 수역은 비교적 분광특성이 뚜렷한 공통점이 있었다. 이 경우 너무 많은 이상치를 제거하면 훈련자료의 분포가 매우 좁은 영역으로 제한되고 이에 따라 조금만 분광특성이 달라도 확률값이 급격하게 떨어지게 되어 다른 유형으로 오분류되는 결과를 가져오게 된다. 즉, 훈련자료는 특징적인 분광특성을 가져야 하지만 너무 과한 경우 오히려 정확도를 떨어뜨리는 원인이 될 수 있다. 시가화지역은 토지피복도상의 토지유형이 다른 유형으로 변화한 경우가 상대적으로 적은 특징을 보였다. 이는 초지나 나대지가 시가화지역으로 변하는 경우는 많지만 시가화지역에서 다른 유형으로 변화되는 경우는 드물기 때문이다. 따라서 상대적으로 적은 이상치를 제거해도 결과에 충분히 반영될 수 있었다고 사료된다.

연구대상지의 경우 초지나 나대지는 도시개발의 중간 단계인 경우가 많다. 이는 토지피복 유형이 짧은 시간 안에 변화할 가능성이 큰 것을 의미하며 최신 영상의 토지피복유형과 불일치하는 경우가 많다. 본 논문의 결과에서도 이들 유형이 가장 많은 이상치를 제거해야 분류 정확도가 높아지는 것으로 나타났다.

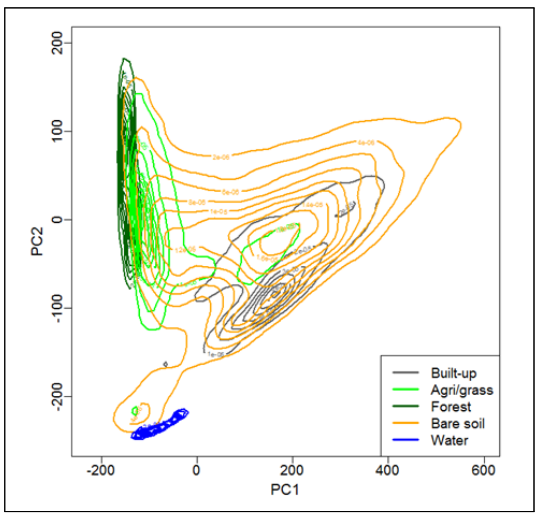
Rodoux and Defourny(2010)는 숲의 변화를 탐지하기 위해 ITA를 이용하였다. 이 연구에서는 alpha는 0.10 이하 값이 사용되었다. 또한 Desclée, Bogaert and Defourny(2006)은 alpha 값으로 0.05 이하를 사용하여 숲의 변화를 탐지할 수 있었다. 위 사례들은 전체 면적에서 변화한 부분이 적은 지역을 대상으로 한 연구였기 때문에 상대적으로 낮은 alpha 값이 필요했던 것으로 판단된다. 하지만 세종시의 경우 GIS자료와 일치하지 않는 면적이 넓었고 구름과 그림자 등으로 가려진 부분이 존재하는 등 정보의 불확실성이 컸기 때문에 alpha 값이 높게 나온 것으로 판단된다.

부트스트랩 기법을 통해 결정된 alpha 값을 이용하여 이상치를 제거하고 훈련자료를 추출하였다. 훈련자료 추출 과정에서는 인위적인 디지타이징 작업이 필요하지 않았고 영상과 GIS자료 및 GT 등의 입력자료만으로 훈련자료를 추출할 수 있었다.

## 2. 영상분류 정확도 평가

이상치 제거 후 자료를 이용하여 추정된 확률분포는 <Figure 7>과 같았다. 이를 기반으로 최대우도법 영상분류를 수행하고 정확도를 평가한 결과가 <Table 3>에 나타나 있다. 전체 정확도는 81%, Kappa 계수는 0.76으로 나타났다. 산림과 수역이 가장 높은 정확

**Figure 7_** Estimated Probability Distribution of Each Land Cover Class Using Refined Training Data




도를 보였으며 농경지/초지에서 가장 낮은 정확도를 나타냈으며 시가화지역 및 나대지 역시 정확도가 낮았다. 또한, 본 논문에서는 추출된 훈련자료를 사용하여 RF 모델을 적용하여 영상을 분류하고 정확도를 평가한 결과는 <Table 4~9>에 나타나 있다.

전체 정확도는 모델에 따라 71~78%로 차이가 났다(kappa 계수: 0.63~0.72). 가장 높은 정확도를 보인 RF6은 주성분 정보와 객체 군집 정보, 객체의 형태 정보 모두를 사용한 모형이었다. 분광 정보 외에 객체의 형태와 관련이 있는 객체둘레, 객체의 면적 대 둘레 비율(Area/perimeter) 정보가 추가된 경우 분류정확도가 증가하였다(RF3, RF6). 객체 군집 ID의 경우는 결과에 미치는 영향이 분명하지 않았다. RF2와 RF5에

**Table 3_** Confusion Matrix of Maximum Likelihood Based Prediction with Refined Training Data

| ML | Maximum Likelihood Classification | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ =0.76 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 20 | 1 | 0 | 5 | 0 | 76.9% |
| Agri/Grass | 0 | 21 | 3 | 0 | 0 | 87.5% |
| Forest | 1 | 0 | 41 | 0 | 0 | 97.6% |
| Bare Soil | 7 | 10 | 0 | 18 | 0 | 51.4% |
| Water | 0 | 0 | 0 | 0 | 14 | 100.0% |
| Producer's Accuracy | 71.4% | 65.6% | 93.2% | 78.3% | 100.0% | 80.9% |

**Table 4_** Confusion Matrix of RF Model 1 Based Prediction with Refined Training Data

| RF1 | PC1+PC2 | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ =0.63 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 21 | 1 | 0 | 7 | 0 | 72.4% |
| Agri/Grass | 0 | 16 | 8 | 2 | 0 | 61.5% |
| Forest | 2 | 0 | 35 | 0 | 0 | 94.6% |
| Bare Soil | 5 | 15 | 1 | 14 | 0 | 40.0% |
| Water | 0 | 0 | 0 | 0 | 14 | 100.0% |
| Producer's Accuracy | 75.0% | 50.0% | 79.5% | 60.9% | 100.0% | 70.9% |

**Table 5_** Confusion Matrix of RF Model 2 Based Prediction with Refined Training Data

| RF2 | PC1 + PC2 + Segment Cluster ID | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ =0.67 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 17 | 0 | 0 | 4 | 0 | 81.0% |
| Agri/Grass | 0 | 19 | 4 | 4 | 0 | 70.4% |
| Forest | 0 | 0 | 40 | 0 | 0 | 100.0% |
| Bare Soil | 11 | 13 | 0 | 15 | 0 | 38.5% |
| Water | 0 | 0 | 0 | 0 | 11 | 100.0% |
| Producer's Accuracy | 60.7% | 59.4% | 90.9% | 65.2% | 100.0% | 74.5% |

**Table 6_** Confusion Matrix of RF Model 3 Based Prediction with Refined Training Data

| RF3 | PC1+PC2+Segment Cluster ID+Peri+Area+Area/Peri | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ = 0.69 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 14 | 1 | 0 | 1 | 0 | 77.8% |
| Agri/Grass | 2 | 19 | 5 | 1 | 0 | 60.0% |
| Forest | 0 | 0 | 39 | 0 | 0 | 93.0% |
| Bare Soil | 12 | 12 | 0 | 21 | 0 | 43.8% |
| Water | 0 | 0 | 0 | 0 | 11 | 100.0% |
| Producer's Accuracy | 50.0% | 59.4% | 88.6% | 91.3% | 100.0% | 75.9% |

**Table 7_** Confusion Matrix of RF Model 4 Based Prediction with Refined Training Data

| RF4 | PC1+PC2+PC3+PC4 | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ =0.66 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 21 | 2 | 0 | 4 | 0 | 77.8% |
| Agri/Grass | 1 | 15 | 4 | 5 | 0 | 60.0% |
| Forest | 0 | 3 | 40 | 0 | 0 | 93.0% |
| Bare Soil | 6 | 12 | 0 | 14 | 0 | 43.8% |
| Water | 0 | 0 | 0 | 0 | 11 | 100.0% |
| Producer's Accuracy | 75.0% | 46.9% | 90.9% | 60.9% | 100.0% | 73.8% |

나타나듯이 객체군집 ID가 정확도를 높이는 경우도 있고 그렇지 않은 경우도 있었다(<Table 5, 8> 참조). 유형별로 보면 생산자 정확도는 농경지/초지가 대부분 낮았다. 오분류된 농경지/초지는 대부분 나대지로 분류되었다. 이는 사용자정확도에서 나대지의 분류정확도가 낮게 나타나는 원인이 되었다. 나대지가 아

**Table 8_** Confusion Matrix of RF Model 5 Based Prediction with Refined Training Data

| RF5 | PC1+PC2+PC3+PC4+Segment Cluster ID | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ =0.63 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 17 | 1 | 0 | 2 | 0 | 85.0% |
| Agri/Grass | 1 | 14 | 4 | 6 | 0 | 56.0% |
| Forest | 0 | 3 | 40 | 0 | 0 | 93.0% |
| Bare Soil | 10 | 14 | 0 | 15 | 0 | 38.5% |
| Water | 0 | 0 | 0 | 0 | 14 | 100.0% |
| Producer's Accuracy | 60.7% | 43.8% | 90.9% | 65.22% | 100.0% | 70.9% |

**Table 9_** Confusion Matrix of RF Model 6 Based Prediction with Refined Training Data

| RF6 | PC1+PC2+PC3+PC4+Segment Cluster ID+Peri+Area+Area/Peri | | | | | |
|---|---|---|---|---|---|---|
| $\hat{K}$ =0.72 | Reference | | | | | |
| Prediction | Builtup | Agri/Grass | Forest | Bare Soil | Water | User's Accuracy |
| Builtup | 20 | 1 | 0 | 1 | 0 | 90.9% |
| Agri/Grass | 1 | 18 | 5 | 3 | 0 | 66.7% |
| Forest | 0 | 0 | 39 | 0 | 0 | 100.0% |
| Bare Soil | 7 | 13 | 0 | 19 | 0 | 48.7% |
| Water | 0 | 0 | 0 | 0 | 14 | 100.0% |
| Producer's Accuracy | 71.4% | 56.3% | 88.6% | 82.6% | 100.0% | 78.0% |

닌 유형이 나대지로 오분류되는 경우는 농경지/초지가 가장 많았고 그 다음이 시가화지역이었다. 농경지/초지와 나대지의 경우 6가지 RF 모델에서 모두 평균보다 낮은 수치가 나타나 매개 변수나 관심영역 추출에서 개선이 필요할 것으로 판단된다. 산림과 수역은 모든 RF 모델에서 약 100%에 해당하는 정확도를 나타내고 있다.

RF 모델은 기계학습법의 일종으로 통계적 모형과 달리 자료의 입력과 출력 사이의 자료 분석과정을 추적하기 어려운 특성을 가지고 있으며 경우에 따라서는 일관적이지 않은 결과를 보일 수도 있다. 본 논문의 경우 모델에 따라 정확도의 편차가 크게 나타났지만 RF5를 제외하면 대체로 입력정보가 많은 경우 정확도가 높아지는 경향을 보였다.

토지피복지도정보를 이용하여 Sentinel-2 영상을 분류한 조기환, 정종철(2018) 연구의 경우 한 시기의 영상을 이용하여 분류한 경우 분류 정확도는 80% 초반이었지만 세 시기(4월, 6월, 9월) 영상을 이용한 경우 정확도가 87%까지 높아졌다. 이는 다중시기 영상을 활용함으로써 계절적 변화를 반영하고 정보의 불확실성을 줄일 수 있었기 때문으로 판단된다. 이에 비해 본 논문에서는 한 시기의 영상만을 이용한 한계가 있다. 분류 정확도를 높이고 그림자나 구름의 영향 등과 같은 불확실성을 줄이기 위해서 다중시기 영상에서 훈련자료를 추출하는 방법으로 발전시킬 필요가 있다.

환경부 토지피복자료는 정확도가 뛰어난[6] 장점이 있지만 항공사진을 기본 자료로 이용하여 분석에 많은 시간과 비용이 소모된다. 이는 갱신 주기에도 영향을 미쳐 토지피복이 변한 후에도 상당 기간 최신 정보가 제공되지 못하는 상황이 발생할 수 있다. ITA를 활용할 경우 사람에 의한 작업이 최소화되어 위성영상을 짧은 주기로 분석하는 것이 가능하게 되고 결과적으로 변화가 발생한 지점을 신속하게 탐지하고 자료를 갱신할 수 있게 되어 시간과 자원을 절약하는 동시에 정확한 토지피복 정보를 제공할 수 있게 될 것이다.

본 논문은 ITA를 이용하여 감독분류를 위한 훈련자료를 자동으로 추출하고 이를 이용한 영상분류 정확도를 평가하는 것을 목적으로 하였다. 이를 위해 추가적인 디지타이징 작업 없이 GIS자료를 이용하여 분류정보를 얻고 이를 영상과 비교하여 확률분포를 추정한 후 확률값에 근거하여 이상치를 제거하였다. 이상치 제거비율 결정 역시 부트스트랩을 활용하여 결정함으로써 인위적인 판단 없이 훈련자료를 자동으로 추출할 수 있었다.

## IV. 결론

본 논문에서는 ITA를 이용하여 과거에 생성된 GIS정보를 바탕으로 최신 영상 중 토지피복유형과 일치하는 지점과 그렇지 않은 지점을 자동으로 분류하여 불일치하는 자료를 제거함으로써 훈련자료를 추출하는 방법을 제시하고자 하였다. 이를 위해 KOMPSAT-3A 영상을 유사한 분광특성을 가진 객체로 구분한 후 하나의 토지피복유형이 80% 이상 우점하는 객체를 선별하였다. 이들 자료를 이용하여 유형별 확률분포를 KDE를 이용하여 추정하였다. 각 자료가 가진 확률 추정값에 근거하여 확률이 낮은 자료를 이상치로 간주하여 제거하는 방식으로 훈련자료를 추출하였다. 이상치로 제거하는 비율은 시가화지역의 경우 8%, 농경지/초지는 16%, 산림은 4%, 나대지는 16%, 수역은 4%였다. 이상치를 제거하는 방식으로 결정된 훈련자료를 이용하여 개발이 활발히 진행 중인 세종시의 토지피복을 분류하였다. 최대우도법의 경우 전체 정확도가 80%, 랜덤포레스트의 경우 78%였다. 유형별 분류정확도는 산림과 수역이 높게 나타났고 시가화지역과 농경지/초지는 대체로 낮게 나타났다.

여러 한계에도 불구하고 본 논문을 통해 영상분석에 중요한 요소인 훈련자료를 수집하는 과정을 수동작업 없이 자동으로 수행할 수 있었다. 본 논문은 영상분류 자동화를 위한 중간 단계 연구로 정확도 측면에서 개선의 여지가 많은 것으로 나타났다. 정확도 개선을 위한 추가적인 연구와 알고리즘을 쉽게 사용할 수 있게 하는 인터페이스 개발이 필요하다고 사료된다.

### 참고문헌


1. 조기환, 정종철. 2018. 토지피복 공간정보를 활용한 자동 훈련지역 선택 기법. 지적과 국토정보 48권, 2호: 171-183. Cho Kihwan and Jeong Jongchul. Automatic selection method of ROI(Region of Interest) using land cover spatial data. *Journal of Cadastre & Land InformatiX* 48, no.2: 171-183.
2. 환경공간정보서비스. https://egis.me.go.kr (2019년 5월 21일 검색).


6) 토지피복지도 작성지침에 따르면 세분류 토지피복지도의 경우 최소 분류 기준은 선형 요소의 경우 폭 3m, 면형 요소의 경우 면적 10m×10m로 한다고 규정되어 있음(https://egis.me.go.kr).

Environmental Spatial Information Service. https://egis.me.go.kr (accessed May 21, 2019).

3. Acharya, T. D., Yang Intae and Lee Dongha. 2016. Land cover classification using a KOMPSAT-3A multi-spectral satellite image. *Applied Sciences* 6, no.11: 371.
4. Attarzadeh, R. and Momeni, M. 2012. Object-based building extraction from high resolution satellite imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 39: 57-60.
5. Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, no.1: 37-46.
6. Congalton, R. G. 1991. A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment* 37, no.1: 35-46.
7. Desclee, B., Bogaert, P. and Defourny, P. 2006. Forest change detection by statistical object-based method. Remote Sensing of Environment 102, no.1-2: 1-11.
8. Huang, X., Weng, C., Lu, Q., Feng, T. and Zhang, L. 2015. Automatic labelling and selection of training samples for high-resolution remote sensing image classification over urban areas. *Remote Sensing* 7, no.12: 16024-16044.
9. Ko Byoungchul, Lee Jihyeon and Nam Jae-yeal. 2012. Automatic medical image annotation and keyword-based image retrieval using relevance feedback. *Journal of Digital Imaging* 25, no.4: 454-465.
10. Liaw, A. and Wiener, M. 2002. Classification and regression by randomForest. *R News* 2, no.3: 18-22.
11. R Core Team. 2017. R: A language and environment for statistical computing. https://www.r-project.org/ (accessed May 21, 2019).
12. Radoux, J. and Defourny, P. 2010. Image-to-map conflict detection using iterative trimming. *Photogrammetric Engineering & Remote Sensing* 76, no.2: 173-181.
13. Zhang, L., Chen, J. and Qiu, B. 2016. Region of interest extraction in remote sensing images by saliency analysis with the normal directional lifting wavelet transformation. *Neurocomputing* 179: 186-201.
14. Venables, W. N. and Ripley, B. D. 2002 *Modern Applied Statistics with S*. 4th Ed. New York: Springer.

• 논문 접수일: 2019. 9. 17.
• 심사 시작일: 2019. 9. 26.
• 심사 완료일: 2019. 12. 7.

## 요약

주제어: 훈련자료, 토지피복지도, 부트스트랩, 영상분류, 위성영상

최근까지 영상분류에서 핵심적인 과정이라 할 수 있는 훈련자료 선정이 주로 수작업에 의존하여 특징적인 지점을 선택하는 방식으로 진행되고 있다. 본 논문에서는 과거에 구축된 공간정보를 바탕으로 Iterative Trimming Algorithm을 이용한 자동화된 훈련자료 추출 기법을 제안하고 활용 가능성에 대해 검정하였다. 2015년에 발표된 세종시 세분류 토지피복지도 정보를 토대로 2018년 KOMPSAT-3A 위성영상 분류를 위한 훈련자료를 추출하였으며 이를 이용하여 토지피복 분류를 실시하였다. 이를 위해 토지 유형별 확률분포를 커널밀도추정을 통해 추정하고 확률이 낮은 자료를 이상치로 간주하여 제거하는 방식으로 훈련자료를 선별하였다. 부트스트랩을 통해 산출된 토지유형별 이상치 제거 비율은 토지 유형에 따라 다르게 나타났다. 이상치 제거 비율은 시가화지역의 경우 0.08, 농업/초지는 0.16, 산림은 0.04, 나대지는 0.16, 수역은 0.04일 때 가장 높을 분류 정확도를 보였다. 이상치 제거 후 선택된 훈련자료를 이용하여 토지피복 분류를 실시하고 정확도를 검정하였다. 최대우도법 분류에 대한 정확도 검정 결과 전체정확도는 약 80%로 나타났고 토지유형별 분류 정확도에는 편차가 있었다. 본 논문을 통해 인위적인 자료수집과정 없이 과거에 제작된 GIS 자료를 활용하여 훈련자료를 수집하고 이를 이용하여 고해상도 영상을 분류할 수 있다는 것을 확인하였다.