Evaluating land cover types from Landsat TM using SAGA GIS for vegetation mapping based on ISODATA and K-means clustering

Polina Lemenkova1*

1 Université Libre de Bruxelles, École polytechnique de Bruxelles (Brussels Faculty of Engineering), Laboratory of Image Synthesis and Analysis (LISA). Building L, Campus de Solbosch, Avenue Franklin Roosevelt 50, Brussels 1000, Belgium. ORCID ID: https://orcid.org/0000-0002-5759-1089

*Corresponding author: polina.lemenkova@ulb.be or pauline.lemenkova@gmail.com

Received 31 March 2021; Accepted 9 November 2021

A B S T R A C T

The paper presents the cartographic processing of the Landsat TM image by the two unsupervised classification methods of SAGA GIS: ISODATA and K-means clustering. The approaches were tested and compared for land cover type mapping. Vegetation areas were detected and separated from other land cover types in the study area of southwestern Iceland. The number of clusters was set to ten classes. The processing of the satellite image by SAGA GIS was achieved using Imagery Classification tools in the Geoprocessing menu of SAGA GIS. Unsupervised classification performed effectively in the unlabeled pixels for the land cover types using machine learning in GIS. Following an iterative approach of clustering, the pixels were grouped in each step of the algorithm and the clusters were reassigned as centroids. The paper contributes to the technical development of the application of machine learning in cartography by demonstrating the effectiveness of SAGA GIS in remote sensing data processing applied for vegetation and environmental mapping.

Keywords: SAGA GIS, mapping, vegetation, K-means, ISODATA, clustering, cartography, machine learning.

1. Introduction

Vegetation mapping is one of the most important tools for environmental monitoring. Using remote sensing data processed by GIS is the fastest way that helps land cover types to be visualized and assessed. There are various GIS applications for thematic vegetation mapping (Klauco et al., 2013a, 2013b; Lemenkova, 2011). The specific geologic setting including volcanism in the southwestern part of Iceland (Figure 1) resulted in the development of erosion prone soils and fragile vegetation (Eckert and Engesser, 2013; Eddudottir et al., 2020). Together with climate impact, ice cover change (Blauvelt et al., 2020; Cabedo-Sanz et al., 2016) and overgrazing, this affected Arctic landscapes and land cover (Lehnhart-Barnett and Waldron, 2020).

The goal of this paper is to present the processing of the Landsat TM image covering the study area i.e. Iceland (Figure 2). Landsat TM images are widely used in environmental studies due to the accessibility, repeatability of survey and coverage (Bryant et al. 2002; Lymburner et al. 2013). The aim is to highlight the distribution of various land cover types. Technical approaches include ISODATA i.e. Iterative Self-Organizing Data Analysis (Memarsadeghi et al., 2007) and K-means image classification (Fard et al., 2020; Peña et al., 1999; Zhao et al., 2020; Bottou and Bengio, 1995), which aim at grouping image pixels into classes of similar properties representing land cover types.

Cartographic processing is based on SAGA GIS (Conrad, 2006; Hengl et al., 2009). Clustering methods are widely used in geosciences for grouping data using similarity properties (Davies and Bouldin, 1979; Filippone et al., 2008; Forgy, 1965; Lemenkova,
The methods are based on the unsupervised approaches of pixel-based classification, which implies the machine learning approach in remote sensing data processing. These classification techniques can be used to monitor environmental changes such as land degradation or deforestation.

2. Materials and methods

The pixel-based classification approach involves the determination of a spectral response (that is, a digital number, DN) for each pixel of a satellite image. Using the selected mathematical algorithm (ISODATA or K-means), it automatically groups pixels into a class based on the similarities of their DN values. Both methods of unsupervised classification are referred to as cluster analysis in SAGA GIS. The theoretical background of cluster analysis is based on the principle of data grouping and sorting by mathematical algorithms (Rubin, 1967). The methodology of this work consists in the following workflow.

2.1. Image Destriping

A Landsat TM image was loaded to SAGA GIS and noise corrected using the ‘Destriping’ procedure by the path ‘Geoprocessing > Grid > Filter > Destriping’. The destriping filter removed the straight parallel lines in the raw Landsat TM raster scene by using two low-pass filters elongated in the stripes direction: the one with a pixel unit wide and the one with a striping wavelength wide (Oiomoen, 2000). The difference indicated a striping error, which was removed from the original Landsat TM image.

2.2. Color Band Composites

The next step included loading Landsat TM bands and generating synthetic images based on the available Landsat bands. The menu of the SAGA GIS used for testing various combinations of the image is presented in Figure 3.
The workflow included the creation of a false color composite (Figure 4) and a natural color composite (Figure 5). For the Landsat TM multiband imagery, three bands in R, G, and B were displayed in color combination from various monochrome bands. A true (natural) color image is composed of the RGB combinations. Using near-infrared bands (NIR), more information (e.g., land-water contrast, vegetation) was added. The blue channel was used for a false color composite. The combination presented in Figure 5 shows bright cyan-colored ice and glacier areas, dark (black) colors for water and natural looking landscapes (green for vegetation areas and brown for bare soils). The false color composite in Figure 4 shows ice covered areas as bright red, useful for glacier mapping.

2.3. ISODATA Clustering

Unsupervised classification was performed using the SAGA GIS path: Geoprocessing > Imagery > Classification > Unsupervised > ISODATA Clustering for grids (Figure 6). The selected bands were 3, 4, 5, 7 as input bands. Afterwards, the K-means cluster grid was reclassified to land cover classes using the SAGA GIS path: Grid > Values > Reclassify Grid Values. Finally, the statistics on the land cover classes was visualized.

ISODATA clustering, an unsupervised pixel classification method by SAGA GIS, was used for detecting and mapping the land cover classes of southwestern Iceland. Image bands were selected to be used in the assignment of bands for ISODATA clustering and the number of eventual output classes was defined as 10.

The SAGA GIS solution of ISODATA clustering was used to solve the problem of a large amount of unlabeled pixels for land cover types. Since the training pixels of supervised classification require fieldwork data observations, unsupervised classification performs better for a distance based data analysis.

Understanding the meaning of land cover types behind the pool of pixels on a Landsat TM scene requires a machine learning algorithm that classifies these pixels into groups based on the patterns it finds. The unsupervised learning of ISODATA conducts an iterative process, analyzing pixels without the intervention of a cartographer. The key approach of ISODATA relies on the assumption that each class has a multivariate normal distribution. Therefore, it uses class means and a covariance matrix for each class. In case of complicated landscapes, there can be many variables in vegetation and mixed land cover types, which can make hand-made supervised classification a difficult process. Instead, the machine-learning classifiers ISODATA and K-means, based on the clustering and association of pixels on a Landsat TM scene, are applied to identify land cover classes automatically.
2.4. K-means Clustering

The K-means method is another widely used approach of unsupervised classification. These classes were assigned as 10 machine-defined classes and then reclassified as post-processing as ‘land cover classes’ by reference to ground data (Figure 7). In contrast to supervised classification, both ISODATA and K-means do not require defining ‘training sites’ of the land cover type. This presents a higher level of machine learning in a cartographic workflow. The automation and methods of machine learning are commonly used in geosciences.

The iterative process of image processing by K-means clustering in SAGA GIS leads to an improvement in the associations of pixels into cluster groups by machine learning. The SAGA GIS divides the pixels from the Landsat TM image and groups them into the assigned number of classes. The K-means based classifications were fast and efficient. However, some drawbacks include errors from misclassified pixels caused by cloudiness.

3. Results and discussions

The resulting maps show the unsupervised techniques of ISODATA (Figure 8) and K-means (Figure 9). The results determine 10 classes within the Landsat TM for optimal classification performance. The maps keep spatial resolution and texture in the image. K-means clustering was reclassified (Figure 10) for the land cover classes with the assigned types.

The difference in performance consists in the mathematical approach of the k-means algorithm, which aims at the placement of the centers that minimizes the average squared distance of each point to its nearest center (Likas et al. 2003). ISODATA tries to treat each class in a multivariate normal distribution, and computes class means and a covariance matrix for each class (Figure 11).

The algorithm of the K-means is implemented in the most straightforward manner, assuming that the number of clusters, k, is much smaller than the total number of pixels in a scene (Pollard, 1982) (Figure 12). Hence, the performance of the algorithm is more time-consuming compared to ISODATA. As with ISODATA, the time of algorithm performance is explained by the time required to compute the cluster center nearest to each pixel in an image (Jainand and Dubes, 1988).

The principles of ISODATA classification based on the iterative approach have the following nature: pixels are being grouped in each iteration of the algorithm: pixels in an image are assigned to their...
closest (nearest) cluster centers. Afterwards, cluster centers are reassigned to be the centroid of these associated points.

The next resulting step includes clusters with very few deleted points. Finally, larger clusters satisfying new conditions are split again, and smaller clusters reassigned according to the proximity of pixels are merged (Tou and Gonzalez, 1974). The algorithms continue until the number of iterations reaches the number of defined classes.

**Figure 10.** Reclassified map of K-means clustering

**Figure 11.** K-means cluster analysis and a histogram

**Figure 12.** ISODATA cluster analysis and a histogram

4. Conclusions

The demonstrated advantage of machine learning in cartographic data processing over supervised classification in image analysis consists in the automation of data processing. Due to the geometric complexity of contours and fragmentation of many landscapes, the patterns and associations of land cover types may be easily overlooked by a human eye.

Better accuracy and reliability result from automation, which is part of machine learning. The machine learning applied in cartographic workflow enables GIS to train on datasets before data processing (Lemenkova, 2019a, 2019b; Schenke and Lemenkova, 2008). Machine-learning approaches in geosciences include advanced statistical analysis, numerical data processing and applied programming (Chavez et al. 1982; Lemenkov and Lemenkova, 2021a, 2021b; Clark and Roush, 1984; Klaušo et al., 2014, 2017; Lemenkova, 2019c, 2019d, 2021a; Jensen, 2005).

The comparison between the ISODATA and K-means approaches showed that ISODATA operates more slowly, particularly with several processed bands, while the K-means algorithm is a faster method. Both algorithms are central to studies on Landsat TM image processing and environmental applications (Esche and Franklin, 2002; Lemenkova, 2020b; Chen et al., 2020; Xu and Wunsch, 2005). Both ISODATA and K-means are popular and widely used supervised classification methods (Kanungo et al., 2002; Forgy, 1965; Arya et al., 2004; Murariu et al., 2018) both in general data analysis and in remote sensing applications and can be recommended in further studies.

The presented work revealed that the use of Landsat TM satellite imagery and various approaches of remote sensing data processing also continues previous applications of Landsat TM (Liu et al. 2010; Zhao et al. 2016; Mondal et al., 2020; Zerrouki et al. 2021). Moreover, clustering is an effective method of image processing in environmental mapping and vegetation monitoring using SAGA GIS, producing both cartographic maps and statistical graphs and tables. However, various approaches in classification techniques (both supervised and unsupervised) applied for satellite images may lead to ambiguities and difficulties to accurately interpret the land cover types. Noise, such as atmospheric effects, doubtless or technical stripes, is still a challenge in the Landsat scenes with existing applications on solving these problems (Mitchell et al., 1977; Iikura, 2002; Deng et al. 2016).

For the selected study area, the final map includes the following land cover types: grass, forest, elevated areas, flat areas, ocean, bare ground, ground, water, ice, mountains. A noted discrepancy between the two approaches can be a recommended direction in further studies on Landsat TM image processing in the optimization of cartographic techniques using remote sensing data by SAGA GIS.

An accurate method for mapping vegetation and detecting land cover types aims to harmonize the existing methods of geospatial data processing using the advanced GIS software combined with various mathematical algorithms of machine learning. Compared to the traditional GIS mapping (Săvulescu and Mihai, 2011; Suetova et al., 2005; Anny et al. 2014; Vilček and Koco, 2018), the machine learning approach enables more rapid, accurate and precise mapping (Lemenkova, 2020c, 2020d, 2021b, 2021c). An outcome of the machine methods in environmental cartography aims at optimizing the techniques of remote sensing data processing for forest and vegetation monitoring (Zaima et al., 2019). This study contributed to the existing research by presenting methods of machine learning cartographic techniques in environmental agricultural studies.
Acknowledgment

This study has been supported by the project No. 0144-2019-0011, Schmidt Institute of Physics of the Earth, Russian Academy of Sciences.

Declaration of competing interest

The author declares no conflict of interest.

References


Lemenkova, P. (2021a). Scripting Languages for
Lemenkova, P. (2020d). Using R packages ‘tmap’, 'raster' and
Lemenkova P. (2019a). Processing oceanographic data by

edo: 10.1016/S0167-8655(99)00069-0


edo: 10.3846/gac.2019.3785


