How the verification of reference samples affects the land cover classification accuracy.

FP CUP 2024
Abstract
Corresponding Author:
adam.wasniewski@igik.edu.pl

Adam Waśniewski[1], Agata Hościło[1], Linda Aune-Lundberg[2]

[1] Institute of Geodesy and Cartography, Centre of Applied Geomatics, Poland
[2] Norwegian Institute of Bioeconomy Research, Division of Survey and Statistics, Norway

Keywords: reference samples, random forest, land cover classification, digital elevation model, Sentinel-2

Abstract

Up to date maps and reliable information on land cover and land use status is important in many aspects of human activities, such as urban planning, management of natural resources, environmental protection and sustainable development. The open data policies and increasing number of reference datasets available on national geoportals popularized the land cover classifications and results in the increasing number of global land cover products. The selection of appropriate and reliable database, knowledge and understanding of the process of reference data collection is crucial for land cover classification accuracy. Selection of reference data and appropriate pre-processing can be a challenging exercise in the process of preparation of reference samples. Reference samples are quite often selected based on the existing land cover products, which may result in the error propagation. Furthermore, collection of in situ data on the ground is expensive, time-consuming and often difficult in particular over a large area. On the other hand, creating reference samples manually based on, for example, aerial orthophotos is time consuming and subjective. Therefore, it is very important to apply the automated method for assessing of thematic and geometric accuracy of the reference samples. In this study, we examined the impact of the selection of the reference samples for the classification accuracy. The land cover classification was carried out using the Random Forest algorithm based on satellite Sentinel-2 data for the Viken county in Norway. The following ten land cover classes were mapped: sealed surfaces, woodland coniferous, woodland broadleaved, low vegetation, permanent herbaceous, periodically herbaceous, mosses, non- and sparse vegetation, water, snow and ice.

The main aims of this study are i) to examine how the selection of the reference samples may affect the classification accuracy, ii) to derived the best possible land cover classification, iii) to examine whether the use of detailed Digital Elevation Model (DEM) can improve the classification accuracy. Firstly, we performed the classification using the automatically selected reference sampling points derived directly from the national databases. Secondly, we focused on automated approach of creating and filtering reference samples to achieve the most reliable set of reference points. We carried out the detailed analysis of the spectral signatures through the analysis of the histogram for the main land cover classes. Due to the definition of classes in the national databases or methodological issues, some of the sampling points had to be removed. Histograms were analysed especially for woodland broadleaved, woodland coniferous, sealed surfaces, mosses and wetland classes. For these classes obtaining the most reliable set of reference samples seem to be the most difficult without filtering. The difficulty arises from the scale and level of generalization of the reference data. Then, the pre-selected reference sampling points were used to perform the classification. The comparison of both classification results revealed that the accuracy of selected classes has increased after the pre-selection of reference samples. Change of the accuracy of woodland broadleaved and woodland coniferous, after filtration of reference samples the overall accuracy increase from 50.6% to 72.5%, and from 88.4% to 93.2%, respectively. The highest increase by 40.7 percentage points (from 33,3% to 74,0%) was achieved for non- and sparse vegetation class. This shows the significance of control of automatically created samples.

The set of pre-selected reference samples was used to classify the land cover over the entire study area. Then, we examined, whether including the DEM in the classification process can improve the classification accuracy. Using the DEM improved the user's (UA) and producer's (PA) accuracy for nine out of ten land cover classes. The highest improvement of PA was observed for classes located at

higher altitudes: low vegetation (from 75.9% to 84.7%) and non- and sparse vegetation (from 81.3% to 85.8%).

The research leading to these results has received funding from the Norway Grants 2014-2021 via the Polish National Center for Research and Development - project InCoNaDa "Enhancing the user uptake of Land Cover / Land Use information derived from the integration of Copernicus services and national databases".