Scientific Exploitation of Operational Missions (SEOM)

S2-4Sci Land and Water

Study 3: Classification

DELIVERABLE 1.24.3

Abstract

Issue: 1

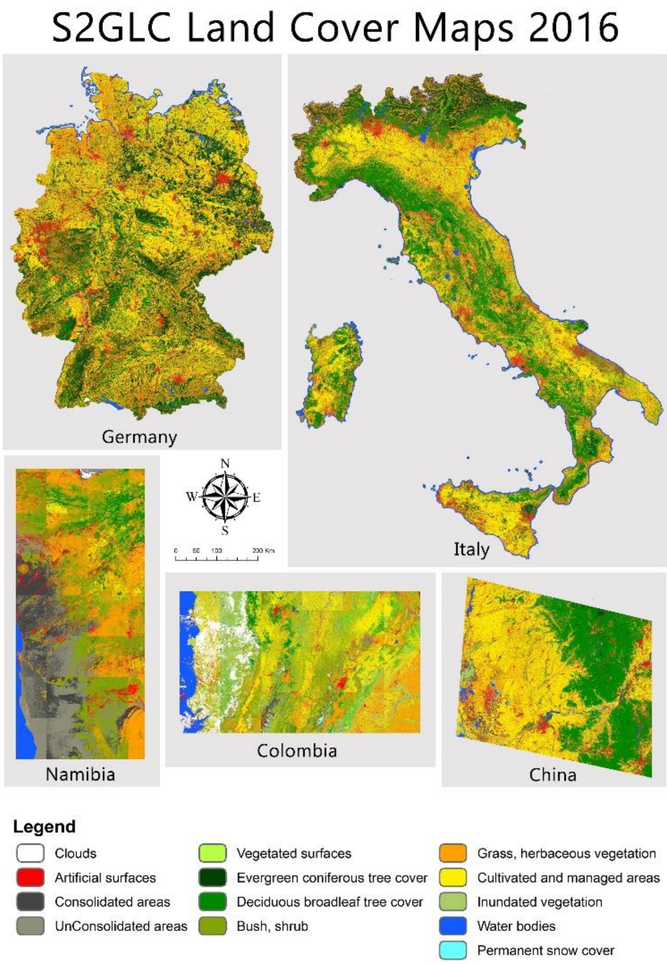
Version: 1

Issue date: 7 March 2018

	Name	Date & Signature
Author(s)	CBK PAN: Radek Malinowski, Artur Nowakowski, Ewa Kukawska, Marcin Rybicki, Michał Krupiński, Stanisław Lewiński	28.02.2018
Input	IABG: Elke Krätzschmar, Sylvia Günther UJ: Denise Dejon, Eric Krüger EOXPLORE: Conrad Bielski	
Reviewer(s)	Stanisław Lewiński	7 March 2018
Approver(s)		

Project Goal

The goal of the ESA SEOM Sentinel-2 Global Land Cover (S2GLC) project was to develop a land cover classification algorithm capable of producing automatically a global fine spatial resolution map based on Copernicus Sentinel-2 Earth Observation imagery. The proposed processing model is fully automated so that the land cover map can be continuously updated and take advantage of the Sentinel-2 temporal and spatial resolution to produce the first global land cover (GLC) map with a resolution of 10 m.



S2GLC Legend

The S2GLC classification legend was defined based on existing legends used within global land cover databases. Land cover classes representing complex structures were not included because of the improved spatial resolution of the Sentinel-2 imagery. The S2GLC map legend identifies the basic land cover classes as well as cultivated areas. Although the agriculture-related class represents land use rather than land cover, it was included due to its importance and existence across all other global land cover databases. The structure of the legend is hierarchical with the third and most detailed level consisting of 15 classes.

Reference Data

Taking into account the spatial resolution of Sentinel-2 data (10 m) and the need to fully automate the classification process, one of the most important and challenging components of the methodology was generating the required training points dataset. For global mapping, the training points can only be selected based on existing databases. Current detailed and reliable open GLC databases are only available at spatial resolutions of 300 m and coarser. The project departed from a common assumption that high accuracy land cover classification results could be obtained solely based on highly precise reference data. However, it was demonstrated that even with the lower uncertainty inherent from the application of coarser reference data, it is still possible to produce high overall land cover accuracy. Therefore, the existing databases were used as the primary source of identifying training reference data for the developed Sentinel-2 classification methodology. Our experiments indicated that the most accurate classification results were obtained when no filtering was applied to the reference data derived from the GLC databases. Additional tests showed that the designed data selection method provides better data than either those collected through innovative data collection strategies such as crowdsourced databases or Sentinel-1 based mask identifying pre-selected land cover classes.

Automated Land Cover Classification

A variety of land cover classification techniques were investigated during the course of the project including object-based and pixel-based, supervised and unsupervised methods. A supervised pixel-based approach was selected because it provided the best results based on classification accuracy, preservation of class details and processing efficiency. Among the supervised methods tested, the Random Forest algorithm outperformed the other classifiers in our tests. The experiments also indicated that the best results were obtained if the classifier training was carried out based on a large number of training samples: between 500 to 1000 pixels per land cover class, per tile. The adopted tile-wise processing approach in part was due to the need for atmospheric correction using the Sen2Cor processor. For each individual Sentinel-2 tile, the potential locations of land cover classes were defined using existing coarser GLC databases. For each class, a thematic mask was defined and then an independent set of randomly designated training points was selected. It was shown that the use of more than one reference land cover database significantly improves the results.

During the course of testing the classification process, it was observed that in many cases the agreement between individual land cover classes from global databases was low. It was also observed that this agreement was correlated to the final classification accuracy. These observations led to the development of a linear model for predicting classification accuracy based only on the agreement of classification results obtained using different global databases as training sources (without any additional validation data). Consequently, for cases where the Sentinel-2 tiles level of the accuracy was predicted to be low, two alternative approaches were developed: (1) the inclusion of local databases and (2) a semi-automatic process for reference data collection with limited human intervention. The following GLC databases were analysed for all testing areas: CCI LC, GlobCover, GLCNMO, MCD12Q1 (MCD). The regional databases available for the analysed test sites (e.g. CORINE LC, Land Cover Classification for Africa by IIASA, Coberturas de la Tierra) were also tested and used as reference data in the final classification approach.

Dedicated functions were applied to analyse the distribution of pixel values in the images from a time series over the analysed period of time. The attempts that were made showed that this type of approach was not the optimal solution due to cloud cover differences across images in a time series. Therefore, the final solution works on each image separately within a time series based on a different set of training data originating from different databases. In the final stage of the classification process, all the results of individually classified Sentinel-2 tiles are aggregated to produce the final output. The proposed aggregation method considers both information on the frequency of a given thematic class occurrence and the value of the prediction score resulting from Random Forest classification. The classification completed over the testing areas confirms the validity of the adopted assumption. Positive final results are obtained even when the image time series is composed of numerous cloudy scenes.

Within the context of the S2GLC project, an additional semi-automatic approach was proposed that could be applied when the results obtained from the automatic method do not produce satisfactory results. Only three to five seed training points per land cover class need to be identified by an operator which minimises his workload. Based on these few seed points, masks with thousands of pixels are automatically generated for selected land cover classes as the training dataset. From these masks the final training samples are randomly selected for the Random Forest classification. Compared to traditional procedures, the proposed solution can significantly reduce the amount of manual intervention. The main automated classification and the class aggregation are then performed in exactly the same manner as described above.

Classification Post-processing

The proposed workflow includes a post-processing procedure to minimise basic classification errors. The initial classification is changed only in the case of pixels classified with low prediction scores and pixels meeting specifically defined neighbourhood conditions. Our approach differs from the commonly used "salt and pepper" smoothing technique and maintains a high degree of spatial details.

Prototype Test Sites

The developed classification workflow was applied to Sentinel-2 images acquired throughout 2016 for five test countries located across four continents: Germany – 360 000 km^2 and Italy – 300 000 km^2 (Europe), an area of 200 000 km^2 in China (Asia), an area of 200 000 km^2 in Colombia (South America) and area of 220 000 km^2 in Namibia (Africa). Depending on the test site, between 12 and 13 classes were identified. A summary of the inputs and results is presented in **Błąd! Nieprawidłowy odsyłacz do zakładki: wskazuje na nią samą.** for the test sites.

Table 1. Sentinel-2 images and existing databases used for classification

Prototype Site	Area km^2	Sentinel-2 tiles	Available images (2016)	Classified images per tile	Databases used for training	Overall accuracy
Germany	357 375	56	1956	10	2 global, 1 regional	85.2%
Italy	301 230	63	2182	10	2 global, 1 regional	72.5%
China	200 750	31	551	6-10	3 global	72.0%
Namibia	235 345	32	1228	10	1 global, 1 regional	56.1%
Colombia	211 705	30	846	3-10	2 global, 1 regional	52.5%
Summary:	*1 306 405*	*212*	*6763*			

Land Cover Map Validation

A land cover map validation was performed on an image tile basis. Initial sampling for the map validation followed a probability-based, stratified random sampling approach to identify at least ten representative sites. This covered approximately 1 000 Sentinel-2 pixels per tile and per class. Among this initial sampling, validation points were reviewed and manually selected. The applied land cover map validation approach followed the ISO Standards 19157 and 2859-1 for samplings strategies.

S2GLC Results

The best S2GLC classification results were achieved for the German and Italian test sites: 85.2% and 72.5% overall accuracy (OA) respectively. These results are in-line with both German and Italian CORINE LC 2012 database validations based on LUCAS points (82.8% and 76.0% of OA respectively) even though the S2GLC classification approach is completely different from the one used to produce the CORINE LC database. The accuracy of the results from the Chinese test site was 72% OA and is also considered relatively high accuracy and characterized with a high degree of details. The results obtained for the test sites in Namibia and Columbia produced lower OA at 56.1% and 52.5% respectively. Disagreement between the existing GLC databases in these areas resulted in having to apply lower quality reference data (compared to Europe) and hampered the collection of reliable training samples. This is considered to be the primary reasons for the obtained lower classification accuracies. Additional problems in the case of the Columbian site are related to difficulties in obtaining cloud free images, high elevation differences producing different lighting conditions within the test area and no seasonal changes in vegetation cover during the year. The latter decreases the usefulness of multi-temporal data because of the lack of change in vegetation throughout the year. The lower results obtained for the Namibian site could be explained by the severe drought occurring in this part of Africa in recent years (2014 - 2016). This contributed to abnormal water regime, i.e. drying up of water bodies and changes in vegetation cover compared to the classes found in the coarse GLC databases. Furthermore, in many areas the difficulty was identifying a clear distinction of the class boundaries, e.g. un-consolidated – grasses, grasses – bush and shrubs, bush and shrubs – tree cover. These areas are transitional zones between the classes and represent a mixture of their components. Another identified issue related to the distinction between non-irrigated areas used for agricultural purposes that closely resemble other classes such as shrubs or grassland.

Conclusions

The essential part of the developed S2GLC classification algorithm is the application of existing low resolution GLC databases to automatically generate training points to produce high spatial resolution land cover maps. The proposed S2GLC solution applies the Radom Forest classifier, a supervised pixel-based approach. Automating the entire land cover classification process makes it possible to map the globe through batch processing. Our uniquely applied solution includes the aggregation procedure which combines results of single tile classifications from multiple dates into a final map product.

The entire classification procedure, including pre- and post-processing are deployed as a dedicated software developed by CBK PAN. The quality of the land cover classification results depends on the training data quality and can be improved by access to regional and local reference datasets.

The solutions and tools developed within the S2GLC project, with adjustments required for specific geographical zones, are able to produce a unique and accurate global land cover database.

The Project Consortium

The project consortium included research institutes and companies with experience in global, pan-European and regional land cover classification. The consortium was led by CBK PAN and included IABG GmbH, Friedrich Schiller University Jena, and EOXPLORE UG. The S2GLC project started in February 2016 and finished in February 2018.

Project website: http://s2glc.cbk.waw.pl/