

EGU2020-1153

<https://doi.org/10.5194/egusphere-egu2020-1153>

EGU General Assembly 2020

© Author(s) 2022. This work is distributed under the Creative Commons Attribution 4.0 License.



Synthetic sampling for spatio-temporal land cover mapping with machine learning and the Google Earth Engine in Andalusia, Spain

Laura Bindereif¹, Tobias Rentschler^{1,2}, Martin Bartelheim^{1,3}, Marta Díaz-Zorita Bonilla^{1,3}, Philipp Gries^{1,2}, Thomas Scholten^{1,2}, and Karsten Schmidt^{1,2}

¹SFB 1070 RESOURCECULTURES, University of Tübingen, D-72074 Tübingen, Germany

²Department of Geosciences, Chair of Soil Science and Geomorphology, University of Tübingen, D-72070 Tübingen, Germany

³Institute of Prehistory, Early History and Medieval Archaeology, University of Tübingen, D-72070 Tübingen, Germany

Land cover information plays an essential role for resource development, environmental monitoring and protection. Amongst other natural resources, soils and soil properties are strongly affected by land cover and land cover change, which can lead to soil degradation. Remote sensing techniques are very suitable for spatio-temporal mapping of land cover mapping and change detection. With remote sensing programs vast data archives were established. Machine learning applications provide appropriate algorithms to analyse such amounts of data efficiently and with accurate results. However, machine learning methods require specific sampling techniques and are usually made for balanced datasets with an even training sample frequency. Though, most real-world datasets are imbalanced and methods to reduce the imbalance of datasets with synthetic sampling are required. Synthetic sampling methods increase the number of samples in the minority class and/or decrease the number in the majority class to achieve higher model accuracy. The Synthetic Minority Over-Sampling Technique (SMOTE) is a method to generate synthetic samples and balance the dataset used in many machine learning applications. In the middle Guadalquivir basin, Andalusia, Spain, we used random forests with Landsat images from 1984 to 2018 as covariates to map the land cover change with the Google Earth Engine. The sampling design was based on stratified random sampling according to the CORINE land cover classification of 2012. The land cover classes in our study were arable land, permanent crops (plantations), pastures/grassland, forest and shrub. Artificial surfaces and water bodies were excluded from modelling. However, the number of the 130 training samples was imbalanced. The classes pasture (7 samples) and shrub (13 samples) show a lower number than the other classes (48, 47 and 16 samples). This led to misclassifications and negatively affected the classification accuracy. Therefore, we applied SMOTE to increase the number of samples and the classification accuracy of the model. Preliminary results are promising and show an increase of the classification accuracy, especially the accuracy of the previously underrepresented classes pasture and shrub. This corresponds to the results of studies with other objectives which also see the use of synthetic sampling methods as an improvement for the performance of classification frameworks.