EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

SFB 1070 RESOURCECULTURES

FACULTY OF SCIENCE
DEPARTMENT OF GEOSCIENCES
CHAIR OF SOIL SCIENCE AND
GEOMORPHOLOGY

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

Laura Bindereif[1*], Tobias Rentschler[1,2], Martin Bartelheim[1,3], Marta Díaz-Zorita Bonilla[1,3], Philipp Gries[1,2], Thomas Scholten[1,2], Karsten Schmidt[1,2]

[1] SFB 1070 RESOURCECULTURES, Eberhard Karls University Tübingen, Germany
[2] Chair of Soil Science and Geomorphology, Department of Geosciences, Eberhard Karls University Tübingen, Germany
[3] Institute of Prehistory, Early History and Medieval Archaeology, Eberhard Karls University Tübingen, Germany

1. INTRODUCTION

Land cover information plays an essential role for resource development, environmental monitoring, and protection. Machine learning approaches based on remote sensing data are very suitable for efficient and accurate spatio-temporal mapping of land cover and change detection. However, most real-world datasets are imbalanced and machine learning methods require sampling schemes with roughly even training sample frequency. Hence, methods to reduce the imbalance of datasets are required. Synthetic sampling methods, such as the Synthetic Minority Over-Sampling Technique (SMOTE, Chawla et al., 2004), were proposed to generate synthetic samples and balance the dataset used in many machine learning applications for more reliable model assessment.

2. MATERIALS AND METHODS

Study area

- Agricultural landscape in the Guadalquivir valley in Andalusia, Spain (approximately 1000 km²)

Datasets

- Landsat 8 OLI, Median composite for 2018, created and downloaded via Google Earth Engine, 7 predictor variables (Band 2 to 7 and NDVI)
- 130 ground truth points from field survey, October 2018, with stratified random sampling
- Classes: arable land (1), plantation (2), pasture (3), forest (4) and shrub (5).
  Urban and water areas were excluded.

Workflow (Fig. 1)

- Import and filtering of Landsat images in the Google Earth Engine (GEE), export as median composite for Sept – Oct 2018
- Classification with Random Forests (RF) and repeated 10-fold cross-validation in R as reference
- Apply SMOTE (UBL package), classification with RF with synthetic samples and repeated 10-fold cross-validation in R

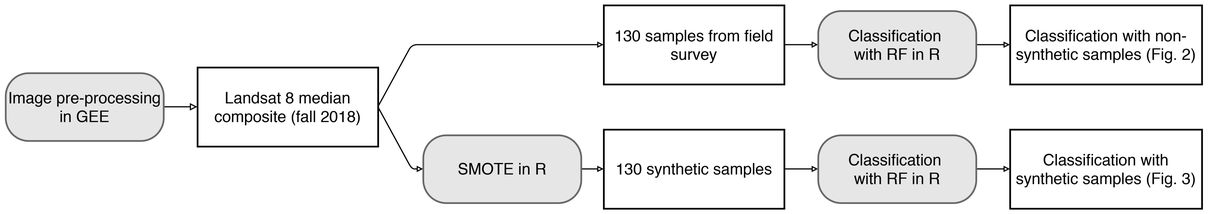

Fig. 1: Workflow of this study using the Google Earth Engine and R

3. RESULTS

The classes 3, 4 and 5 are the minority classes while class 1 and 2 can be considered as the majority classes (Tab.1).

Tab. 1: Precision and recall of the classes with and without SMOTE.

Dataset	Class	1	2	3	4	5	Sum
Original	Number of samples	48	47	13	16	7	131
	Proportion	0.37	0.36	0.1	0.12	0.05	
SMOTE	Number of samples	26	26	26	26	26	130
	Proportion	0.2	0.2	0.2	0.2	0.2	

The SMOTE algorithm produced 26 samples per class. Due to the use of random samples in the SMOTE algorithm, the classification result and the accuracy vary. Therefore, we used 100 iterations and computed the mean values of the accuracy, kappa and confusion matrix.

The overall accuracy of the classification of the original dataset is 0.6 and Cohen's kappa is 0.41. The confusion matrix (Tab. 2) shows that the minority classes have a higher class error rate than the majority classes. The overall accuracy of the SMOTE dataset is 0.63 and kappa is 0.53. Tab. 2 shows that the classes 3, 4 and 5 have a lower class error rate when the SMOTE samples are used. The class error rate of class 1 and 2 increased.

Tab. 2: Confusion matrix of the predictions with field samples (left) and with the SMOTE samples (right)

		Reference					
		1	2	3	4	5	Error rate
Predicted	1	38	9	1	0	0	0.21
	2	14	30	2	1	0	0.36
	3	5	6	1	1	0	0.85
	4	0	4	0	9	3	0.44
	5	0	3	1	3	0	1

		Reference					
		1	2	3	4	5	Error rate
Predicted	1	16	5	4	0	0	0.39
	2	7	9	5	2	4	0.66
	3	2	2	19	1	2	0.26
	4	0	1	2	18	5	0.31
	5	0	2	1	4	19	0.28

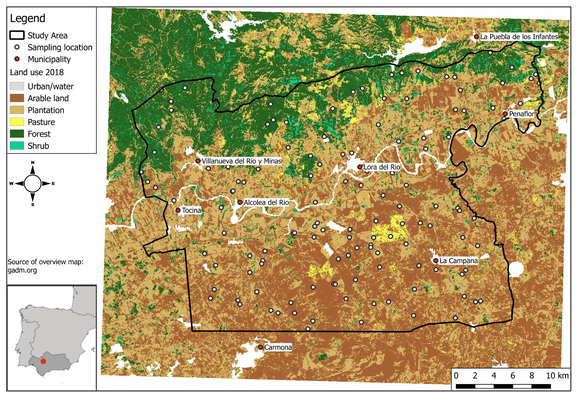

Fig. 2: Classification with field samples and locations of the samples

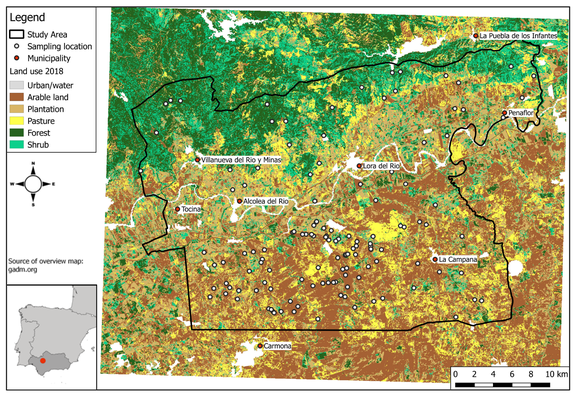

Fig. 3: Classification with SMOTE samples and locations of the samples

4. DISCUSSION AND CONCLUSION

(1) The Google Earth Engine was used for data acquisition and not for classification because the implemented algorithm of RF in Google Earth Engine is less transparent and more difficult to adjust than in R. Further, there is no SMOTE-Algorithm available for the Google Earth Engine yet.

(2) The classification with field samples (Fig. 2) shows an overestimation of the majority classes while the minority classes are underestimated.

(3) Using SMOTE does not provide much higher overall accuracy but the predictions of the minority classes gain in accuracy. The accuracy of the predictions of the majority classes decrease and are less overestimated. This is caused in the fact that most classifiers are built to create a preferably high overall accuracy which can be achieved through the optimization of the classifier for the majority classes.

(4) The spectral signatures of the land cover classes in the study area tend to be similar in autumn and, therefore, are hard to distinguish. Seasonal changes of the spectral signatures are very high in this area wherefore further studies will focus on classification approaches incorporating the seasonal land cover changes.

CONTACT

* Laura Bindereif
laura.bindereif@uni-tuebingen.de

ACKNOWLEDGEMENTS

This study was funded by the German Research Foundation (DFG) through the Collaborative Research Center SFB 1070 RESOURCECULTURES (subprojects Z, S and A02).

FURTHER READING

Bindereif, L. 2019: Analysis and mapping of spatio-temporal land use dynamics in Andalusia, Spain using the Google Earth Engine cloud computing platform and the Landsat archive. Unpublished bachelor's thesis. University of Tübingen, Germany. **Bindereif, L.; Rentschler, T.; Bartelheim, M.; Díaz-Zorita Bonilla, M.; Gries. P.; Scholten, T.; Schmidt, K. 2019**: Analysis and mapping of spatio-temporal land use dynamics in Andalusia, Spain using the Google Earth Engine cloud computing platform and the Landsat archive. Poster, EGU 2019. **Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W. 2002:** SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16 (1), 321-357. **Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. 2017**: Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202: 18-27.