IST 718 – Group project

Land Cover mapping in North Algeria using Sentinel-2 data

A. AMRICHE, S. BANGALORE MUNIBAIRE GOWDA, N. R. SHENOY, Y. ZHOU

THE iSCHOOL Syracuse University

INTRODUCTION

Land Cover and Land Use (LCLU) products are used for a wide range of research and development applications including environmental monitoring, natural resources management, urban planning, and socio-economic studies. Scientists and practitioners use various types of remote sensing data, adopt different approaches, and consequently obtain LCLU products with different mapping accuracies. Current research benefits from technological advances and freely available of remote sensing data, and focuses on the development of advanced classification approaches and machine learning algorithms. As such, this project will focus on fine-tuning and optimizing the Land Cover mapping process through the selection of the most accurate and generalizable method. The expected final classification scheme should be able to adapt and obtain high accuracies in different types of study areas (tropical, humid, arid, desert, etc.). This is most important when processing very large datasets (terabytes of satellite imagery) to generate land cover maps at the national, continental, or global scales, where environmental changes are more pronounced.

OBJECTIVES

- Training multiple classifiers for comparison
- Testing multiple configurations for each classifier to select the best performing parameters
- Run the best classifier and parameters using 100 iterations to generate better performance generalization evaluation.

STUDY AREA

The Mitidja plain is located in North Algeria and characterized by it fertile soils and high agricultural value. It was selected for the purpose of this study to cover a critical need for land cover information to support decision makers and sustainable land management in the area.

MATERIALS AND METHODS

The data used in this study consists of Ground Control Points (GCPs), which are used in remote sensing mapping process. This dataset contains 3000 instances that were used in the classification process, i.e., training, validation, and testing. Each point (or instance) has geographic coordinates, a land cover label (such as forest, urban, water, etc.), the reflectance from the spectral bands of the satellite Sentinel-2, as well as a large number of features (2909), generated from the original 13 bands. The additional features include spectral indices and texture features.

Land Cover Labels:

1, Water

2, Bare land

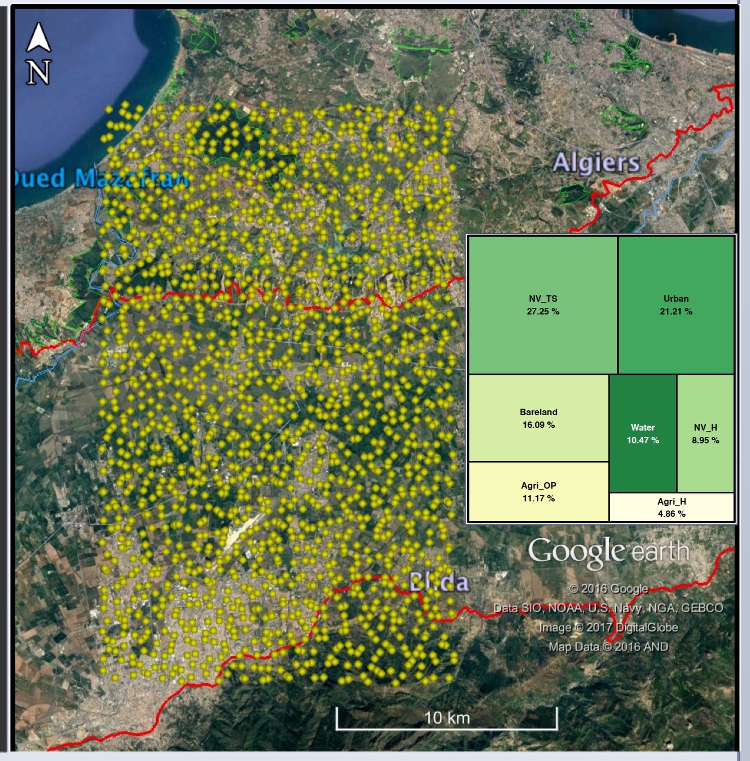
3, Urban / built up (roads, pavement, sidewalks ... etc.)

4, natural vegetation (trees and shrubs)

5, natural vegetation herbaceous

6, agriculture (orchards)

7, agriculture (herbaceous crops)



As part of the preprocessing, dimensionality reduction was applied externally using the correlation-based feature selection (CFS), by Hall (1999), to reduce the number of features to 44. The reduced dataset was used in our analysis. The following steps were undertaken:

- Classification using Logistic Regression (LR)
- Classification using Linear Support Vector Machine (LSVM)
- Classification using Random Forest (RF)

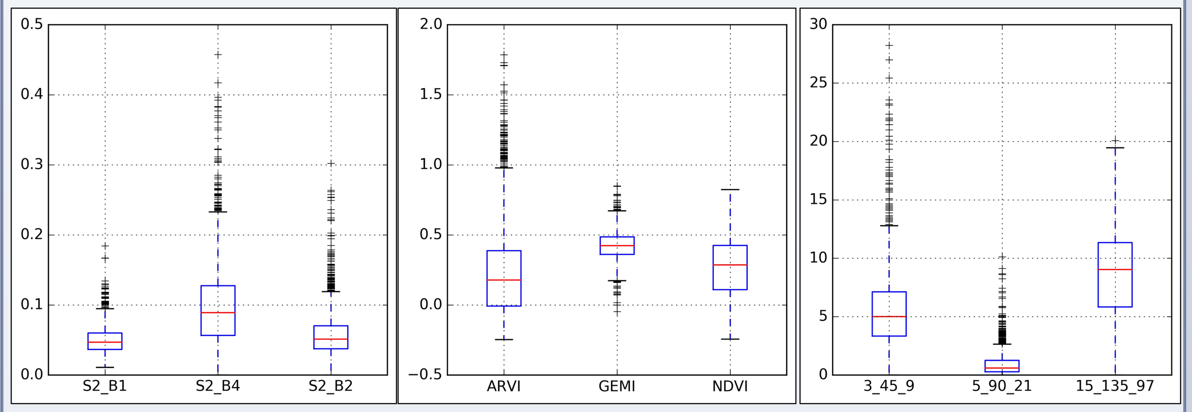
The following individual classifier parameters were tested:

Model parameters (P)	Classifiers LR	RF	SVM
Model 1	Reg. P. = 0 Elastic Net P. = 0	Num. Trees = 10 Max. Depth = 5	Reg. P. = 0.01
Model 2	Reg. P. = 0.02 Elastic Net P. = 0.2	Num. Trees = 100 Max. Depth = 15	Reg. P. = 0.5
Model 3	Reg. P. = 0.1 Elastic Net P. = 0 .4	Num. Trees = 200 Max. Depth = 5	Reg. P. = 1.0

The training, validation, testing split used was 80%, 10%, and 10% respectively. The assessment of classifier performance is reported through the Area Under Curve (AUC).

RESULTS

- Sample descriptive statistics of input features



Sentinel-2 bands | Spectral indices | Texture features

- Validation AUC values for the three tested classifiers and their configurations. The best RF also gets a Kappa = 0.6

Computed AUC	Classifiers LR	RF	SVM
Model 1	0.7507	0.7136	0.5654
Model 2	0.7128	**0.7587**	0.5654
Model 3	0.5309	0.7170	0.5654

- Best model generalization performance:

The best RF is re-trained with 90% of the data and tested on 10% of the data. Then running 100 iterations and storing AUC values.

Generalization AUC 100 iterations	Classifiers: Best RF (90% training, 10% testing)
Max AUC	0.7954
Min AUC	0.7023
Mean AUC	**0.7521**

- Feature importance for the best RF model (at validation stage).

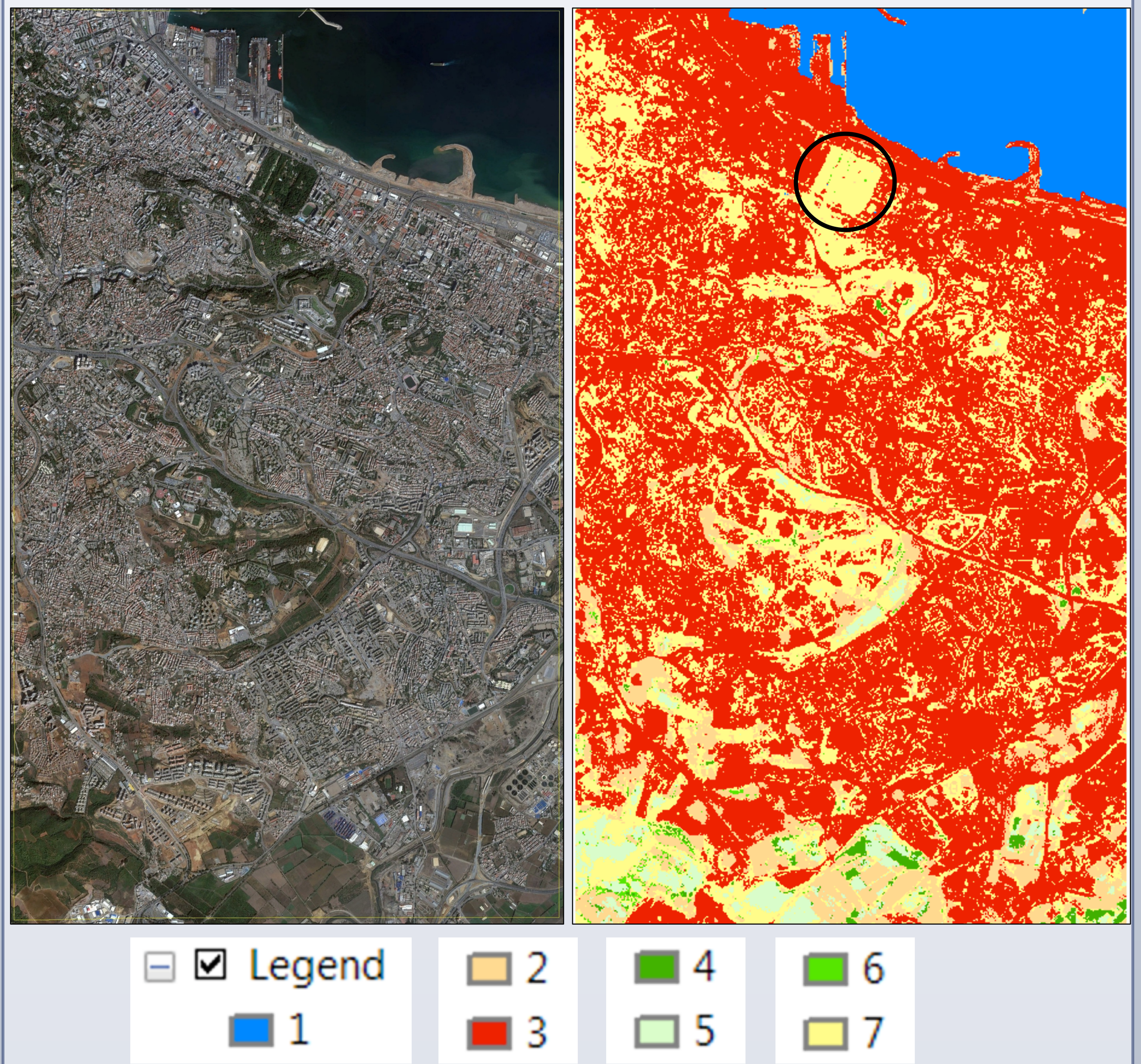
In the top 10 features, we have 6 spectral indices, 3 Sentinel-2 original bands, and 2 texture features.

	feature	importance
5	ARVI	0.043151
10	IPVI	0.042584
19	RVI	0.035737
3	S2_B11	0.034268
6	CI	0.030234
4	S2_B12	0.029862
29	9_45_97	0.029690
0	S2_B1	0.029584
17	PSSRA	0.028396
30	9_90_11	0.027953

30	9_90_11	0.027953
25	5_90_62	0.027678
9	GNDVI	0.027677
15	NDVI	0.027488
26	7_90_15	0.026115
16	NDWI2	0.026109
2	S2_B4	0.025276
21	3_45_9	0.024847
14	MSAVI2	0.023310
39	15_90_13	0.022307
13	MSAVI	0.021809

RESULTS: Land Cover map

The best performing Random Forest classifier was utilized to produce a land cover map using a small image subset of 940*565 pixels. Given the original image spatial resolution (10m), this subset covers an area 5311 hectares.



Examination of the generated land cover map against the corresponding Google Earth imagery show that our classifier performed well in identifying urban areas. However, a large majority of the error in classification is most likely attributed to the overestimation of agricultural areas, which seems to be the second dominant class. This does not make sense because this sample image subset covers the center of the capital of Algeria and is densely urbanized. The misclassification is likely to be grassland, but in the case of the highlighted black circle, it was a botanical garden that contains trees.

CONCLUSIONS

While the classifier performed well with some classes, it still makes too many potentially costly misclassification errors. There is a need to investigate further classifiers such as Neural Network as well as exploring advantages of clustering and image segmentation techniques to further improve the performance.

REFERENCES

Hall, M. A. (1999) 'Correlation-based Feature Selection for Machine Learning'. PhD Thesis. University of Waikato, 178p.