Training samples from open data for satellite imagery classification: using K-means clustering algorithm

Cláudia M. Viana
Universidade de Lisboa
Institute of Geography and Spatial Planning
R. Branca Edmée Marques, 1600-276
Lisbon, Portugal
claudiaviana@campus.ul.pt

Inês Girão
Universidade de Lisboa
Institute of Geography and Spatial Planning
R. Branca Edmée Marques, 1600-276
Lisbon, Portugal
inesgirao@campus.ul.pt

Jorge Rocha
Universidade de Lisboa
Institute of Geography and Spatial Planning
R. Branca Edmée Marques, 1600-276
Lisbon, Portugal
jorge.rocha@campus.ul.pt

Abstract

To create a land use/land cover (LULC) map from a satellite image, we can follow a supervised classification approach if we know what classes exist in the study area and if we have representative training samples for each class. However, in heterogeneous biophysical environments, the wide range of spectral signatures among LULC classes can bias the classification results. In this study, we generated training samples from the official 2015 Portuguese Land Cover Map (COS). In spite of the viability of this source of information (official reference data), we faced some problems with corrupted data and an unbalanced number of training samples per class. As such, we explored the K-means clustering technique in order to understand whether the data had critical issues and to select the most representative training samples by class for satellite imagery classification. We investigated the potential of this technique for LULC classification in a predominantly rural region characterized by a mixed agro-silvo-pastoral environment, which means there is a broad range of spectral signatures for each LULC class. Two image classifications for 2015 were performed using the random forest classifier. The first was done by using the most representative training samples selected from the statistical analysis, and the other was done by using the full generated training set (original training set). Ultimately, the present study demonstrates the improvements in overall accuracy between both image classifications (+8%), showing that the applied methodology has a positive impact on the results.

Keywords: land use/land cover, training set, clustering, Landsat, classification, random forest

1 Introduction

For image classification approaches, pixel-based techniques (unsupervised and supervised) have been largely preferred in Landsat products (Huang *et al.*, 2015; Lu, Ma & Xia, 2017). For case studies at a regional scale, Landsat TM/ETM+ is the most frequently used product. With medium/high spatial resolution, this product is often classified as pixel-based because it provides high-accuracy values (Lu & Weng, 2007). However, the performance of the supervised classification technique is highly dependent on the quality and quantity of the training set used to train the classification model, which may affect the accuracy of the image classification (Lippitt *et al.*, 2008; Brodley & Friedl, 1999). For most supervised classifiers, such as maximum likelihood classification (MLC), multi-layer perceptron (MLP), support vector machine (SVM), or random forest (RF), not having sufficient and representative training data can be detrimental to the image classification results (Lu & Weng, 2007).

Training samples are usually acquired from expert knowledge, field surveys, or through the visual interpretation of other products, e.g., high-resolution images from Google Earth and aerial photographs (Lu & Weng, 2007). However, collecting training samples from fieldwork has high associated costs in terms of money and time, and collecting training samples through visual interpretation can be difficult and cause bias in the representativeness of the samples (Usman, 2013). Nevertheless, various approaches have been used to overcome the insufficient training samples, such as semi-supervised learning or, more recently, active learning (Lu, Ma & Xia, 2017; Huang *et al.*, 2015). However, both approaches are dependent on pre-existing labelled samples, which requires user expertise and proprietary software.

Furthermore, in heterogeneous biophysical environments, the representativeness of each land use/land cover (LULC) class may not exist in the training set because of the spectral confusion among LULC classes (Lu & Weng, 2007). Accordingly, it is essential to explore ways to obtain numerous, high-quality training samples to allow remote sensing applications in such environments.

Therefore, the present study attempts to explore the popular technique of K-means clustering in order to select the most representative training samples by class for satellite imagery classification. We investigate the potential of this technique for LULC classification in a predominantly rural region characterized by a mixed agro-silvo-pastoral environment, which means there is a broad range of spectral signatures for each LULC class. Landsat-8 data based on the Normalized

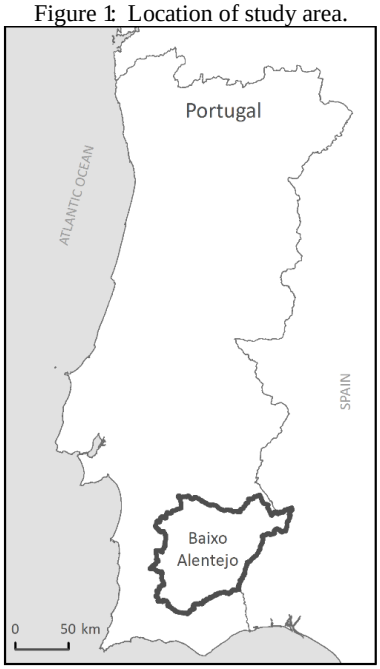
Difference Vegetation Index (NDVI), the Normalized Difference Built-up Index (NDBI), and the Normalized Difference Water Index (NDWI) were used. The official, open data Portuguese Land Cover Map (COS) from 2015 was used to generate the training sets to train the RF classifier. Finally, two image classifications for 2015 were performed; for one of them, the most representative training samples selected from the statistical analysis were used and for the other, the full generated training set (original training set) was used. The classification results were later validated by performing classification accuracy assessment (confusion matrix).

2 Study area and data

2.1 Study area

To apply the methodology, we chose Baixo Alentejo region, located in the southeast of Portugal, with about 8,505 km² (Figure 1). This region is characterized by a vast landscape of wheat, cork oaks, and olive trees, where the dominant land use is mixed agro-silvo-pastoral. The landscape of this region is distinctive, with both fragmented parcels and more compact parcels of land due to the different cropping calendars and field geometries. Dispersed settlements are the predominant urban form in this region.

Figure 1: Location of study area.



2.2 Satellite Image selection and Pre-Processing

Landsat-8 satellite image located on path 203 and row 34 from 2015/03/31 was downloaded from the official website of the United States Geological Survey (USGS). All the images are at Landsat Surface Reflectance Level-2; therefore, no atmospheric correction was needed.

Since we are applying this methodology to a heterogeneous rural area to improve the identification of differences between different LULC classes with similar spectral signatures we computed three indices: (1) NDVI, (2) NDBI, and (3) NDWI. To derive these indices, we used the formulas shown in Table 1.

Table 1: Formulas to derive NDVI, NDBI and NDWI indices

Vegetation index name	Formula
NDVI	$NDVI = \frac{TM\ Band\ 5 - TM\ Band\ 4}{TM\ Band\ 5 + TM\ Band\ 4}$
NDBI	$NDBI = \frac{TM\ Band\ 6 - TM\ Band\ 5}{TM\ Band\ 6 + TM\ Band\ 5}$
NDWI	$NDWI = \frac{TM\ Band\ 3 - TM\ Band\ 5}{TM\ Band\ 3 + TM\ Band\ 5}$

2.3 Reference dataset

The official Portuguese Land Cover Map (COS) of 2015 is produced by the Portuguese General Directorate for Territorial Development (DGT). The dataset is freely available for download from the DGT website. The spatial representation of COS is in polygons with a spatial resolution of 0.5 meters, a minimum mapping unit (MMU) of 1 hectare and is production method is by visual interpretation (i.e., air-photo maps). COS uses hierarchical and *a priori* nomenclature system (5 levels - 225 classes) and, in fact, the COS's first three levels are similar to the three CORINE Land Cover (CLC) map levels.

2.4 Training and validation test

Following COS nomenclature levels 1, 2 and 3, and based on our expert knowledge, we identified seven major LULC classes as the most representative of the selected study area: 1) Non-vegetated surfaces; 2) Herbaceous temporary; 3) Herbaceous permanent; 4) Vineyards; 5) Olive Orchards; 6) Forest and semi-natural surfaces; 7) Water bodies (Table 2).

To generate the training set, we first converted the COS map to raster with a 30-meter resolution (spatial resolution of Landsat-8 satellite image); then, we created a point for each pixel centroid.

Finally, we randomly selected one thousand points per LULC class and used them as the training set for the analysis – we used the remaining points to validate the classification.

Table 2: LULC class correspondence

LULC classes	COS nomenclature code
Non-Vegetated surfaces	1.1.1, 1.1.2., 1.2.1., 1.2.2., 1.2.4., 3.3.2
Herbaceous Periodic	2.1.0., 2.1.3.
Herbaceous Permanent	2.3.1.
Vineyards	2.2.1.
Olive groves	2.2.3.
Forest and semi-natural surfaces	3.1.1., 3.1.2., 3.2.2.
Water bodies	5.1.1., 5.1.2.

3 Methods

The clustering process is schematically described in Figure 2 and is essentially divided into three distinct processes: 1)

classification-trimmed likelihoods calculation (ctlcurves); 2) cluster computation; and 3) mean discriminant factor value calculation (DiscrFact). The variables used for the creation of clusters were NDVI and NDWI indices (the same variables that will be used for the image classification).

3.1 K-means clustering

We used the K-means clustering algorithm to select the most representative training samples for each LULC class. We performed this technique in the R environment using the *tclust* package (Fritz, García-Escudero & Mayo-Iscar, 2012). This package enabled us to apply a clustering method, and prior to the creation of clusters, to determine the most suitable parameters regarding the optimal number of clusters and the number of sample points to be trimmed. The *tclust* package handles various types of clustering methods. For this study, we used the trimmed k-means method that Cuesta-Albertos et al. (1997) introduced. This method is a simple algorithm that uses unsupervised learning to solve known clustering issues; it works well with large datasets.

To create the clusters, as mentioned before, we used NDVI, NDWI, NDBI as well as band two to band five (Landsat-8) as variables in the attempt to spectrally characterize each selected LULC class. After we extracted this information, we calculated the optimal number of clusters using the silhouette method technique (Figure 3A), defining the number of samples that should be trimmed out using the correlated trait locus (CTL) (Arends *et al.*, 2016) curves technique (Figure 3B). In this way, we were not blindly choosing the best value for the parameters that constituted the clustering method.

Thereafter, we implemented the trimmed k-means clustering technique by setting the parameters, in *tclust* function, of *rest* to "eigen" to simultaneously control the relative group sizes and the deviation from sphericity in each cluster. Also, equal weights was set to "TRUE", to avoid the creation of one cluster that is really well-determined but that actually does not represent the LULC class, and to achieve, as much as possible, heterogeneous clusters since no LULC has a pure and unique spectral signatures. For a correct classification, we needed some variability in the range of values to avoid eliminating potentially important information (Fritz, García-Escudero & Mayo-Iscar, 2012).

The first step after applying the method was to observe the discriminant factor value for each cluster, in each class, in order to select the most representative cluster(s) to use in the classification. To analyse these values, we used the *tclust* package as well as the *DiscrFact* function, creating a graphical display that allowed us to know which cluster presented the value for discriminant factor further from zero, and therefore this should be selected (Fritz, García-Escudero & Mayo-Iscar, 2012).

3.2 Random forest classifier

We selected random forest method as our image classifier, since it is very well-known within the remote sensing community, especially because of the high accuracy assessment values achieved in these satellite image classifications (Belgiu & Drăguț, 2016). We performed this classification in the R environment using the RandomForest package. We fixed the number of trees at 500.

3.3 Classification accuracy assessment

We performed the accuracy assessment using the remaining points as described in Section 2.4. We photo-interpreted 50 points per class to ensure they really represented the LULC class. We then validated the classification results with these points. We evaluated the accuracy of the classifications in terms of user and producer accuracy metrics (Congalton, 1991). We also used the R software to derive omission and commission errors.

4. Results

4.1 Classfication accuracy assessment

Figure 2: Workflow representing the proposed methodology

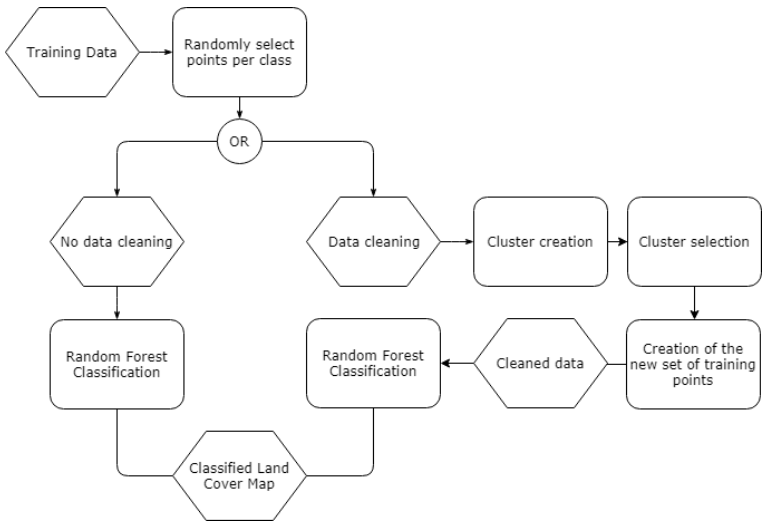
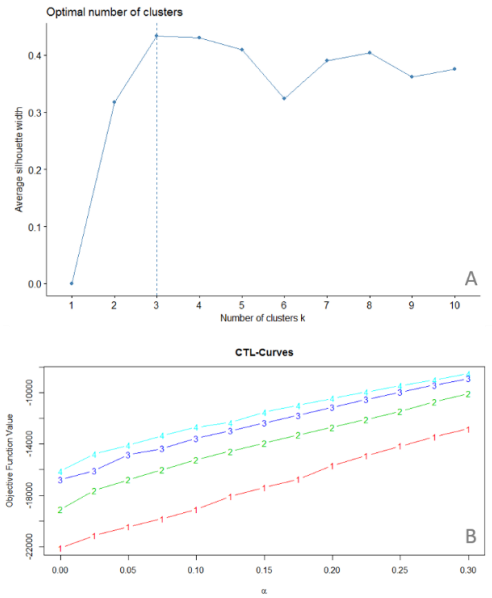


Figure 3. Graphical representation of silhouette method and ctl curves and summary of these components for each LUCL class



Classes	Nº of clusters	Trimmed points (%)
1	4	15
2	2	20
3	4	25
4	3	20
5	4	20
6	3	15
7	4	15

For the first stage of our study, we computed a confusion matrix for each image classification to evaluate classification accuracy. Table 3 presents the accuracy assessment for the original generated training set. The results showed that the overall accuracy was 65.5%; however, there was wide variation among the accuracy values for each LULC class.

In particular, 1) water bodies and forests, and 2) semi-natural surface classes were highly accurate (90% and 74% user accuracy, respectively), suggesting that the areas that we classified as those classes closely matched those in the reference datasets (COS). While the producer's accuracy rates were not quite as high (77% and 70%, respectively), they were

Table 3. Confusion matrix with the validation points (classification with original training samples)

		COS 2015							
		Non-Vegetated surfaces	Herbaceous temporary	Herbaceous permanent	Vineyards	Olive groves	Forest and semi-natural surfaces	Water bodies	User accuracy
Classified map	Non-Vegetated surfaces	30	3	1	13	2	0	1	60
	Herbaceous temporary	5	28	11	4	0	1	1	56
	Herbaceous Permanent	4	3	29	0	3	9	2	58
	Vineyards	12	2	0	31	4	0	1	62
	Olive groves	4	1	3	5	30	5	2	60
	Forest and semi-natural surfaces	0	1	3	2	1	37	6	74
	Water bodies	1	0	2	1	0	1	43	90
	Producer accuracy	54	74	59	55	75	70	77	65.5

Table 4. Confusion matrix with the validation points (classification with cleaned training samples)

		COS 2015							
		Non-Vegetated surfaces	Herbaceous temporary	Herbaceous permanent	Vineyards	Olive groves	Forest and semi-natural surfaces	Water bodies	User accuracy
Classified map	Non-Vegetated surfaces	42	1	0	3	3	0	1	84
	Herbaceous temporary	0	37	7	3	1	2	0	74
	Herbaceous Permanent	1	7	29	1	3	9	0	58
	Vineyards	6	2	1	33	4	4	0	66
	Olive groves	2	3	5	1	34	5	0	68
	Forest and semi-natural surfaces	3	0	3	1	1	42	0	84
	Water bodies	1	1	2	2	0	4	40	79
	Producer accuracy	76	73	62	75	74	64	97	73.3

still high enough to suggest that these classes were correctly shown on the image classification. All the remaining classes had user accuracy values below 62%; however, the producer accuracies for herbaceous temporary and olive groves were 74% and 75%, respectively.

Table 4 presents the accuracy assessment for the most representative training samples selected from the statistical analysis; the results showed a high agreement (73.3%). The increase of user accuracy values for most of the classes is noteworthy. In particular, non-vegetated surfaces were highly accurate, with user and producer accuracy rates of 84% and 76%, respectively. Forest and semi-natural surfaces also had high user accuracy values (84%), while the producer accuracy rates were lower, at 64%. User accuracy for water bodies decreased to 79%, while the producer accuracy increased to 97%. Figure 4 presents the LULC map with the highest classification results.

Discussion and conclusions

To create a LULC map from a satellite image, we can follow a supervised classification approach if we know what classes exist in the study area and if we have representative training samples for each class. However, even respecting these two "rules" in such heterogeneous biophysical environments, the wide range of spectral signatures among LULC classes can bias the classification results (Lu & Weng, 2007).

Generating training samples from a LULC map (such as the COS) from a governmental institution that we considered a reliable source of information can still have associated problems, including an unbalanced number of training samples per class or corrupted data. In such cases, some interpretation and selection (based on expert knowledge and statistical analysis, among others) should be done in order to understand if the data have critical issues (Lippitt *et al.*, 2008; Brodley & Friedl, 1999).

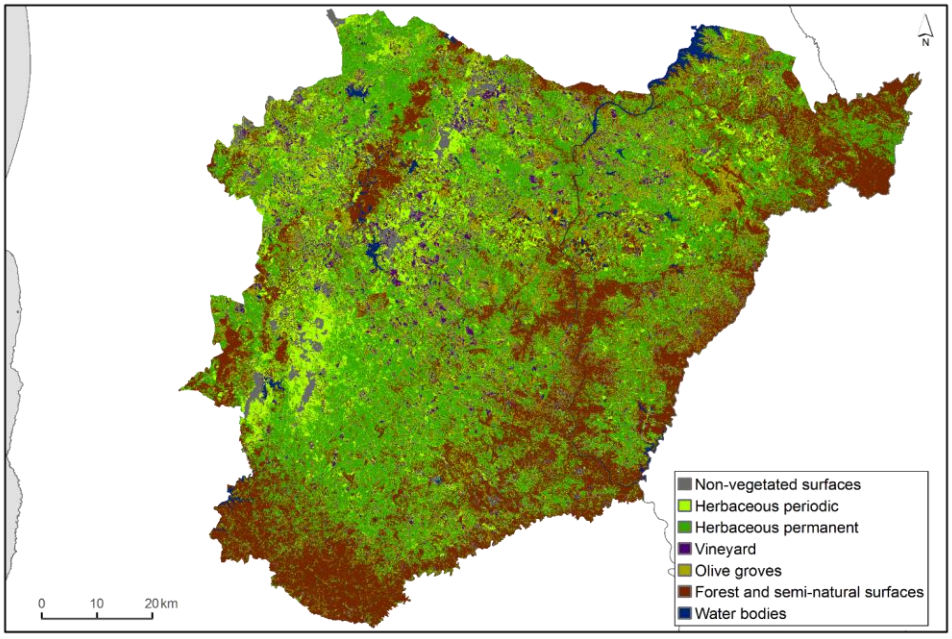
Our purpose was to demonstrate how the pre-cleaning of training samples can have a positive impact on the classification results, especially when there is a wide range of spectral signatures for each class. For example, in past years in Alentejo, one olive grove area has increased substantially, suggesting a change in emphasis to intensive farming without fallow (Viana & Rocha, 2018). This super-intensive olive grove is a particular type of production with its own spectral signature characteristics that differ from the "usual" olive grove since the trees are small and have less space between them. As such, the same LULC class can have different spectral signatures, so it is difficult to discern the difference between the olive grove class and the forest and semi-natural surface class (Table 4).

Additionally, some of the errors identified in the calculation of the confusion matrix can actually be helpful in understanding certain dynamics. As an example, we have some locations where some pixels were classified as herbaceous temporary, and we observed a perfectly opposite shape classified as a body of water, meaning that the field was actually being heavily irrigated during the time that the image was captured. In addition, the vineyard class presents some confusion with the non-vegetated surfaces class, mainly due to how vineyard planting is organised—in lines with spaces of bare soil between them.

One other issue is that croplands have different behaviours throughout the year; some may mature during spring and others in summer, autumn, or winter, so it is expected that different classes will be easier to identify during different times of the year. Based on our results, we argue that using only an image for a specific date may be not be the best approach; it would be interesting to use images for the whole year to see if it helps the classifier to output better classifications.

For future studies, it would be interesting to see if we have good results in case we use the obtained training samples to classify other images of the same area but in a different time (e.g. same month but in 2016). Ultimately, our study

Figure 4 LULC map classification



demonstrates the improvement of the overall accuracy between both image classifications (+8%), showing that the applied methodology had a positive impact on the results. The cluster analysis we performed on R was efficient and straightforward, proving itself to be promising approach.

Funding

This research was funded by the FCT - Portuguese Foundation for Science and Technology [grant number SFRH/BD/115497/2016 to Cláudia M. Viana and grant number SFRH/BD/138979/2018 to Inês Girão].

Acknowledgments

We acknowledge the GEOMODLAB - Laboratory for Remote Sensing, Geographical Analysis and Modelling – of the Center of Geographical Studies/IGOT for providing the required equipment and software.

References

Arends, D., Li, Y., A. Brockmann, G., C. Jansen, R., et al. (2016) Correlation Trait Loci (CTL) mapping phenotype network inference subject to genotype. The Journal of Open Source Software., 1 (6), 87.

Belgiu, M. & Drăguţ, L. (2016) Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing., 114, 24–31.
Brodley, C.E. & Friedl, M.A. (1999) Identifying Mislabeled Training Data. Journal of Artificial Intelligence Research, (11), 131–167.

Congalton, R.G. (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37 (1), 35–46.

Cuesta-Albertos, J.A., Gordaliza, A. & Matrán, C. (1997) Trimmed k-means: An attempt to robustify quantizers. Annals of Statistics, 25 (2), 553–576.

Fritz, H., García-Escudero, L.A. & Mayo-Iscar, A. (2012) tclust: An R Package for a Trimming Approach to Cluster Analysis. Journal of Statistical Software,47 (12).

Huang, X., Weng, C., Lu, Q., Feng, T., et al. (2015) Automatic labelling and selection of training samples for high-resolution remote sensing image classification over urban areas. Remote Sensing, 7 (12), 16024–16044.

Lippitt, C.D., Rogan, J., Li, Z., Eastman, J.R., et al. (2008) Mapping Selective Logging in Mixed Deciduous Forest: A Comparison of Machine Learning Algorithms. Photogrammetric Engineering & Remote Sensing, 74 (10), 1201–1211.

Lu, D. & Weng, Q. (2007) A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing, 28 (5), 823–870.

Lu, Q., Ma, Y. & Xia, G.S. (2017) Active learning for training sample selection in remote sensing image classification using spatial information. Remote Sensing Letters, 8 (12), 1210–1219.

Usman, B. (2013) Satellite Imagery Land Cover Classification using K-Means Clustering Algorithm Computer Vision for Environmental Information Extraction. Sci. & Engg., 63 (October 2013), 18671–18675.

Viana, C.M. & Rocha, J. (2018) Spatiotemporal analysis and scenario simulation of agricultural land use land cover using GIS and a Markov chain model. In: A. Mansourian, P. Pilesjö, L. Harrie, & R. von Lammeren (eds.). Geospatial Technologies for All: short papers, posters and poster abstracts of the 21th AGILE Conference on Geographic Information Science. 2018 Lund, Sweden, 12-15 June 2018. p.