Improving image classification accuracy: a method to incorporate uncertainty in the selection of training sample set

Luísa M S Gonçalves
Polytechnic Institute of Leiria
Department of Civil Engineering, Portugal
Institute for Systems and Computers Engineering at Coimbra (INESC Coimbra)
Coimbra, Portugal
luisa.goncalves@ipleiria.pt

Cidália C Fonte
Institute for Systems and Computers Engineering at Coimbra (INESC Coimbra)
Coimbra, Portugal
Department of Mathematics - University of Coimbra
Coimbra, Portugal
cfonte@mat.uc.pt

Hugo Carrão
Remote Sensing Unit (RSU)
Portuguese Geographic Institute (IGP)
Lisboa, Portugal

Mario Caetano
Remote Sensing Unit (RSU)
Portuguese Geographic Institute (IGP)
Lisboa, Portugal
Research Centre for Statistics and Information Management (CEGI), Institute for Statistics and Information Management (ISEGI),Universidade Nova de Lisboa, Campus de Campolide
Lisboa, Portugal
{hugo.carrao, mario.caetano}@igeo.pt

Abstract— **The automatic production of land cover maps using multispectral remote sensing images requires the use of learning classifiers for mapping the imagery data into a set of discrete classes. A group of classifiers commonly used are the supervised classifiers. The first stage of a supervised classification consists on the identification of training areas in the satellite image for each class, which are then used as descriptors of the spectral characteristics of the different classes. The classification results are therefore influenced by the sample pixels selected as training sets. This paper proposes an automatic method to assist the selection of training samples for mapping land cover from satellite images with the aid of ancillary information, namely elder or contemporaneous maps with lower spatial resolution, the Normalized Difference Vegetation Index and information provided by the classification uncertainty. It is shown that more accurate outputs may be derived with this methodology and some conclusions are drawn.**

Keywords: ***training samples, soft classification, measures of uncertainty, classification accuracy***

I. INTRODUCTION

The supervised classification of multispectral images requires the use of a training set, formed by a group of samples that describe the spectral characteristics of all classes in the nomenclature. This set is usually obtained by identifying pixels in the image corresponding to each of the classes, which will be used as classes' descriptors. This approach however requires the human intervention to choose the regions considered to describe properly each class.

Several studies have highlighted the usefulness of the additional information provided by soft classifiers together with the application of uncertainty measures (Maselli, Conese and Petkov, 1994; Ricotta, 2005; Bo and Wang, 2008; Gonçalves, Fonte, Júlio and Caetano, (in press)). For example, Gonçalves, Fonte, Júlio and Caetano (2009a) have shown that uncertainty information can be used as indicator of the classifier difficulty to assign only one class to the spatial unit. Gonçalves, Fonte, Júlio and Caetano (2009b) developed a method to incorporate the uncertainty in the classification process to improve the output accuracy. However, the application of the uncertainty information in the full process of automatic production of thematic maps, using multispectral remote sensing images, is still a field of investigation.

The main contribution of this study is to evaluate the usefulness of the uncertainty information to assist the automatic selection of the training sets.

The methodology proposed in this paper is to use knowledge about the land cover distribution in the study area to identify the regions where the training sample set will be collected. In detail, we plan to use the information available in other raster land cover maps with lower spatial resolution or prior knowledge about the land cover classes obtained from vector maps. The main advantage of this procedure is that the classification process may be completely automated. The main problem raised by this approach is that many of the sample pixels chosen will actually not represent the expected class, because the used map has a lower spatial resolution, it is itself a generalization of the true land cover to match required technical specifications, such as a Minimum Mapping Unit (MMU), or because the map is not contemporaneous of the used satellite images and therefore other land cover classes may exist in the chosen pixels. So, the main need of this approach is to identify a method to eliminate the pixels of training sample that actually do not

describe properly the class to be classified. The methodology proposed uses *a priori* NDVI information and information about the classification uncertainty to exclude the undesirable pixels from the training set.

II. DATA

Two data sets were used in this work: SPOT- 4 images of the region of Leiria (Portugal) of 2006 and the vector CORINE Land Cover (CLC) cartography of 2006.

The SPOT-4 images have 4 spectral bands, namely (1) green (0.50-0.59 μm), (2) red (0.61-0.68 μm), (3) near infrared (0.78-0.89 μm) and (4) short-wavelength infrared (1.58-1.75 μm), and a spatial resolution of 20m.

The CLC cartography has a scale of 1:100 000 and a MMU of 25ha (see Fig. 1). The nomenclature is structured in three hierarchical levels and has 44 classes at the third and most detailed level. The study area includes 19 classes of the CLC nomenclature and the area is mainly occupied by discontinuous urban fabric (code 112) (11%), complex cultivation patterns (code 242) (17%), coniferous forest (code 312) (19%) and mixed forest (code 313) (13%). The land occupied by agriculture includes also vineyards (code 221) (0.06%) and permanently irrigated land (code 212) (0.04%).

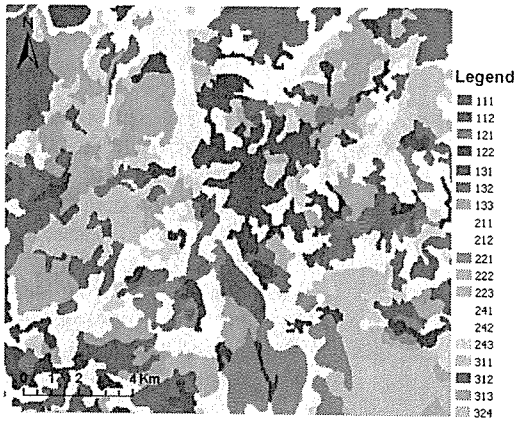

Figure 1. CORINE Land Cover cartography of the study area.

III. METHODOLOGY

The methodology used consists of the following steps: 1) establishment of the protocol and selection of the training and testing sample pixels using the CLC cartography; 2) computation of the NDVI; 3) exclude from the training set sample pixels that, according to the NDVI, correspond to vegetation and are in classes that do not include vegetation, and sample pixels that are not vegetation and are in sample classes corresponding to vegetation classes; 4) perform a classification with the redefined training sample; 5) extract the degree of uncertainty of the classification for the pixels in the training sample; 6) eliminate the training sample pixels with low levels of uncertainty; 7) classification of the image with the cleaned training sample; 8) evaluate the classification accuracy.

To determine if the proposed methodology improves the classification results, a classification was made with the initial set of training sample points, identified in step 1 (named Classification 1) and its accuracy was evaluated. In the sequence, the classification performed with the proposed methodology is named Classification 2 and its accuracy was compared with that attained with Classification 1. The classifications accuracy evaluation was made with a stratified random sampling by selecting 100 pixels per class considering the entire image scene.

The initial training set was obtained with a stratified random sample of 1000 points per class, collected over the CLC polygons (Fig. 2), and the process was completely automated. Then, a land cover classification using the Maximum Likelihood Classifier was performed. The classes used in this classification were: Artificial Areas (AA); Barren Areas (BA), Irrigated Herbaceous Crops (IHC); Forest Areas (FA), Heterogeneous Agriculture Areas (HA) and Shrub Lands (SL). These classes were obtained generalizing the CLC nomenclature (see Table I).

TABLE I. USED NOMENCLATURE, OBTAINED FROM THE CORINE LAND COVER NOMENCLATURE.

CLC CODE	CLC LABEL	New Code	Study Label
111	Continuous urban fabric	1	Artificial Areas
112	Discontinuous urban fabric	1	Artificial Areas
121	Industrial or commercial units	1	Artificial Areas
122	Road and rail networks and associated land	1	Artificial Areas
131	Mineral extraction sites	2	Barren
132	Dump sites	2	Barren
133	Construction sites	2	Barren
211	Non-irrigated arable land	6	Heterogeneous Agriculture
212	Permanently irrigated land	3	Irrigated Herbaceous Crops
221	Vineyards	6	Heterogeneous Agriculture
222	Fruit trees and berry plantations	6	Heterogeneous Agriculture
223	Olive groves	6	Heterogeneous Agriculture
241	Annual crops associated with permanent crops	6	Heterogeneous Agriculture
241	Complex cultivation patterns	6	Heterogeneous Agriculture
243	Land principally occupied by agriculture, with significant areas of natural vegetation	6	Heterogeneous Agriculture
311	Broad-leaved forest	4	Forest Areas
312	Coniferous forest	4	Forest Areas
313	Mixed forest	4	Forest Areas
323	Sclerophylous vegetation	5	Shrublands
324	Transitional woodland/shrub	2	Forest Areas or Barren

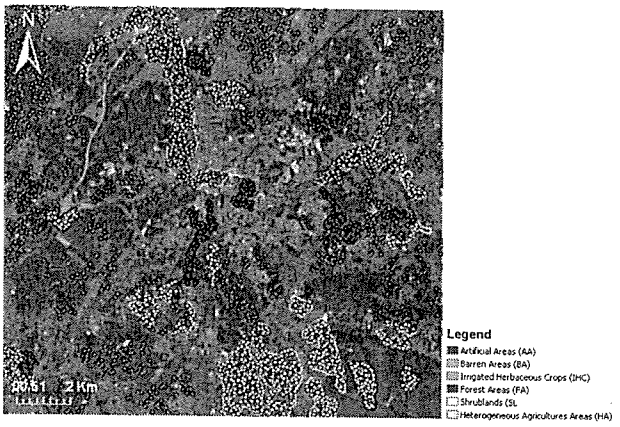

Figure 2. SPOT-4 image (321) of the study area overplayed with the sample sites initially included in the training set.

The uncertainty was evaluated for each pixel using the uncertainty measure E, developed by Chow (1970), and given by (1)

$$E(x) = 1 - p(x) \quad (1)$$

where $p(x)$ is the largest degree of probability of the probability distributions assigned to pixel x.

IV. RESULTS

The accuracy assessment for both classifications was made with an error matrix and was undertaken with the same testing datasets. The Global Accuracy was computed as well as the User's Accuracy (UA) and the Producer's Accuracy (PA) for all classes. Fig. 3 shows the results of the preliminary classification made with the 1000 points per class (Classification 1) and Table II the confusion matrix obtained for this classification. Fig. 4 shows the uncertainty of Classification 1.

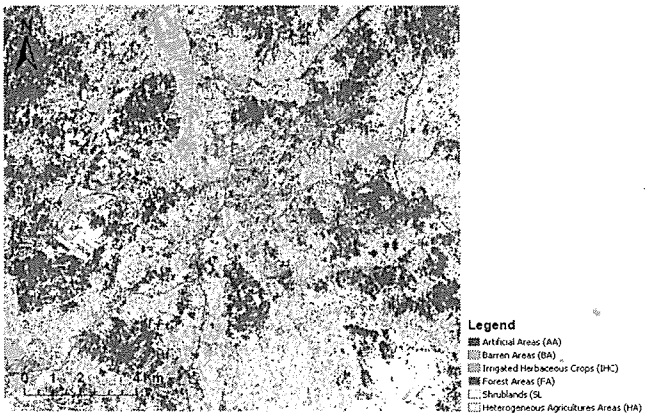

Figure 3. Classification results obtained with the 1000 points per class chosen automatically from the CLC cartography (Classification 1).

TABLE II. CONFUSION MATRIX OBTAINED WITH THE 1000 POINTS PER CLASS CHOSEN AUTOMATICALLY FROM THE CLC CARTOGRAPHY (CLASSIFICATION 1)

		Reference Label of Classification 1						User Accuracy (%)
		AA	BA	IHC	FA	SL	HA	
MAP Label	AA	38	14		1	1	1	69.09
	BA	24	29			1	2	51.79
	IHC		3	50	17	7	10	57.47
	FA	7	7		90	4	1	82.57
	SL	6	11		18	78	17	60.00
	HA	18	38	10	5	21	71	43.56
Producer Accuracy (%)		40.66	28.43	83.33	68.70	69.64	69.61	59.33%

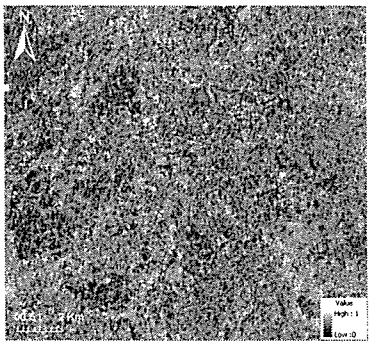

Figure 4. Spatial distribution of uncertainty obtained in Classification 1.

The results obtained with the initial training set show that, according to the UA, Forest Areas (FA) (82.6%) is the class better classified, and according to the PA, it is Irrigated Herbaceous Crops (IHC) (83.3%). The classes with smaller PA are Barren Areas (BA) (28.4%) and Artificial Areas (AA) (40.9%), which means that these are the classes with more omission errors, which also presented a great confusion between each other. The classes with smaller UA are Heterogeneous Agriculture Areas (HA) (43.6%), Barren Areas (51.8%) and Irrigated Herbaceous Crops (IHC) (57.5%) and are therefore the classes with more commission errors.

Some of the problems of Classification 1 derive from the fact that the used vector map is a generalization of the true land cover, obtained by visual analysis performed by a human interpreter, and therefore other land cover classes exist in the chosen pixels. Fig. 5 shows an example of the procedure used to filter the training set for the class Irrigated Herbaceous Crops (IHC), using the NDVI and the uncertainty. In Fig. 5a) a small area of the image is shown, overlaid with the initial training set. In Fig. 5b) the obtained NDVI values are shown overlaid with the sample pixels' locations. Since class IHC is a vegetation class, pixels with low values of NDVI were excluded from the training set (points in red). Fig. 5c) shows the uncertainty value per pixel overlaid with the training set, where only points corresponding to values of uncertainty lower that 0.25 were considered.

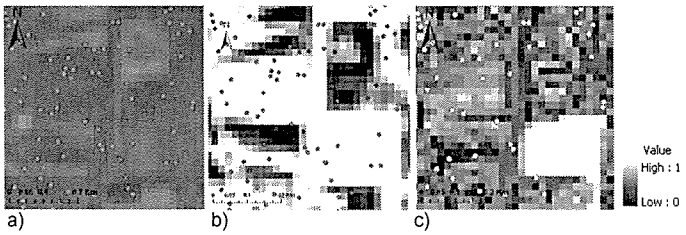

Figure 5. Extract of a region belonging to Irrigated Herbaceous Crops: a) SPOT image overlaid with the initial training set; b) NDVI information overlaid with training set (points that do not include vegetation are shown in red and excluded from the training sample); c) uncertainty information overlaid with the final sample.

Fig. 6 shows the values of the mean uncertainty per class obtained for the training sample after its first redefinition using the NDVI. It can be seen that low values were obtained for all classes (always less than 0.33). The classes that show slightly higher levels of uncertainty are BA, AA and IHC.

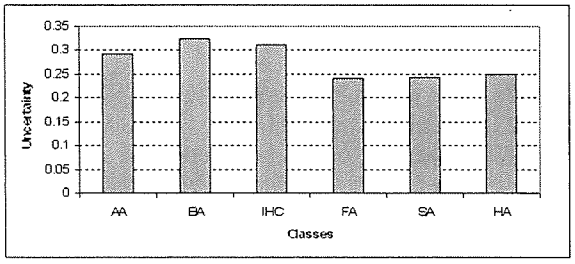

Figure 6. Mean uncertainty per class of the training set obtained after its preliminary redefinition using NDVI information.

The classification was than redone and the results evaluated computing the accuracy indices for the testing set.

Table III presents the confusion matrix obtained. The resulting global accuracy is 71.5%, which reveals an increase of 12% when comparing with Classification 1 accuracy. Fig. 7 and Fig. 8 allow the comparison between the results for the UA and PA before and after the redefinition of the training areas.

Fig. 9 shows the final results obtained with the classifier using the proposed methodology.

TABLE III. CONFUSION MATRIX OBTAINED FOR THE CLASSIFICATION 2 (OBTAINED AFTER FILTERING THE TRAINING SET).

		Reference Label of Classification 2						User Accuracy (%)
		AA	BA	IHC	FA	SL	HA	
MAP Label	AA	65	32	0	0	3	0	65.00
	BA	22	63		1	3	11	63.00
	IHC			62	15	13	11	61.39
	FA				97	3		97.00
	SL	2	2		10	74	12	74.00
	HA	4	5	5	3	15	68	68.00
Producer Accuracy (%)		69.89	61.76	92.54	76.98	66.67	66.67	71.50%

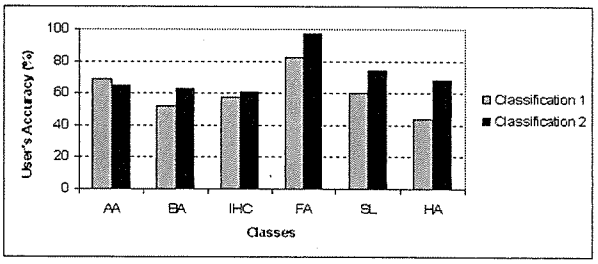

Figure 7. User's accuracy of Classification 1 (classification with the training points chosen automatically from the CLC cartography) and Classification 2 (obtained after filtering the training set).

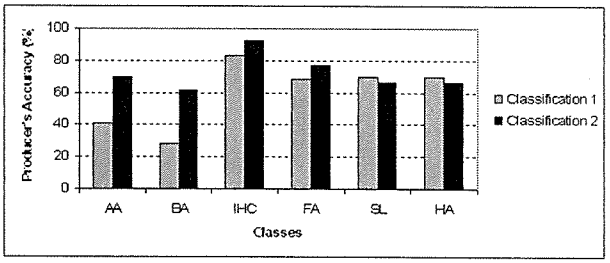

Figure 8. Producer's accuracy of Classification 1 (classification with the training points chosen automatically from the CLC cartography) and Classification 2 (obtained after filtering the training set)

V. CONCLUSIONS

The results obtained with this study showed that the use of the NDVI and the information about the classification uncertainty proved to be valuable in the process of filtering undesirable pixels from a training set obtained automatically from a map with larger MMU. This approach allowed the improvement of the classification accuracy in 12% after the redefinition of the training areas.

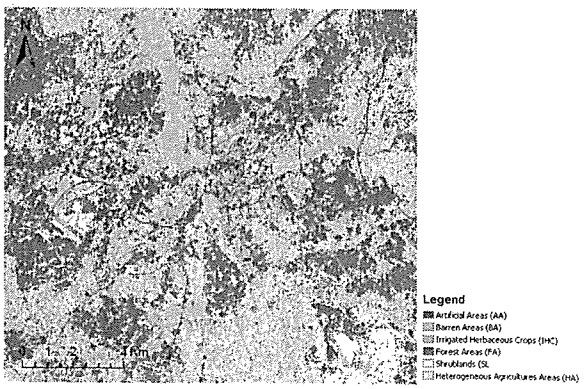

Figure 9. Classification results obtained with the redefined training sample (Classification 2).

The proposed methodology may be applied to obtain improved training sets, through a fully automated approach, using available prior knowledge about the land cover. In detail, we showed that maps with larger MMU, contemporaneous or not of the base satellite images, may be used to obtain, from remote sensing images, new maps with higher resolution. For all the reasons mentioned, it can be stated that the proposed approach is quite promising, deserving therefore further investigations.

REFERENCES

Bo, Y.C. and J.F. Wang. A general method for assessing the uncertainty in classified remotely sensed data at pixel scale. In J. Zhang and M. F.Goodchild (Eds) *Spatial Uncertainty, Proceedings of Accuracy 2008, 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science* (pp. 186-194) Sciences World Academic Press.

Chow, C.K. (1970). On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*. 16, 41–46.

Gonçalves, L. M., Fonte, C. C., Júlio, E. N. B. S., Caetano, M. (2009a). Evaluation of remote sensing images classifiers with uncertainty measures. In Rodolphe Devillers and Helen Goodchild (Eds) *Spatial Data Quality From Process to Decisions* (pp.163-177). Boca Raton: CRC Press Taylor & Francis Group.

Gonçalves, L. M., Fonte, C. C., Júlio, E. N. B. S., Caetano, M. (2009b). A method to incorporate uncertainty in the classification of remote sensing images. *International Journal of Remote Sensing*. 30 (20), 5489–5503.

Gonçalves, L. M., Fonte, C. C., Júlio, E. N. B. S., Caetano, M., (in press) Evaluation of soft possibilistic classifications with non specificity uncertainty measures. *International Journal of Remote Sensing*.

Maselli, F., Conese, C., Petkov, L. (1994). Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. *ISPRS Journal of Photogrammetry and Remote Sensing*. 49, 13-20.

Ricotta, C., (2005). On possible measures for evaluating the degree of uncertainty of fuzzy thematic maps. *International Journal of Remote Sensing*. 26, 5573-5583.