DERIVING CLASSIFICATION RULES FROM MULTIPLE REMOTELY SENSED URBAN DATA WITH DATA MINING

Sheeren D.[1], Puissant A.[2], Weber C.[3], Gançarski P.[1] and Wemmert C.[1]

1. UMR 7005 – CNRS / ULP Strasbourg, LSIIT, Machine Learning and Data Mining Group
   Mail: sheeren ; gancarski ; wemmert @lsiit.u-strasbg.fr
2. FRE 2795 – CNRS / University of Caen, Department of Geography, Geosyscom Laboratory
   Mail: anne.puissant@unicaen.fr
3. UMR 7011 – CNRS / ULP Strasbourg, Image & Ville Laboratory
   Mail: christiane.weber@lorraine.u-strasbg.fr

ABSTRACT

In a context of urban planning, it is necessary to support the identification and the formalization of the urban elements. Very often, it requires some complementary aspects of a set of images and also ancillary data. However the lack of methods enabling the combination of several sources is still compelling. In general, the use of several sources of remotely sensed data in a classification procedure results in data fusion upstream or fusion of the results.

Since the appearance of VHR-images, object-oriented methods have been defined to image analysis. This approach involves segmenting images into homogeneous regions and characterizing objects with a set of features related to spectral signatures, and to spatial and contextual properties. The main issue in this approach is the definition of the knowledge base classification. Generally, the relevant information is not well-formalized and it is difficult to grasp knowledge directly from domain experts. The experts are rarely able to supply an explicit description of the knowledge they use for objects identification.

In this paper, we propose to use data mining techniques to derive automatically a set of classification rules from remotely sensed data. Knowledge is extracted from a VHR-image (Quickbird MS) following an object-oriented approach. We also investigate the possibilities of acquiring matching rules from multiple classified images. These rules can help to improve the classification accuracy. They can also be used for building a multi-scale database. Experiments show the effectiveness of the proposed approach. Our first results indicate that the performance of the learnt rules is acceptably good.

1. INTRODUCTION

Urban planning requires identification, localization and formalization of the urban elements (impervious surfaces, vegetation, water). Very often, the identification step requires some complementary aspects of a set of images and also ancillary data: seasonal to discriminate mineral surfaces (zones of agriculture or not), spectral to supplement the range of the effective spectral answers and finally spatial to take into account (1) the relationships between the studied area and (2) the adequacy between the resolution of the pixel and objects of interest. However the lack of methods facilitating the selection of useful data, the improvement of extraction of knowledge and the interpretation assistance adapted to the needs is still compelling.

Classical methods (pixel-based classification) do not allow simultaneous and complementary approaches. In general, the use of several sources of remotely sensed data in a classification procedure results in data fusion upstream or fusion of the results. Moreover, in the traditional pixel-based classification methods only the pixels spectral information is used to extract urban objects. This approach can not satisfy high resolution images classification accuracy.

Since the appearance of VHR-images (Very High Resolution), the current tendency is the development of object oriented-methods [1,2]. This approach involves segmenting an image into objects

(group of pixels). These objects have geographical features such as shape and length, and topological properties, such as adjacency, inclusion etc. These features characterize the objects. They can be called upon in the classification process.

While there are some studies comparing object-oriented and pixel-based classification techniques, only few works focus on the development of the knowledge base for classifying urban areas. The main difficulties for this task are to define the rules taking into account the expert knowledge. The experts are rarely able to supply an explicit description of the knowledge they use for objects identification.

In this context, this paper proposes a knowledge acquisition method based on data mining techniques to define generic classification rules of urban objects for remotely sensed imagery. It also studies the possibilities of extracting rules from multiple images to improve the classification accuracy.

The paper is structured as follows. In the next section, principles of data mining techniques are detailed (2). Then, the study area and the data used are presented (3). We expose in section 4 the learning procedure to extract generic classification rules from one image. These rules are derived from a VHR-image (Quickbird MS). Finally, we investigate the possibilities of extracting matching rules between classes of two images (Quickbird MS and Landsat ETM+) in section 5.

2. BUILDING A KNOWLEDGE BASE WITH MACHINE LEARNING

In a knowledge base classification approach, the main difficulty consists in the acquisition of classification rules. In general, it is rather difficult to draw knowledge from domain experts. The experts are rarely able to supply an explicit description of the knowledge they use for objects identification. In addition, acquiring knowledge in this way takes usually a long time. This is a well-known problem within the artificial intelligence community. It has led to the emergence of machine learning techniques [3] that we propose to use for this study. These techniques can help to extract knowledge automatically from the raster data.

Machine learning techniques can be divided in two basic categories: unsupervised learning (or learning from observations) and supervised Learning (or learning from examples). We are interested by the second category for which the algorithms are guided by domain knowledge. Supervised learning requires to provide a set of training examples given by an expert to derive a general classification model. The expert gives some examples in the form of, on one hand, a description of an object and on the other hand, a classification of this object. Learning algorithms build automatically some rules for these examples to explain the classification from the object description (the target or classification function). These rules can then be used to classify unknown examples: this is an *inductive* process.

As we intend to learn interpretable rules and build a reusable knowledge base, we use symbolic supervised machine learning tools for this study (in opposition to numerical approaches). One very prominent symbolic algorithm is the C4.5 classifier proposed by [4]. It enables to create a decision tree providing the shortest optimal description possible for a classification into the given classes. This algorithm is based on the entropy measure and requires an attribute-value list as representation language for the input data. Results can easily be used, validated and if necessary, revised by an expert. This algorithm has been retained for our experiments.

Decision tree classifiers have already proved to be a great practical value for classification tasks in remote sensing [5]. Compared to other classification methods like maximum likelihood procedure (ML) or artificial neural networks (ANN), decision trees are generally more performant [6]. In the context of our object-oriented approach, decision trees help us to build automatically a knowledge base classification.

3. STUDY AREA AND DATA

The examined area for making our experiments is situated in the urban area of Strasbourg (France). This area has an extent of 42 km². It is representative of western cities and is characterised by many different objects with a diverse range of spectral reflectance values. The data used are:

- a non classified QUICKBIRD MS image (4 bands) with a spatial resolution of 2,8 m (May 02).
- a classified LANDSAT TM image (7 bands) with a spatial resolution of 30 m (April 2001)
- a set of ancillary vector data derived from the BDTOPO database with a 1 m resolution provided by the French Mapping Agency

In the first experiments, we have focused on the extraction of knowledge to classify the QUICKBIRD image using the decision tree approach (section 4). In the second experiments, we have studied the possibilities of combining the QUICKBIRD image with the LANDSAT image classified in order to improve the classification results and understand the relationships between the images of different resolution (section 5). The vector data were used as an additional layer to enrich the training sets of spatial and contextual properties for the learning procedure.

4. DERIVING CLASSIFICATION RULES FROM ONE IMAGE

We illustrate in figure 1 the general process we have followed to discover rules to classify the QUICKBIRD image. It is composed of several steps including image segmentation, feature extraction, definition of learning examples, rules acquisition (application of decision tree classifier), classification of the image and validation of the results.

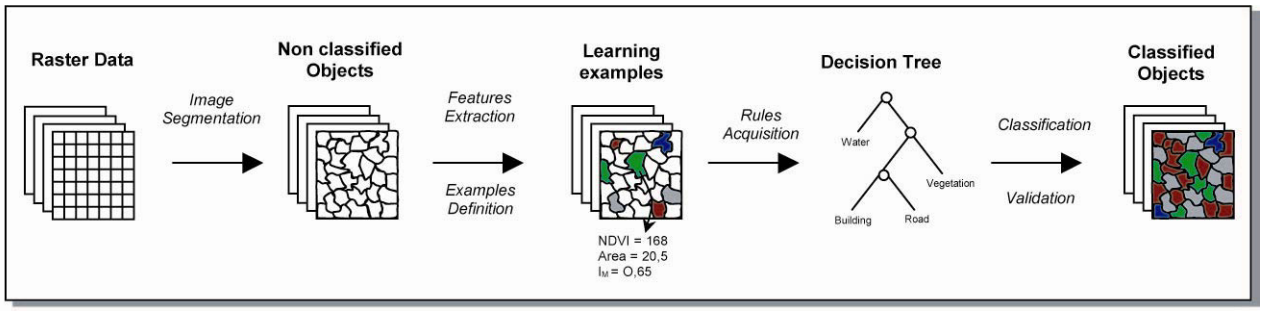

Figure 1. Steps of the rules acquisition process

4.1. Image segmentation

The first step of the process was the QUICKBIRD VHR-image segmentation. The aim of this step was to subdivide the image into separated regions corresponding to the objects of interest. The *eCognition* object-oriented image analysis software developed by Definiens Imaging was used for this task. The segmentation method proposed in this system follows an ascending region growing approach. The algorithm is based on homogeneity criteria in combination with local and global optimization techniques to extract regions from raster data [7].

The method was parameterized to create regions that correspond to a particular class hierarchy of elementary urban objects:

- Level 1: water (1), vegetation (2), shadow (3), bare soil (4), mineral area (5);
- Level 2: grass (2.1), tree (2.2), road (5.1), building (5.2);
- Level 3: continuous built-up area (5.2.1), residential area (5.2.2), commercial and industrial area (5.2.3), collective building (5.2.4)

In order to obtain homogeneous regions with representative shape and dimension, we used data (roads, buildings and hydrographic network) from a vector topographical database for the segmentation process. The eCognition software enables to take into account thematic layers to constraint the process of creating objects. Results are illustrated in figure 2.

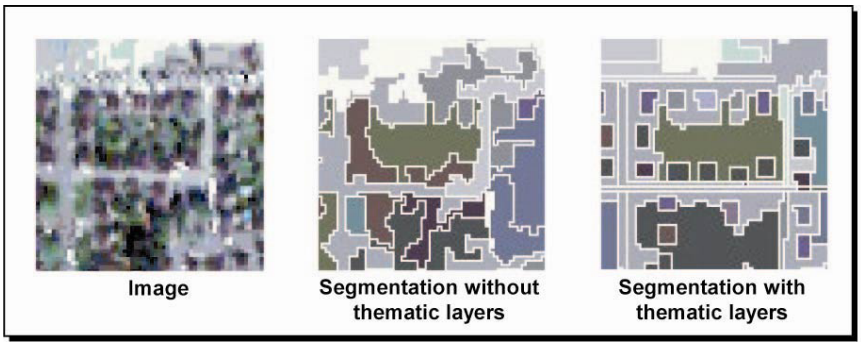

Figure 2. Taking shape and dimensions properties into account for the classification process supposes to create homogeneous and representative regions.

4.2. Feature extraction and definition of learning examples

After the segmentation, each region was characterised by means of a set of features. These features were selected to distinguish the objects of the class hierarchy.

The spectral reflectance of pixels composing the objects is the first criterion that we used. Several features were retained: the mean spectral value of the objects in the four bands of the image (red, green, blue and near infrared), the mean value of the NDVI index (Normalized Difference Vegetation Index), and the mean value of the SBI index (Soil Brightness Index). The rules that we acquired to identify the objects of the first and the second levels of the class hierarchy were only based on this knowledge. It enables to discriminate elementary land cover classes.

For the third level of the class hierarchy, additional information was derived. Spectral signatures were not sufficient to determine the functional character of the buildings in urban environment. The corresponding classes have very close spectral values and therefore overlap in the feature space. Spatial and contextual information are more relevant to separate residential buildings, collective buildings, industrial and commercial buildings, and continuous built-up area. Thus, we also computed this kind of features to recognize the building objects. Several shape properties were selected: area, perimeter, diameter (length of the major axis), compactness (I_M = Miller's index) and solidity (I_S =ratio of the area to the convex hull area). The percentage of vegetal area (P_v) in the surrounding of the buildings was also retained as a relational feature. A buffer with a 20 meters radius was computed for this attribute.

The learning examples were defined from these characterized regions. 50 objects of each class were visually interpreted and labelled interactively by a domain expert to create the training sets. The rules acquisition procedure is detailed below.

4.3. Rules acquisition

We tried to learn rules with the C4.5 decision tree classifier in two ways. First, in one step, by using directly all the examples labelled in one of the classes defined (water, vegetation, road…). Second, in several steps, by learning rules enable to distinguish successively the objects of one class from all others, starting with the objects easiest to identify. That means that we first tried to acquire rules to discriminate water from the other classes. These last ones were merged and labelled as being non water. Then, we learned rules to identify vegetation from the set of examples labelled as non water, and so on.

In practice, the rules acquisition procedure in several steps gives more accurate results. The hypothesis space of the classifier is reduced with this approach. By merging classes, we abstract them and then simplify the resolution of the problem. Several rules have been learned (the decision trees have been converted into decision rules):

SPECTRAL RULES: Entire range of values [0…255]

Class Hierarchy - Level 1:

Rule 1: IF NDVI < 38.23 and IBS > 14.67 THEN Class = Water
ELSE Class = Non Water
Rule 2: IF NDVI < 169.14 THEN Class = Vegetation
ELSE Class = Non Vegetation
Rule 3: IF GREEN < 15.65 THEN Class = Shadow
ELSE Class = Non Shadow
Rule 4: IF NIR > 59.25 and BLUE < 57.86 THEN Class = Bare Soil
IF RED > 101.24 THEN Class = Bare Soil
ELSE Class = Mineral

Class Hierarchy - Level 2:

Rule 5: IF 60.2 < BLUE < 130.8 THEN Class = Road
ELSE Class = Building
Rule 6: IF GREEN > 30.4 THEN Class = Grass
ELSE Class = Tree

SPATIAL RULE:

Class Hierarchy - Level 3:

Rule 7: IF AREA > 5203 m² and I_M > 0.3 THEN Class = Industrial or Commercial Building
IF AREA < 436.8 THEN Class = Residential Building
IF AREA < 1254.9 THEN Class = Collective Building
IF P_V > 11.9 and AREA < 1803.2 THEN Class = Collective Building
IF I_S < 0.43 THEN Class = Collective Building
ELSE Class = Continuous Built-up Area

4.4. Classification and validation

In order to assess the quality of the rules, we introduced them in the *eCognition* software and applied them to label each region of the image (figure 3). The relevance of the rules was estimated computing the confusion matrix and Kappa statistics. Results are detailed in table 1.

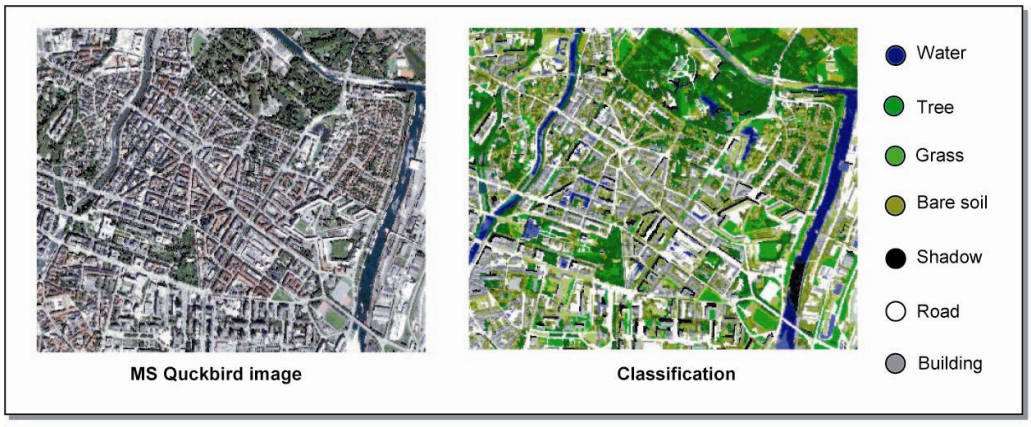

Figure 3. Classification of level 1 and 2 with the learning rules.

The overall accuracy of the obtained classification is 79.7% with Kappa value of 75.5%. An error matrix was also defined to assess the rules intended to classify buildings according to their functional membership. An overall classification accuracy of 80.2% was computed.

Table 1. Accuracy assessment of the classified QUICKBIRD image using the learning rules

Ref. / User	Water	Shadow	Bare Soil	Building	Road	Tree	Grass	User's Accuracy
Water	31235	94	0	403	80	0	0	89.8%
Shadow	1171	4023	0	57	0	0	0	79.8%
Bare Soil	2060	472	11698	3927	2037	508	0	78.3%
Building	303	451	1522	9657	224	0	0	63.1%
Road	0	0	1726	1261	12567	0	0	84.3%
Tree	0	0	0	0	0	11352	6771	88.5%
Grass	0	0	0	0	0	962	13934	67.3%
Producer's Accuracy	98.2 %	76.6%	56.5%	79.4%	80.8%	62.6%	93.5%	
Overall Classification Accuracy: 79.7% - Kappa: 75.5%								

These results show that machine learning methods seem to be appropriated to acquire relevant classification rules. This method can help to define a knowledge base in an automatic way. However, the performance of the rules could probably be improved. Some classes are spectrally too close. Additional spatial and contextual features should be used. Textural information should be also considered. It can contribute to the improvement of the results [8,9]. Finally, further investigations should be carried out in order to determine if the number of learning examples used is sufficient since the quality of the learnt hypothesis can be affected by the size of the training data set [6].

5. DERIVING CLASSIFICATION RULES FROM MULTIPLE IMAGES

Another approach that seems very promising to improve the classification accuracy is the use of several sources of remotely sensed data in the classification procedure.

The objects identification should be facilitated by taking into account some complementary aspects of a set of images. Objects can be more or less quite visible in an image according to the period during which the image was acquired. Thus, it should be useful to combine images dated from different seasons to improve the interpretation of the objects.

The definition of links between objects from images of different resolutions can also be used to better extract and identify the objects in the images. For instance, that could be the case to recognize the parks in an urban area. The objects composing the parks could be individualised in a VHR-image (e.g. trees, lawns, bushes, alleys, lakes…). In a HR-image, these elements are mixed and the parks could be extracted directly from the image (using textural information). We can imagine therefore defining more precisely the limits and the extent of the parks in the HR-image using the shape and the dimension properties of the elements extracted in the VHR-image. We can also consider to precise the semantic of the elements in the VHR-image knowing the link they have with the parks in the HR-image. The elements could be viewed as being components of a complex object. The spatial relationships between the elements could therefore be considered and the structure of the park could be defined. The experiments presented in this section follow this approach.

We investigated the possibilities of discovering rules of correspondence between several classes belonging to two images: the QUICKBIRD image and a classified LANDSAT image. In particular, we tried to learn rules enable to predict the class of the built-up areas in the LANDSAT image from the classes of the buildings in the QUICKBIRD image. These rules could be used to asses the consistency of the LANDSAT image classification. They could also be used to detect updates in the images since they do not have the same dates. Finally, the definition of these rules constitutes the first step towards the development of a multi-scale database.

The LANDSAT image used is composed of 18 land cover classes. These classes are related to wooded vegetation, agriculture, transportation networks and built-up areas. For this study, we only retained the built-up areas which are subdivided into 4 classes: high density, medium density, low density and industrial. In the QUICKBIRD image, we selected the residential buildings, the collective buildings, the industrial and commercial buildings, and the continuous built-up areas.

A decision tree classifier has been trained to discover links between these classes. The learning examples have been defined interactively by identifying visually the correspondence relationships between the objects of the two images. The label value of each learning example corresponds to one of the classes of the built-up areas of the LANDSAT image. The attribute value of each learning example corresponds to one of the classes of the buildings of the QUICKBIRD image. 30 learning examples have been selected (for each label value). The learnt rules are given below:

```
MULTI-SCALE RULES:

Rule 1: IF Quickbird's building = Residential Building
        THEN Landsat's Built-up Area = low density
Rule 2: IF Quickbird's building = Continuous built-up Area
        THEN Landsat's Built-up area = high density
Rule 3: IF Quickbird's building = Industrial and Commercial Building
        THEN Landsat's Built-up area = industrial
Rule 4: IF Quickbird's building = Collective built-up Area
        THEN Landsat's Built-up area = medium density
```

The predictive accuracy of these rules was 76.4%. This value has been computed by the cross-validation method. The principle of this method is as follows: the set of examples is divided in *k* subsets; the learning process is undertaken on *k-1* subsets and a rate of error is computed on the last subset. This is done for each subset. Then a global evaluation of rules is estimated by the mean of the rates of errors for each subset in terms of predictive accuracy. This method is considered as a good approach to estimate the predictive accuracy of the rules when another training set can not be provided for the validation [3]. It is the case here. We cannot directly apply these rules on the images since the matching procedure is not automated yet. We are now investigating this aspect [10].

The predictive accuracy of these rules is acceptable and satisfies an expert visual evaluation (figure 4). However, other relationships can sometimes appear for the residential and collective buildings. A correspondence between a residential building and a built-up area of medium density is possible and consistent. In some cases, a collective building can also be connected to a built-up area of high density. The training set was composed of this kind of examples but they were considered as noisy data for the learning algorithm. The algorithm only learned the most frequent relationships between the classes. Consequently, it is necessary to revise the rules in this case before using them.

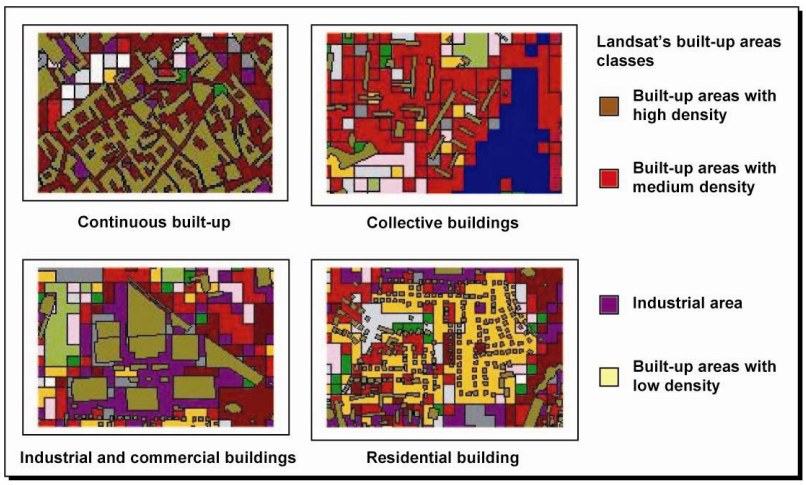

Figure 4. Individual buildings coming from the vector topographical database used for the QUICKBIRD image segmentation, overlaid on the LANDSAT's built-up areas.

6. CONCLUSION

We have presented an approach to classify remotely sensed data using data mining techniques and in particular, decision tree classifier. The experiments have showed that decision tree is a suitable technique to build automatically a knowledge base classification in an object-oriented paradigm. The learnt rules have enabled to classify a QUICKBIRD image with a relatively good performance.

However, the quality of the classification rules could probably be improved using more spatial and contextual features. Nevertheless, making this information relevant supposes to extract homogeneous regions with representative shape and dimension. This depends on the quality of the segmentation procedure. In this study, we have suggested to introduce thematic layers coming from a vector database to constraint the image segmentation.

If additional features are used to characterize the regions, more learning examples shall be defined to train the classifier because the hypothesis space will be greater and more complex. The size of the training set and the complexity of the learning examples can affect the quality of the hypothesis [3].

We make the assumption that the combination of different images coming from several sources can also contribute to improve the image analysis (in particular, the segmentation and the classification procedures). The experiments related to the acquisition of matching rules between classes of the QUICKBIRD and LANDSAT images are a first step towards the validation of this assumption. The learning approach we performed was intended to acquire rules enabling to predict the classes of the regions of one image from the classes of the regions of another image. This approach seems to be promising. It could be used to assess the consistency between multiple image classifications. It could also be followed to build a multresolution database. The research prospects for making this approach operational concern the image matching procedure. Methods for automating the relationships computation have to be defined. Our efforts tend to this direction.

7. ACKNOWLEDGEMENTS

This study is a part of the research project FoDoMuST (multi-strategies data mining to extract and identify urban elements from remote sensing images database) which is financed by the ACI "*Masses de Données*" (2004-2007).

REFERENCES

1 Herold M., Scepan J., Muller A., Gunter S. 2002. Object-oriented mapping and analysis of urban landuse/cover using Ikonos data. In: 22nd Earsel Symposium "Geoinformation for European-wide integration".

2 Benz U.C., Hofmann P., Willhauck G., Lingenfelder I. and Heynen M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information, ISPRS Journal of Photogrammetry & Remote Sensing, 58: 239-258.

3 Mitchell T.M. 1997. Machine Learning (McGraw-Hill International Editions), 414 p.

4 Quinlan J.R. 1993. C4.5: Programs for machine learning (Morgan Kaufmann), 302 p.

5 Friedl M.A. and Brodley C.E. 1997. Decision tree classification of land cover from remotely sensed data, Remote Sensing of Environment, 61: 399-409.

6 Pal M. and Mather P.A. 2003. An assessment of the effectiveness of decision tree methods for land cover classification, Remote Sensing of Environment, 86: 554-565.

7 Baatz M. and Schäpe A. 2000. Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. In: 12th Angenwandte Geographische Informationsverarbeitung, edited by Strobl J., Blaschke T., Griesebner G., 12-23.

8 Debeir O., Van den Steen I., Latinne P., Van Ham P. And Wolff E. 2002. Textural and Contextual land-cover classification using single and multiple classifier systems, Photogrammetric Engineering and Remote Sensing, 68: 597-605.

9 Puissant A., Hirsch J. and Weber C. 2005. The utility of texture to improve per pixel classification for high spatial resolution imagery, International Journal of Remote Sensing, 26: 733-745.

10 Vauglin F. and Bel Hadj Ali A. 1998. Geometric matching of polygonal surfaces in GIS. In: ASPRS-RTI Annual Conference, 1511-1516.