GENETIC ALGORITHMS AND DATA MINING APPLIED TO OPTICAL ORBITAL AND LIDAR DATA FOR OBJECT-BASED CLASSIFICATION OF URBAN LAND COVER

F. Leonardi [a], C. M. Almeida [b, *], L. M. G. Fonseca [c], L. R. Tomás [d], C. G. Oliveira [b]

[a] Geopixel Ltda., R. M. Egydio Pinto, 165 – Cj. 12, 12245-190, São José dos Campos, Brazil - fernando@geopx.com.br
[b] INPE, Division for Remote Sensing, 12227-010 São José dos Campos – SP, Brazil - (almeida,cleber)@dsr.inpe.br
[c] INPE, Division for Image Processing, 12227-010 São José dos Campos – SP, Brazil - leila@dpi.inpe.br
[d] Institute for Research, Administration and Planning - IPPLAN, R. Augusto E. Ehlke, 181, 12243-110, São José dos Campos, SP, Brazil - liviatomas@gmail.com

KEY WORDS: Laser Scanning, Decision Tree, Semantic Network, Semi-Automated Classification

ABSTRACT:

The study of the urban environment has raised great interest among researchers and practitioners involved with the use of remote sensing, in face of the challenges for its investigation and the complexity of its targets. Although they have great potential for studies of urban environments, the high-resolution images present difficulties for automatic extraction of information because they are characterized by high spatial and spectral heterogeneity for the same segment, which greatly complicates segmentation and classification processes. Thus, new concepts and analyses have been used for mapping the urban space. Object-based image analysis and multiresolution segmentation have been quite efficient in the discrimination of urban targets in high spatial resolution images. One technique that can assist the classification process is data mining, which can be used to explore large data sets, identify and characterize patterns of interest, and hence, support the precise extraction of useful information. In this context, this paper proposes a methodology jointly employing cognitive approaches (semantic net, object-based image analysis) and data mining (genetic algorithms and decision trees) for the classification of urban land cover from optical orbital and airborne laser data. To assess the efficacy of the methodology and ensure the accuracy of the produced maps, the steps undertaken in this study were subject to quality control. The results were presented and discussed, indicating a satisfactory accuracy in the generated mapping products, demonstrating the reliability of the methodology for mapping land cover in urban areas.

1. INTRODUCTION

Urban areas are dynamic systems of great complexity, for they materialize the results of human action over the natural environment. In this sense, it is crucial that maps of these areas be elaborated and continuously updated. The information extracted out of these products can be used to guide medium- and long-term investments planning of a given municipality, monitor its increasing demands for technical infrastructure and social equipments, besides supporting the elaboration of public policies in compliance with environmental guidelines and targeted to provide a sound quality of life to its inhabitants.

Aerophotogrammetric survey is one of the oldest and most traditional sources for the generation of such maps. However, this a costly procedure, executed on demand, what renders remote sensing a more advantageous source of information, in face of its synoptic view, systematic acquisition and comparatively lower costs.

More in-depth and thorough studies of urban areas could only be realized after the advent of high spatial resolution images in the year 1999, with the successful launching of the first high spatial resolution satellite - IKONOS II. Information extraction on these images were based either on manual or semi-automatic methods. A great number of automatic and semi-automatic classification methods has been developed since the release of the first images acquired by orbital remote sensing. Nevertheless, the automatic and/or semi-automatic classification of urban land cover/land use in high spatial resolution images poses new challenges.

Experiments conducted by Pinho (2005) and Araújo (2006) demonstrated that both traditional and object-based image classification methods result in confusion between classes with similar spectral behavior, as for instance, French tiles and clay bare soil, asphalt and dark asbestos-cement tiles, as well as trees and grass vegetation. Considering that these confusing classes present distinct elevation values, this work was committed to insert high accuracy elevation data obtained by airborne laser scanning in the classification experiments, so as to minimize confusion between them.

It is though worth stating that increasing accuracy is not the only goal of an object-based image analysis, but speeding up the definition of optimal segmentation parameters and the elaboration of a semantic network through automation is as well envisaged. For this purpose, genetic algorithms have been used to assess the best set of segmentation parameters and a decision tree algorithm has been employed to explore the input dataset in order to unravel patterns that could be of use in the generation of the knowledge model.

In brief, this paper proposes a methodology jointly employing cognitive approaches (semantic net, object-based image analysis) and data mining (genetic algorithms and decision trees) for the classification of urban land cover from optical orbital and airborne laser data. In the remainder of this paper, a brief overview on the study area is provided in Section 2. Section 3 describes the data acquisition, pre-processing and methods adopted in this work, and Section 4 presents the results followed by a critical evaluation on the potential and drawbacks of the input data and methodological procedures in Section 5.

* Corresponding author.

Finally, some conclusive remarks and directions for future work are drawn in Section 6.

2. STUDY AREA

The selected study area concerns a central sector of Uberlandia city, located in the southeastern State of Minas Gerais, Brazil (Fig. 1a and 1b). The city is located 550 km away from Belo Horizonte, the state capital, and has the following coordinates 18º 55' 07" S and 48º 16' 38" W. Uberlandia city presents a cluster of high-rise buildings in its central neighborhoods, within which the study area is contained (Fig. 1c and 1d). The municipality had a population of 608,369 inhabitants in 2007. The city itself is located on a mildly undulated terrain, with a mean altitude of 1,000 m above sea level and it presents a hot and tropical climate, with a mean annual rainfall of 1,500 mm and an average temperature of 22 ºC.

The occupation patterns of the central neighborhoods in Uberlandia are very diverse, comprising green areas, one- and two-storey buildings as well as high-rise residential and business buildings. The urban land cover materials found in the area are manifold and can be approximately categorized in: trees, grass, light bare soil, dark bare soil, light French tiles, dark French tiles, metallic roofs, light concrete decks, light asbestos cement tiles, medium to dark concrete decks or asbestos cement tiles, swimming-pool, and asphalt.

3. METHODS

3.1 Data Acquisition and Pre-processing

The data used comprised: i) five IKONOS II images, of which four correspond to multispectral bands (B, G, R, NIR) with 4,0 m of spatial resolution, and one is panchromatic, with 1,0 m resolution, all of them acquired on June 27, 2008, with 11 bits of radiometric resolution and a viewing angle of 8.2º; ii) a digital surface model (DSM) and a digital terrain model (DTM) obtained by a laser scanning air survey with the ALTM 2025, with a raw elevation accuracy of 0.15 m; iii) a digital height model (DHM), generated from the subtraction between the DSM and DTM; iv) a vector file in shape format, containing the streets network of Uberlandia city, given by its Municipal Government; and v) 55 GPS points collected in rapid static mode with GPS Hipper.

Initially, the four IKONOS II multispectral bands were pan-sharpened with the respective panchromatic band using the HIS method with the cubic convolution interpolator. Among several pan-sharpening methods which have been evaluated, the HIS proved to yield the best results. After this procedure, the image was subject to an orthorectifying process, based on GPS points collected in the field and evenly distributed over the study area. In total, 55 GPS points were collected, with an approximate planimetric (horizontal) accuracy of 0.030 m, and an elevation (vertical) accuracy of nearly 0,021 m. All of the points were processed using the UTM projection, South Zone 22, Datum WGS 84, based on the MGUB and UBER stations from the Brazilian Network for Ongoing Monitoring.

After the acquisition and processing of field data, the image was orthorectified in absolute mode, using such GPS points, the sensor attitude and ephemeris data (rendered available in the image metadata files), and the elevation data derived from the laser scanning air survey accomplished with ALTM 2025.

Fig. 2a shows the spatial distribution of 25 tie points (extracted ouf of the 55 points collected in the field) used for orthorectifying the IKONOS II image. The great majority of selected points are placed over the central portion of the scene, precisely where the study area is located, so as to assure a better geometric fit in this region.

Before executing the orthorectification, the elevation values obtained with the GPS points, which actually correspond to geometric heights (related to the reference ellipsoid) had to be converted in orthometric heights, which are those related to the Earth´s geoid (or the mean sea level).

The coordinates obtained after the field data processing were exported to a software named MAPGEO, which uses the planimetric (horizontal) coordinates to extract the value of the undulation of the geoid (difference between the reference ellipsoid and the Earth´s geoid) in the surveyed points.

After this procedure, the image was finally orthorectified using the Rational Function Model (RFM), the LiDAR-derived DTM, and the GPS points. The orthoimage was generated using the bilinear interpolator and was referenced in the UTM projection, Datum WGS 84.

The orthoimage represents a product corrected in relation to distortions caused by the sensor tilt, Earth curvature and ground relief, and hence, it has been actually used in the urban land cover classification. The orthorectified image was further subject to statistical validation procedures, based on the selection of a different set of GPS points positioned in easily identifiable locations in the scene. None of such points has been previously used for orthorectifying the image, so as to avoid trend errors in the validation process.

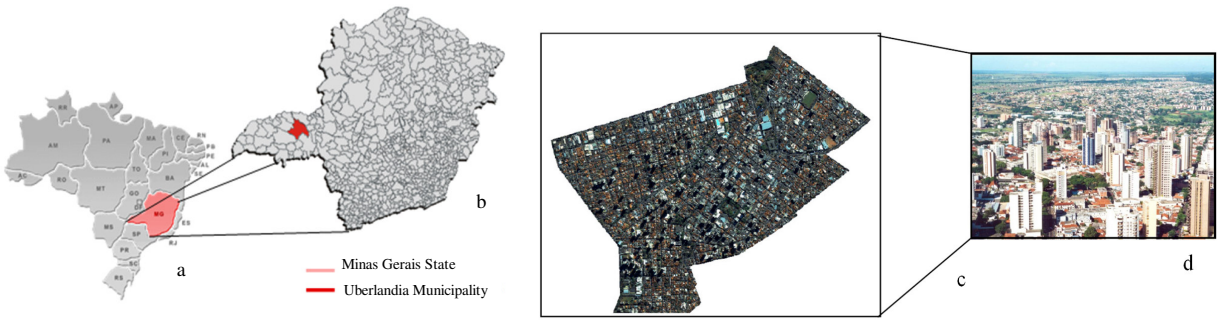


Figure 1. Study area: a. Brazil b. Minas Gerais State and Uberlandia Municipality (in red) c. Study area (central sector of Uberlandia city) in true color composition of IKONOS II images (1B_2G_3R) d. Aerial view of study area

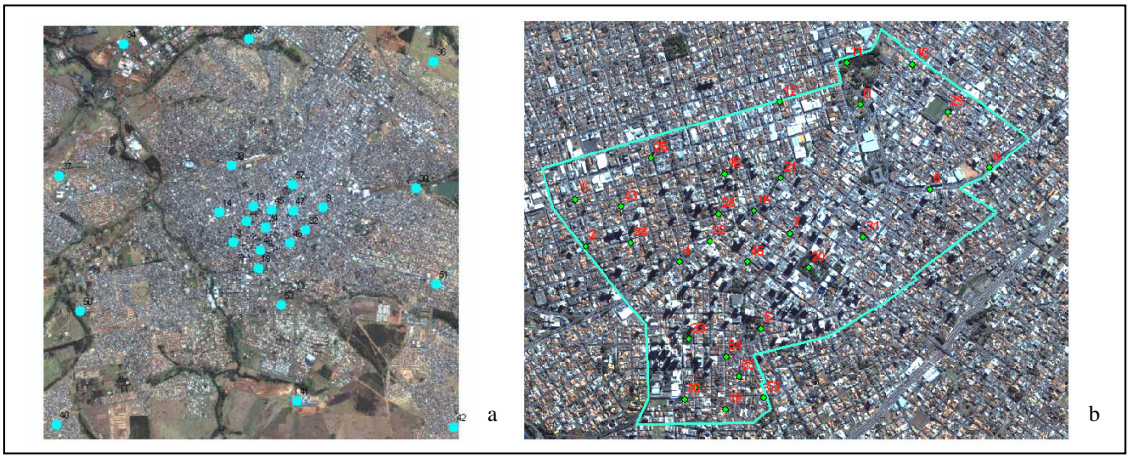


Figure 2. IKONOS II image orthorectification: a. The entire scene comprising the study area, with tie points (blue dots) used for orthorectifying the image, with their respective ID number in black b. Zoom in the scene, showing the study area boundaries (light blue line) and the validation points (green dots) with their respective ID number in red

Fig. 2b illustrates the spatial distribution of points used for validating the orthoimage. According to a method proposed by Galo and Camargo (1994), the validation tests are based on a 10% level of significance and they comprise both trend and precision analyses. The trend analysis is based on a t-Student distribution and refers to an analysis of discrepancies between the observed coordinates in the cartographic product and the reference coordinates, calculated for each sample point. The precision analysis, on its turn, concerns the comparison of the standard deviation related to the discrepancies with the expected standard deviation for the desired class by means of a hypothesis test.

The LiDAR DSM data were processed with the module TerraScan of TerraSolid software through Axelsson's progressive TIN densification algorithm, which extracts points directly located on the terrain surface by constructing an iterative TIN. This network was then converted in a regular grid representing the study area DTM.

The same statistical tests applied for validating the orthoimage were as well applied in the validation of the LiDAR-derived DSM and DTM. According to USGS guidelines, the minimum amount of checking points for calculating the root mean square error (RMSE) of a DEM (DTM or DSM) is 28, out of which 20 must be located in the central part of it and 8 in its borders.

Out of the 55 GPS points collected in rapid static mode in the field, 42 of them were actually used for assessing the DSM and DTM accuracies, since they were located in easily identifiable places, well distributed over the scene and lying within the study area boundaries (Fig. 3a. and 3b). The subtraction between the DSM and DTM generated the digital height model (DHM) (Fig. 3c), which has been effectively used for the classification purposes reported in this paper.

3.2 Definition of Segmentation Parameters Using Genetic Algorithms

The segmentation parameters for the object-based land cover classification were defined in a plug-in named Segmentation Parameters Turner (SPT), developed by the Lab for Computer Vision of the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. SPT uses a genetic algorithm to identify optimal values, within a given search space, for all segmentation parameters required by the OBIA platform Definiens Developer, i.e. scale factor, weights for each input band, color and shape parameters as well as compactness and smoothness parameters. The most satisfactory value is determined by an objective-function which assesses the

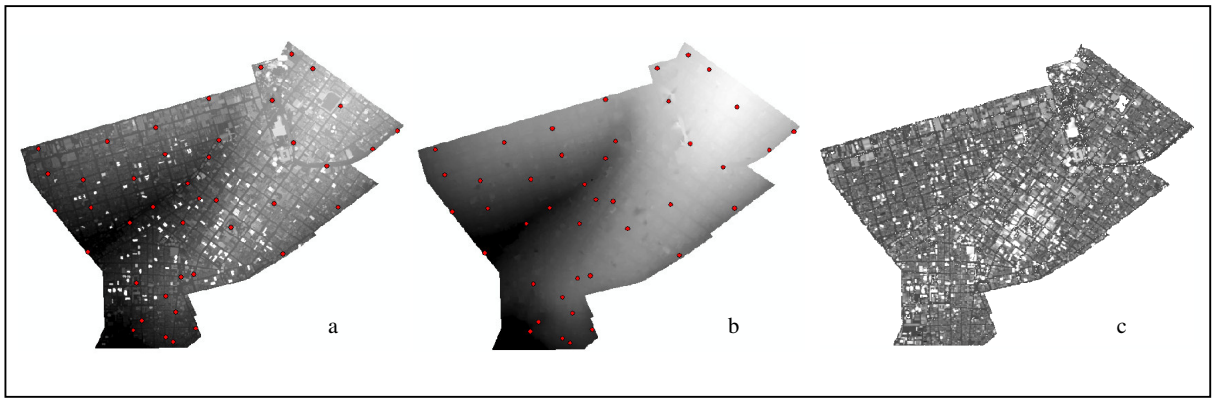


Figure 3. a. Digital surface model (DSM) of the study area with validation points (red dots) b. Digital terrain model (DTM) of the study area with validation points (red dots) c. Digital height model (DHM) of the study area

degree of agreement between the segmentation results and the reference samples, consisting of segments manually delimited by the interpreter (Fig. 4a.). In mathematical terms, given a set of reference segments S, a disparity function F, and a vector of parameters P, the genetic algorithm (GA) aims at finding the optimal set of segmentation parameters P_{opt} defined by Equation 1 (Fredrich and Feitosa, 2008):

$$P_{opt} = \arg_P \min ([F(S,P)]) \; . \qquad (1)$$

The user has to define the GA internal parameters, like number of experiments, population size, number of generations, initial and final gap (Fig. 4a). Nevertheless, SPT has as its default parameters values that have proved to generate the best results. Data are expected to converge after a certain amount of generations (Fig. 4 b), producing a segmentation at the end (Fig. 4c). Several disparity functions have been implemented in SPT. In this particular work, the *Larger Segments Booster* - LSB function has been used. LSB favours results that closely match the reference samples with the minimum possible number of segments. Let S_i be a set of pixels belonging to i-th segment of S, $SO(P)$ the set of segments produced by the segmentation that contains pixels of S, and $SO(P)_i$ the set of $SO(P)$ which members own at least 50% of theirs pixels in S_i. It is worth stating that:
- the number of pixels in $SO(P)_i$ that do not belong to S_i are considered *fpi* (false positive or omission errors);
- the number of pixels in $SO(P)_i$ that also belongs to S_i are considered *vpi* (trye positive);
- the number of pixels in S_i that do not belong to $SO(P)_i$ are considered *fni* (false negative or comission errors);
- the number of pixels in the boundaries of $SO(P)_i$ contained within S_i - or internal boundaries of references - as b_i; and
- the number of empty $SO(P)_i$ as *NS*.

The LSB function is given by Equation 2:

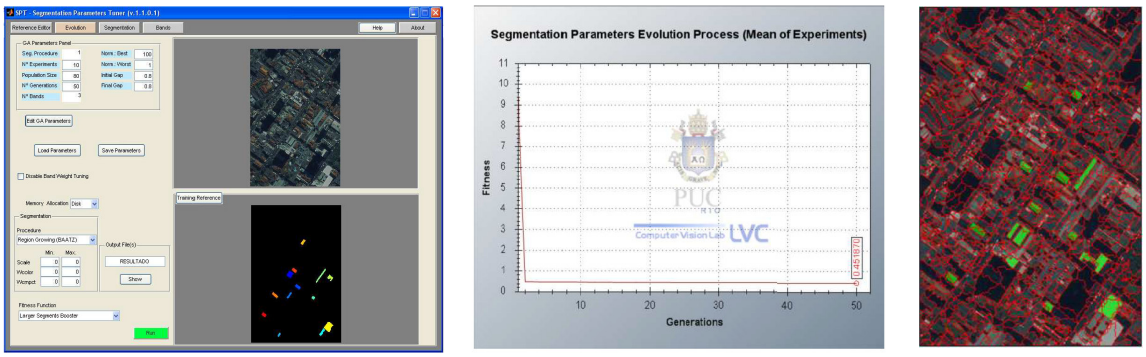
$$F(S,P) = 1/n \left[NS + \sum_{SO(P)_i \neq 0} \frac{fp_i + fn_i + b_i}{\# S_i} \right]. \qquad (2)$$

3.3 Data Mining for Automating the Elaboration of the Semantic Network

The semantic network in this object-based urban land cover classification was elaborated through data mining using the C4.5 algorithm, created by Quinlan (1993) and implemented as the tree.J48 classifier in the data mining platform WEKA, developed by the University of Waikato, New Zealand. This algorithm builds decision trees based on training samples and through a recursive procedure of data partitioning. The trees are expressed as a flowchart, where each internal node executes a test with a given attribute, the branch (or arc) represents the test result, and the external node (or leaf) accounts for the expected class. For each node, the algorithm chooses the best attribute to separate the data in individual classes. The attributes that are not included in the tree are regarded as irrelevant. If the number of samples and/or their class descriptive ability are not appropriate, the decision tree will incorrectly classify many objects. Big trees tend to data overfitting, while very small trees end up by missing important attributes of the data. The algorithm always strives to produce less complex and smaller trees, for they are more easily understandable and show a better performance. For this end, it uses entropy in order to assess to what extent the node is informative. Small entropy values mean that less information will be used to describe the data.

4. EXPERIMENT DESIGN: OBJECT-BASED CLASSIFICATION

The optimal parameters provided by SPT drove the segmentation of the four pan-sharpened IKONOS II image bands in the Definiens Developer 7.0.4. This segmentation level was used to collect the samples for the decision tree training in the WEKA platform. Eleven classes of urban land cover were defined: bare soil, light French tiles, dark French tiles, metallic roofs, swimming pool, light concrete deck or light asbestos cement tiles, medium to dark concrete deck or medium to dark asbestos cement tiles, asphalt, shadow, trees, and grass. All attributes existent in the Definiens Developer platform together with customized ones related to arithmetic transforms of image bands were added to the training set, what totalized 355 attributes, exported as file with CSV extension. In WEKA, the input dataset was subject to a preliminary filtering for removing noise and inconsistencies. The number of training samples per class was very diverse, but they tried to be representative of the spectral and textural heterogeneity of the concerned classes. A minimum of 55 objects (sample units) was set to be considered during the decision tree processing.
After a certain number of consecutive training experiments in WEKA, the final decision tree was produced (Fig. 5), considering only five attributes out of the 355 initially selected. This decision tree was then converted into a hierarchical semantic network inside the Definiens Developer 7.04 platform.



a b c

Figure 4. a. Genetic algorithms parameters definition (left), inset of IKONOS II image (upper right corner) and respective reference samples in different colors (bottom right corner) of SPT GUI b. Fitness curve showing convergence c. Segmentation results

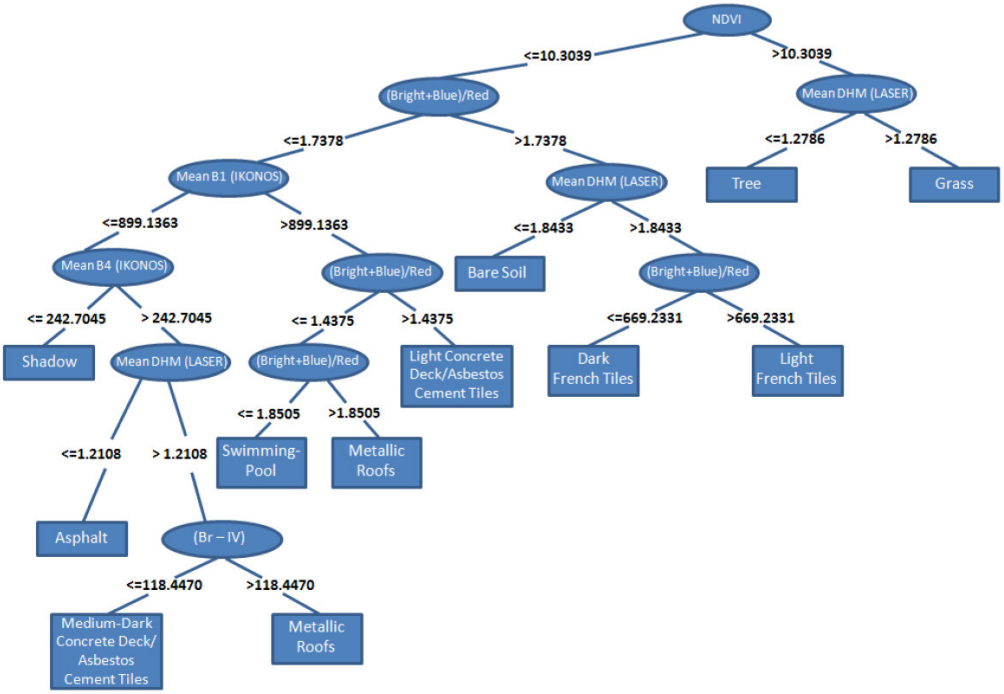


Figure 5. Decision tree generated with the C4.5 algorithm for the study area, considering optical orbital data from IKONOS II images and a DHM extracted from laser scanning

Figure 6. Object-based classification of urban land cover for the study area and its respective hierarchical semantic network derived from the generated decision tree

CLASSES		References											Total
		A	B	C	D	E	F	G	H	I	J	K	
Classification	A - Swimming Pool	**1**	0	0	0	0	0	0	0	0	0	0	**1**
	B - Bare Soil	0	**8**	0	0	0	3	0	2	1	0	4	**18**
	C - Trees	0	0	**39**	7	0	0	0	0	0	1	0	**47**
	D - Grass	0	0	4	**12**	0	0	0	0	0	0	0	**16**
	E - Light French Tiles	0	0	0	0	**11**	4	0	0	6	0	0	**21**
	F - Dark French Tiles	0	0	0	0	0	**40**	0	1	3	3	10	**57**
	G - Metallic Roofs	0	0	0	0	0	0	**18**	2	2	4	4	**30**
	H - Asphalt	0	1	0	0	0	0	1	**62**	5	3	8	**80**
	I - Light Concrete Deck/Asbestos Cement Tiles	0	0	0	0	0	0	0	1	**30**	7	0	**38**
	J - Medium to Dark Concrete Deck/Asbestos Cement Tiles	0	0	2	0	0	5	0	6	4	**94**	6	**117**
	K - Shadow	0	0	0	0	0	0	1	7	0	0	**67**	**75**
	Total	**1**	**9**	**45**	**19**	**11**	**52**	**20**	**81**	**51**	**112**	**99**	**500**
	Global Accuracy: 0.7640	**Kappa Index : 0.7344**							**Variance of Kappa: 0.00045**				

Table 1. Error matrix and agreement indices for the object-based classification of urban land cover

5. RESULTS AND DISCUSSION

In the trend analysis, the validation of the orthoimage showed that it presented a mean error of 0.0071 m in the E component, and 0.0008 m in the N component, without trends in both components, and a standard error of 0.66m, compatible with a 1:2,000 scale. In the precision analysis, the results also confirmed the orthoimage is up to the highest cartographic accuracy standard of a 1:2,000 scale.

The statistical tests for the DSM and DTM demonstrated that the mean elevation error lied around 0,41 m and 0,48m, and the RMSE about 0,48 m and 0,47 m, respectively. In both cases, the presence of trend in the H direction was observed, revealing systematic influences in this component. This trend was further removed by means of algebraic manipulations. The precision analysis revealed that the DSM and DTM were up to the highest cartographic accuracy standard of a 1:5,000 scale.

As to the optimal segmentation parameters, SPT provided a scale factor of 11, a color factor of 0.57, a compactness factor of 0.64, and a weight of 0.06 for the blue band, of 0.52 for the green band, and of 0.42 for the red band.

The generated decision tree showed a concise and logical structure, relying only on five attributes: the Normalized Difference Vegetation Index (NDVI); the ratio of the sum of the mean of brightness and the mean of the blue band to the mean of the red band; the mean of the DHM; the mean of the blue band; the mean of the near infrared band; and the difference between the mean of brightness and the mean of the near infrared band. There was only one repetition of a leaf node, namely the class metallic roofs, what is considered a very satisfactory result. It is worth remarking that the mean of the DHM has been used to differentiate trees from grass, bare soil from dark and light French tiles, and medium to dark concrete deck or asbestos cement tiles from asphalt, as expected.

The classification result generated from the application of this decision tree is shown in Fig. 6. For validating this classification, 500 random points, collected through stratified sampling based on the share of the expected areas of each class, were taken into account for assessing the global accuracy and the Kappa index (Congalton, 1991). The validation results are presented in Table 1, indicating a reasonable amount of omission errors of shadow, which has been in some cases wrongly classified as dark French tiles, asphalt and medium to dark concrete deck or asbestos cement tiles. There were also meaningful commission errors of shadow, to which asphalt objects have been assigned. The global accuracy achieved 76% and the Kappa index reached 73%, what is regarded as a very good accuracy accordingto a ranking of Landis and Koch (1977).

6. CONCLUSIONS

The use of the genetic algorithm routine implemented in SPT showed a satisfactory performance for the automatic assessment of the optimal segmentation parameters. Nevertheless, the shape complexity of some targets, the internal spectral variability of certain classes, and the diverse conditions of ageing and maintenance of some roof classes found in the study area led to an over-segmentation of some targets.

In face of the massive number of available attributes (spectral, geometric, topological, textural) in the Definiens platform, the data mining techniques proved to be crucial for handling and exploring this great amount of information. Adding height information derived from laser scanning to the multispectral images helped in the discrimination between targets with similar spectral behavior but diverging elevation values. It is worth mentioning that all methodological stages in this work were subject to a quality control, aiming at assessing the quality and accuracy obtained by each generated product. The employed methods demonstrated to be applicable for the classification of urban land cover.

REFERENCES

Araújo, E. H. G., 2006. *Multitemporal analysis of QuickBird scenes*. Master´s Thesis. INPE, São José dos Campos, Brazil.

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

Fredrich, C. M. B., Feitosa, R. Q., 2008. Automatic adaptation of segmentation parameters applied to inhomogeneous objects detection. In: *GEOBIA 2008*, Alberta, Canada.

Galo, M., Camargo, P. O., 1994. Utilização do GPS no controle de qualidade de carta. In: *Congresso Brasileiro de Cadastro Técnico Multifinalitário*, Florianópolis (SC), Brazil.

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): pp. 159-174.

Pinho, C. M. D., 2005. *OBIA applied to high spatial resolution images*. Master´s Thesis. INPE, São José dos Campos, Brazil.

Quinlan, R., 1993. *C4.5: programs for machine learning*. Morgan Kaufmann, San Francisco.