

# Evaluating the Performance of Decision Tree Algorithms CART and C4.5 for an Object-Based Urban Land Cover Classification

Paulo Roberto da Silva Ruiz<sup>a1</sup>, Cláudia Maria de Almeida <sup>a</sup>, Camila Souza dos Anjos Lacerda <sup>b</sup>, Thales Sehn Körting <sup>a</sup>, Leila Maria Garcia Fonseca <sup>a</sup>

<sup>a</sup>National Institute for Space Research - INPE, São José dos Campos, SP, Brazil

<sup>b</sup>Institute for Advanced Studies - IEAV, São José dos Campos, SP, Brazil

Received on August 01, 2016 / accepted on \*\*\*\*\*, 2016

## Abstract

This paper aims at evaluating the performance of two decision trees generated by data mining using the algorithms CART and C4.5 for a region-based classification of urban land cover. A pan-sharpened WorldView-2 image was used for classification, from which statistical attributes were extracted. The study area concerns a high standard residential neighbourhood in Campinas city, located in Sao Paulo State, southeast of Brazil. The Kappa index obtained by the classification based on CART algorithm was 0.70, while the result using C4.5 algorithm achieved 0.75. The structures of the two generated decision trees are quite alike, although the employed statistical attributes differ from one another. Considering the attained Kappa indices, one can state that both classifications presented satisfactory and similar quality accuracies.

**Keywords:** Remote sensing, images classification, data mining.

## 1. Introduction

The acquisition of information about urban areas from high resolution remote sensing data has continuously increased in recent years. New data sources are required to meet the demands of several urban studies, with a direct impact on the planning of cities and, consequently, on the daily life of its citizens. The techniques meant for such data processing are of crucial importance, since they are responsible for the systematic generation of land cover and land use maps.

The launch of satellites with high resolution sensors introduced new possibilities for the remote sensing of the urban environment, but imposed at the same time numerous challenges. These sensors have substantially improved the images spatial resolution. However, the advancement in spatial

---

<sup>1</sup>E-mail Corresponding Author: paulo.ruiz@inpe.br

38 and spectral resolutions ended up by hampering the classification of urban  
39 targets, especially in the traditional pixel per pixel approaches [1] [2]. Ad-  
40 ditionally, the refinements in the resolutions of the multi and hyperspectral  
41 images require better computational resources for their processing, handling  
42 and storage.

43 An approach to ease the automation of image processing is data mining.  
44 This technique allows the exploration of a data set, in order to highlight pat-  
45 terns of interest that assist in the generation of knowledge [3]. Experiments  
46 conducted by [4] show that data mining techniques applied to the detection  
47 of changing patterns in remote sensing images can provide good results.

48 Considering that data mining techniques have proven to be robust for  
49 the handling of massive data set, this work aims to evaluate the performance  
50 of two data mining approaches, the decision tree algorithms Classification  
51 and Regression Trees (CART) and C4.5, for an object-based urban land  
52 cover classification relying on statistical attributes.

53

54

## 55 2. Methodology

56

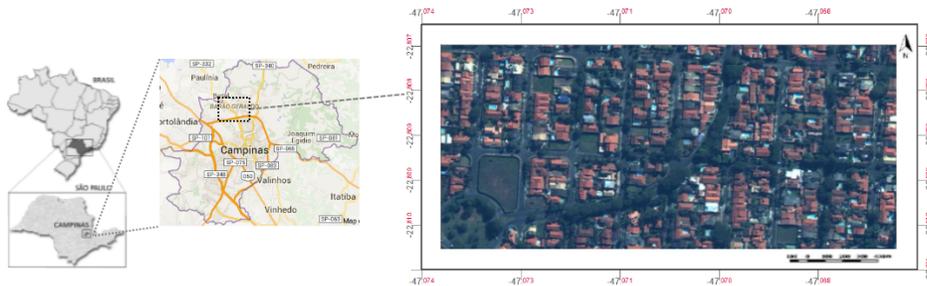
### 57 2.1 Materials

58 A multispectral scene of the WorldView-2 sensor was employed for the  
59 analysis. The multispectral data own 1.85 m of spatial resolution operating  
60 in eight spectral bands (coastal blue, blue, green, yellow, red, red edge, near-  
61 infrared-I, near-infrared-II) and were acquired in July of 2010. They were  
62 used in combination with a panchromatic image with spatial resolution of  
63 0.50 m. The image comprise a small study area located in the vicinities  
64 of the Campinas State University campus, in the southeastern state of São  
65 Paulo, Brazil (Figure 1).

66

67

68



69 **Figure 1** - Study area: Neighborhood in the vicinities of the Campinas  
70 State University campus, São Paulo state, southeast of Brazil.

69 The images preprocessing was made in ENVI [5] version 4.7. For the  
70 segmentation process and land cover classification, eCognition Developer [6]  
71 version 8.7, was used. Samples of the land cover classes were extracted in  
72 eCognition and were further exported to the software WEKA 3.6.12 [7] for  
73 executing the data mining process.

74

## 75 **2.2 Methods**

76

### 77 **2.2.1 Preprocessing**

78 The multispectral images were pansharpened with the panchromatic im-  
79 age by means of the Gram-Schmidt method in ENVI 4.7. This processing  
80 begins by simulating the existence of a panchromatic band through the  
81 multispectral bands. A transformation of Gram-Schmidt is applied and the  
82 simulated panchromatic band is employed as the first band of the result-  
83 ing set, being later replaced by the panchromatic band. Finally, an inverse  
84 transform is applied to provide the fused image [8].

85

### 86 **2.2.2 Segmentation and Features Extraction**

87 The first step in image classification is the segmentation process, which  
88 consists in partitioning the image into regions that show some degree of ho-  
89 mogeneity with respect to their content and the purpose of classification. In  
90 this work, the multiresolution segmentation was used, that includes several  
91 parameters defined by the user, like compactness, shape and scale factor  
92 [9]. In this work, these parameters have been empirically defined, so as to  
93 produce image regions whose boundaries were as close as possible to the  
94 boundaries of the targets of interest.

95 To achieve better results, land cover classes were defined prior to seg-  
96 mentation. This procedure was based on visual interpretation of the pan-  
97 sharpened Worldview-2 image, trying to identify the main materials used  
98 in paved roads and buildings roofs and also the main types of natural fea-  
99 tures, such as vegetation, for example. In this step, 11 classes were defined:  
100 asphalt, dark gray roof, light yellow roof, dark yellow roof, dark ceramic  
101 roof, light ceramic roof, swimming-pool, yellow quartzite, grass, trees and  
102 shadow. After segmentation, representative samples of classes were acquired  
103 based on the visual identification of segments associated with different types  
104 of land cover. The samples file, exported in CSV format, comprised the fol-  
105 lowing attributes: mean, standard deviation, minimum and maximum pixel  
106 values, brightness and maximum difference (max. diff.) of each segment.  
107 To calculate the max. diff., the minimum mean value belonging to an object

108 is subtracted from its maximum value. To get the maximum and minimum  
109 value of all layers, the objects are compared with each other. Subsequently,  
110 the result is divided by brightness [10].

111

## 112 **2.2.3 Data Mining and Decision Tree Algorithms**

113 For generating the decision tree, the algorithms CART and C4.5 imple-  
114 mented in WEKA 3.6.12 were used. Decision trees are diagrams built in  
115 sequence, which make use of inductive learning. Such algorithms help in  
116 decision making for solving a given complex problem. They work in a recur-  
117 sive way so as to generate a tree-based data structure that aids in sorting  
118 and classifying unknown samples [11]. The characteristics of the decision  
119 tree algorithms used in this work will be explained next.

120

### 121 **2.2.3.1 Classification and Regression Trees (CART)**

122 The CART method is technically known as binary recursive partitioning.  
123 The process is called binary, because the parents are always divided exactly  
124 into two child nodes, and recursively, because the process can be repeated  
125 treating each child node as a parent node. The main stages of CART are:  
126 define the set of rules for dividing each tree node; decide when the tree is  
127 complete; associating each terminal node to a class or to a predictive value  
128 for the regression [12].

129 In order to split a node into two child nodes, the algorithm will always  
130 ask questions that only admit as answers "Yes" or "No". The next step is  
131 to sort each rule of division based on quality criteria. The standard criteria  
132 used for sorting is the Gini Index, which is based on the calculation of  
133 entropy [13].

134 In the CART procedure, instead of determining when a node is terminal  
135 or not, the algorithm continues expanding the tree until it is no longer  
136 possible to do it, such as when the minimum number of samples in a leaf  
137 node is achieved. After all the terminal nodes are found, the tree finally  
138 acquires its maximum size [12].

139 The sorting algorithm used in the decision tree was the tree.SimpleCart,  
140 a Java transcription of the original CART algorithm implemented in WEKA  
141 3.6.12.

142

### 143 **2.2.3.2 C4.5 Algorithm**

144 The algorithm C4.5 is the result of a paper published in 1993 by Quinlan  
145 [11]. It is different from the CART because it is not mandatory to do a  
146 binary division, what leads to smaller trees. Trees with these characteristics

147 are more easily understood and prone to have greater accuracy, and so the  
148 algorithms strive generating trees as small as possible.

149 The algorithm C4.5 also builds decision trees from training samples. The  
150 trees are expressed by a flowchart, where the internal node represents a test  
151 with the attribute, the branch represents the result of the test, and the ex-  
152 ternal (leaf) node displays the expected class. For each node, the algorithm  
153 chooses the best attribute to separate the data in individual classes. The  
154 attributes that do not appear in the tree are irrelevant. For the construc-  
155 tion of smaller trees, the algorithm uses the entropy to measure the extent  
156 to which the node is informative. Low values of entropy indicate that less  
157 information will be used to describe the data [4].

158 In sum, the algorithm C4.5 is considered as the one that provides the  
159 best result in the assembly of decision trees from a set of training data. It  
160 is implemented as the classifier tree.J48 in WEKA 3.6.12

161

#### 162 **2.2.4 Classification of Urban Land Cover**

163 Decision trees generated in WEKA were implemented in eCognition De-  
164 veloper, by converting the decision rules supplied by the trees in crisp thresh-  
165 olds, generating classifications based on semantic networks containing sta-  
166 tistical descriptors (attributes). Decision trees were generated from data  
167 mining by means of the CART and C4.5 algorithms.

168

#### 169 **2.2.5 Evaluation of the accuracy of land coverage classification**

170 For the accuracy assessment of classifications, 140 random samples were  
171 collected in the image. The samples refer to pixels, precisely to avoid the  
172 bias that would occur if they were selected samples segments, given the  
173 observed variability in the size of the segments. Due to the reduced area of  
174 some classes, it has not been possible to observe significant points similarly  
175 for all 11 classes.

176 Matrices of error were them built for the obtained classifications. These  
177 matrices indicate errors of omission, i.e. samples that have not been classi-  
178 fied according to the reference classes, and commission errors, which relate  
179 to samples erroneously classified as belonging to other classes. The fol-  
180 lowing indices were calculated: global accuracy, producer's accuracy, user's  
181 accuracy, Kappa [14].

182

183

184

185

### 3. Results and Discussion

#### 3.1 Database and Segmentation

The pansharpened WorldView-2 image is shown in Figure 2-A. The segmentation process was performed by the multiresolution algorithm in eCognition, with compactness set to 0.65 and shape to 0.2, and the default weight of 1 being applied to all multispectral bands. The same segmentation was used for the two classifications. The result of the segmentation can be seen in Figure 2-B.

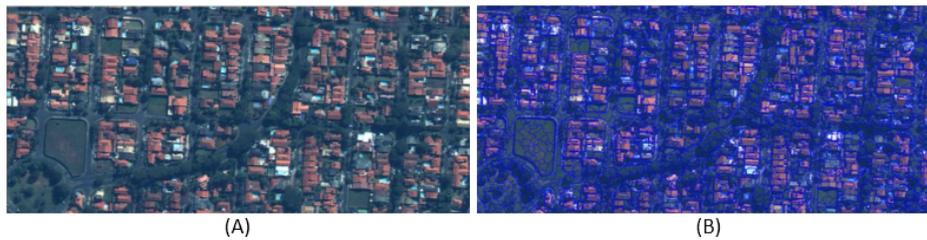


Figure 2 - (A) Pansharpened image and (B) segmentation.

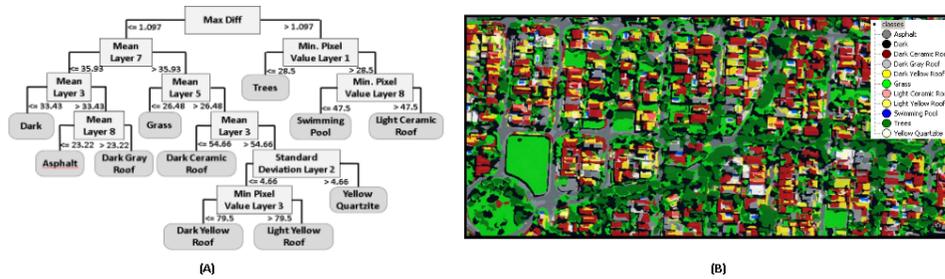
#### 3.2 Decision Trees and Classification of Urban Land Cover

The decision trees trained with the set of statistical descriptors by means of the CART and C4.5 algorithms will be presented in the following sections.

##### 3.2.1 CART

The set of statistical descriptors were mined by CART in WEKA using the implemented SimpleCart algorithm. Figure 3-A shows the decision tree generated by CART based on these descriptors.

Initially, the tree divides itself into two major branches based on the max. diff. attribute. In the left branch, the mean of layer 7, mean of layer 3 and mean of layer 8 were used to separate the very dark classes (asphalt, shadow and dark gray roof) from the classes with intermediate tone (grass, dark ceramic roof, yellow quartzite, dark yellow roof and light yellow roof) with the aid of mean of layer 5, mean of layer 3, standard deviation of layer 2 and minimum pixel value of layer 3. In the right branch, the minimum pixel value of layers 1 and 8 were used to separate the classes of light tone: trees, swimming-pool and light ceramic roof. Figure 3-B shows the image classification result obtained from CART.

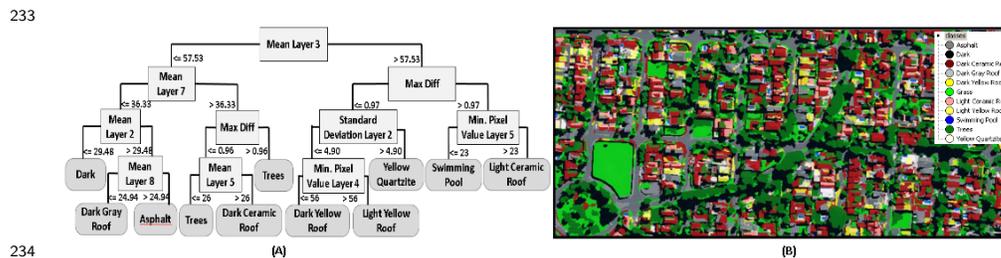


216

217 **Figure 3** - (A) Decision tree generated by tree.SimpleCart (CART) with  
 218 WEKA and (B) final classification results by tree.SimpleCart (CART).

### 219 3.2.2 C4.5

220 The tree.J48 algorithm of WEKA was used to generate the decision tree  
 221 by C4.5. Figure 4-A shows the decision tree generated by C4.5 based on  
 222 the statistical descriptors. The mean of layer 3 divides the tree into two  
 223 branches. The left branch presents the darker classes, being subdivided by  
 224 the mean of layer 7. Shadow, asphalt and dark gray roof were separated  
 225 by the mean of layer 2 and the mean of layer 8. In turn, grass, trees and dark  
 226 ceramic roof were separated by the max. diff. and the mean of layer 5. The  
 227 right branch presents the light classes. Yellow quartzite, dark yellow roof  
 228 and light yellow roof were separated by the max. diff., standard deviation  
 229 of layer 2 and minimum pixel value of layer 4. Swimming-pool and light  
 230 ceramic roof were separated by the max. diff. and the minimum pixel value  
 231 of layer 5. Figure 4-B presents the image classification result obtained from  
 232 C4.5.



234

235 **Figure 4** - (A) Decision tree generated by tree.J48 (C4.5) with WEKA  
 236 and (B) final classification results by tree.J48 (C4.5).

### 237 3.3 Accuracy of Land Coverage Classification

238 Table 1 refers to the confusion matrix presented by the CART classifica-  
 239 tion. The index of global accuracy achieved was 74%, and the Kappa index,  
 240 0.70. Only the dark gray roof class presented errors of omission around 50%,

attaining a producer's accuracy of 55%, due to confusion with asphalt and shadow. Regarding the user's accuracy, the dark gray roof class presented a very low index, 55%, due to confusion with asphalt, dark ceramic roof and shadow. Grass also had a low value of user's accuracy, 50%, due to confusion with shadow and trees. In general, all classes presented values of user's and producer's accuracies above 50%. Table 2 refers to the confusion matrix presented by the C4.5 classification. The index of global accuracy achieved was 79%, and the Kappa index, 0.75. Asphalt showed a low value of producer's accuracy, 59%, due to confusion with grass and shadow. Regarding the user's accuracy, the class grass presented the smallest index, 58%, due to confusion with asphalt, dark ceramic roof and trees. Nevertheless, all classes obtained values of user's and producer's accuracies above 50%.

**Table 1 - Confusion matrix of the CART classification.**

Classes		Reference Samples										Total Classified	
		Asphalt	Dark	Dark Ceramic Roof	Dark Gray Roof	Dark Yellow Roof	Grass	Light Ceramic Roof	Light Yellow Roof	Swim. Pool	Trees		Yellow Quartz.
Classification	Asphalt	15			3	2							20
	Dark	2	15	1	2		5				1		26
	Dark Ceramic Roof	3		14									17
	Dark Gray Roof	2	2	1	6								11
	Dark Yellow Roof					6						1	7
	Grass		3				8				5		16
	Light Ceramic Roof			1				9					10
	Light Yellow Roof			1					3				4
	Swimming Pool									5			5
	Trees		1									18	19
	Yellow Quartzite											5	5
	<b>Total Collected</b>		22	21	18	11	8	13	9	3	5	24	6
Producer's Accuracy		68%	71%	78%	55%	75%	62%	100%	100%	100%	75%	83%	74%
User's Accuracy		75%	58%	82%	55%	86%	50%	90%	75%	100%	95%	100%	

**Table 2 - Confusion matrix of the C4.5 classification.**

Classes		Reference Samples										Total Classified	
		Asphalt	Dark	Dark Ceramic Roof	Dark Gray Roof	Dark Yellow Roof	Grass	Light Ceramic Roof	Light Yellow Roof	Swim. Pool	Trees		Yellow Quartz.
Classification	Asphalt	13											13
	Dark	6	21	1							2		30
	Dark Ceramic Roof			12			2						14
	Dark Gray Roof			2	7						2		11
	Dark Yellow Roof				1	8							9
	Grass	3		2			11				3		19
	Light Ceramic Roof							9					9
	Light Yellow Roof								3				3
	Swimming Pool									5		2	7
	Trees			1	1							17	19
	Yellow Quartzite				2							4	6
	<b>Total Collected</b>		22	21	18	11	8	13	9	3	5	24	6
Producer's Accuracy		59%	100%	67%	64%	100%	85%	100%	100%	100%	71%	67%	79%
User's Accuracy		100%	70%	86%	64%	89%	58%	100%	60%	100%	89%	67%	

#### 4. Conclusions

This work presented a comparison between two classifications of urban land cover. We evaluated the CART and C4.5 data mining algorithms for

260 the generation of decision trees. Considering the Kappa indices of both  
261 experiments, one can state that the two classifications presented satisfactory  
262 and similar quality accuracies.

263 The decision trees generated by the two algorithms are similar, although  
264 they relied on different statistical descriptors. The CART algorithm required  
265 more statistical descriptors to separate classes, so the classification presented  
266 a lot of confusion among the darkest objects, mostly asphalt, dark grey roof,  
267 grass, shadow and trees. In turn, the algorithm C4.5 required less statistical  
268 descriptors and attained better accuracy in the classification.

269 The use of data mining techniques for the exploration of large amount  
270 of data proved to be crucial, as it contributed to the selection of the best at-  
271 tributes and their respective thresholds to characterize the land cover classes.  
272 Despite using only statistical descriptors, the results of the classifications  
273 presented in this work were satisfactory. Other descriptors, in addition to  
274 the statistical ones, could be used in the data mining process, what would  
275 probably increase the classification accuracy. Also, a good segmentation  
276 is able to correctly delimit the objects, making it possible to obtain more  
277 refined samples, and hence, contribute to a more accurate data mining pro-  
278 cess.

279 Finally, the use of data mining by means of decision trees for remotely  
280 sensed data classification has proved to be advantageous, for it allows the  
281 automation of procedures for selecting attributes and defining decision rules,  
282 avoiding the subjectivity of the interpreter. The evaluated decision tree  
283 algorithms also showed to be effective for handling large volumes of data  
284 and suitable for object-based image classification.

## 285 References

- 286 [1] Campbel, J. Introduction to remote sensing, 2007, New York: The Guil-  
287 ford Press.
- 288 [2] Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q.  
289 Per-pixel vs. object-based classification of urban land cover extraction  
290 using high spatial resolution imagery, Remote sensing of environment  
291 115-5: 11451161, 2011.
- 292 [3] Körting , T.S., Fonseca, L.M.G., Câmara, G. GeoDMA - Geographic  
293 Data Mining, *Analyst. Computers & Geosciences* 57: 133145, 2013.

- 294 [4] Silva, M.P.S. Mineração de padrões de mudança em imagens de Senso-  
295 rramento Remoto, 2006, Tese (Doutorado em Sensoriamento Remoto)  
296 INPE, S. J. Campos.
- 297 [5] ENVI. ENVI 4.7 Reference Guide, ITT Visual  
298 Information Solutions, 2009. [Online]. Available:  
299 [http://www.exelisvis.com/portals/0/pdfs/envi/Reference Guide.pdf](http://www.exelisvis.com/portals/0/pdfs/envi/Reference%20Guide.pdf).
- 300 [6] Trimble. eCognition Developer. 8th ed. [Online]. 2014. Available:  
301 <http://www.ecognition.com>
- 302 [7] The University of Waikato. WEKA: Waikato Environment for Knowl-  
303 edge Analysis, 3.6.12, Hamilton, New Zeland. 2014.
- 304 [8] Laben, C.A., Brower, B.V. Process for enhancing the spatial resolution  
305 of multispectral imagery using pan-sharpening. US6011875 A. 2000.
- 306 [9] Baatz, M., Schape, A. Multiresolution segmentation: an optimization  
307 approach for high quality multi-scale image segmentation, *Angewandte*  
308 *Geographische Informationsverarbeitung*. XII: 1223, 2009.
- 309 [10] Zerbe, L.M., Liew, S.C. Reevaluating the traditional maximum NDVI  
310 compositing methodology: The Normalized Difference Blue Index, *geo-*  
311 *science and Remote Sensing Symposium*. Proceedings. IEEE Interna-  
312 tional, 4: 2401-2404, 2004.
- 313 [11] Quinlan, J. C4.5: Programs for Machine Learning, 1993, Morgan Kauf-  
314 man.
- 315 [12] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. *Classifica-*  
316 *tion and Regression Trees*, 2nd Edition, 1984, Pacific Grove, CA,  
317 Wadsworth.
- 318 [13] Lamas, A.I. *Sistemas Neuro-Fuzzy Hierárquico BSP para Previsão de*  
319 *Extração de Regras Fuzzy em Aplicações de Minerações de Dados*, 2000.  
320 *Dissertação de Mestrado*, PUC-Rio.
- 321 [14] Congalton, R., Green, K. *Assessing the Accuracy of Remotely Sensed*  
322 *Data: Principles and Practices*, 2009, Boca Raton: CRC/Taylor &  
323 Francis.