



HAL
open science

Structural knowledge learning from maps for supervised land cover/use classification: Application to the monitoring of land cover/use maps in French Guiana

Meriam Bayouhd, Emmanuel Roux, Gilles Richard, Richard Nock

► To cite this version:

Meriam Bayouhd, Emmanuel Roux, Gilles Richard, Richard Nock. Structural knowledge learning from maps for supervised land cover/use classification: Application to the monitoring of land cover/use maps in French Guiana. *Computers & Geosciences*, 2015, 76, pp.31-40. 10.1016/j.cageo.2014.08.013 . hal-01369822

HAL Id: hal-01369822

<https://hal.science/hal-01369822>

Submitted on 21 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Structural knowledge learning from maps for supervised land cover/use classification: Application to the monitoring of land cover/use maps in French Guiana

Meriam Bayoudh^{a,c,*}, Emmanuel Roux^a, Gilles Richard^b, Richard Nock^{c,d}

^a*ESPACE-DEV, UMR228 IRD/UMII/UR/UAG, Institut de Recherche pour le Développement*

^b*Institut de Recherche en Informatique de Toulouse (IRIT), UMR5505 CNRS/INPT/UPS/UT1/UTM, Toulouse, France*

^c*Université des Antilles et de la Guyane, France*

^d*Centre d'Etude et de Recherche en Economie, Gestion, Modélisation et Informatique Appliquée (CEREGMIA), Martinique, France*

Abstract

1 The number of satellites and sensors devoted to earth observation has be-
2 come increasingly elevated, delivering extensive data, especially images. At
3 the same time, the access to such data and the tools needed to process
4 them has considerably improved. In the presence of such data flow, we need
5 automatic image interpretation methods, especially when it comes to the
6 monitoring and prediction of environmental and societal changes in highly
7 dynamic socio-environmental contexts. This could be accomplished via arti-
8 ficial intelligence.

*Corresponding author. Maison de la Télédétection, 500 rue Jean-François Breton, 34093 Montpellier Cedex 5, France. Tel: +33 (0)4 67 91 72 60

Email addresses: meriam.bayoudh@ird.fr (Meriam Bayoudh), emmanuel.roux@ird.fr (Emmanuel Roux), richard@irit.fr (Gilles Richard), Richard.Nock@martinique.univ-ag.fr (Richard Nock)

9 The concept described here relies on the induction of classification rules that
10 explicitly take into account structural knowledge, using Aleph, an Induc-
11 tive Logic Programming (ILP) system, combined with a multi-class clas-
12 sification procedure. This methodology was used to monitor changes in
13 land cover/use of the French Guiana coastline. One hundred and fifty-eight
14 classification rules were induced from 3 diachronic land cover/use maps in-
15 cluding 38 classes. These rules were expressed in first order logic language,
16 which makes them easily understandable by non-experts. A ten-fold cross-
17 validation gave significant average values of 84.62%, 99.57% and 77.22% for
18 classification accuracy, specificity and sensitivity, respectively. Our method-
19 ology could be beneficial to automatically classify new objects and to facili-
20 tate object-based classification procedures.

Keywords: Supervised classification, Machine learning, Inductive Logic
Programming (ILP), Geographic Information System, Land cover map.

21 **1. Introduction**

22 The availability of remotely sensed Earth observation data, taken from
23 aircrafts (including drones) and satellites, is constantly increasing. This ob-
24 viously comes from the increasing number of Earth observation satellites
25 and sensors. In fact, a recent report (Zaiche and Smith, 2011) estimates
26 that the number of satellite launches will be 50% higher during the next
27 ten years, when compared to the last decade. In particular, 200 govern-
28 mental Earth observation satellites will be launched during that period. At
29 the same time, as an increasing number of countries and/or organizations
30 distribute remotely sensed data for free, the evolution in data distribution

31 and use policies contributes to the use of huge volumes of data. Thus, data
32 processing and interpretation have become a serious challenge for engineers
33 and researchers. Therefore, classical procedures cannot continue to be used,
34 and new approaches are needed to automatically update the land cover/use
35 maps that provide essential information to decision makers.

36 In this context, several studies have formally represented and introduced ex-
37 pert knowledge for automatic image classification and interpretation. For
38 instance, Suzuki et al. (2001) built a system for satellite image classification
39 based on expert knowledge. More recently, Forestier et al. (2012) built a
40 knowledge-base of urban objects, allowing the interpretation of high spatial
41 resolution images in order to assist urban planner with mapping tasks. Re-
42 cent studies devoted to expert knowledge formalization for automatic image
43 interpretation have been directed towards ontologies. Hudelot et al. (2008)
44 proposed an ontology of spatial relations to guide medical image interpre-
45 tation, which is then enriched by fuzzy representations of concepts. Within
46 the remote sensing framework, both Durand et al. (2007) and Andres et al.
47 (2012) propose ontology-based automatic procedures for image processing.

48 A complementary approach to expert knowledge formalization is knowledge
49 extraction from data. Such approach is utilized by all existing supervised
50 image classification procedures, which first require a learning phase with de-
51 limitation and labeling (allocation to a class) of regions in the image. How-
52 ever, most methods consider only pixel information within such regions to
53 separate and characterize the different classes. Structural aspects, i.e., infor-
54 mation arrangement in space, are essentially taken into account by computing
55 textural indexes within the same regions. To our knowledge, there is no op-

56 erative tool that provides general and efficient classification rules exploiting
57 structural knowledge at a higher semantic level, particularly at the object
58 level within the object-oriented image analysis (Blaschke, 2010), when such
59 knowledge is more robust and expressive than at the pixel level.

60 Automatically learning such structural knowledge within the supervised frame-
61 work, however, requires the delimitation and labeling of many more regions
62 than with pixel-based approaches, and would consequently entails important
63 expert efforts. One solution would be to take advantage of existing maps
64 resulting from different types of expertise already acquired (e.g., expertise in
65 remote sensing, image processing, environment, ecology, etc.).

66 Thus far, very few studies have proposed to learn structural knowledge from
67 maps.

68 Malerba et al. (2003) implemented INductive GEographic iNformation Sys-
69 tem (INGENS) to assist with topographic map interpretation. INGENS con-
70 sists of a prototypical extended Geographic Information System (GIS) with
71 inductive learning capabilities. GIS classical functionalities are used to ex-
72 tract relevant concepts and features from spatial database, and the integrated
73 inductive system allows finding rules to automatically recognize complex ge-
74 ographical contexts that are defined by the presence of specific geographical
75 objects and their spatial arrangement in predefined spatial windows (cells).

76 It is devoted to support map interpretation and geographical information re-
77 trieval by enriching geographical queries, but not to automatic classification
78 in the context of large datasets. In fact, such automatic procedures require
79 a quantitative evaluation that has not been performed with INGENS.

80 Vaz et al. (2007) use an Inductive Logic system called APRIL (Fonseca et al.,

81 2006) to learn classification rules from both a detailed map provided by
82 botanists and CORINE Land Cover (CLC) maps of the same zone. Such
83 rules are intended to automatically disaggregate CLC map information that
84 is considered too generic within the application framework. Here again, the
85 precision of the system is not provided.

86 Inductive learning of structural features from maps has been applied to the
87 prediction of particular events that partially depend on landscape charac-
88 teristics. Vaz et al. (2010) propose a system that predicts wildfires from
89 information on past fires and from compositional and structural features of
90 the land use. However, the performance of the predictions, estimated by a
91 10-fold cross validation, does not seem to allow operational use.

92 Finally, Chelghoum et al. (2006) automatically transformed spatial relation
93 information stored in multi-tables into first-order logic, and used S-TILDE
94 (Spatial Top-down Induction Logical DEcision tree) to induce classification
95 rules. They applied their method for spatial prediction of shellfish contamina-
96 tion in the Thau lagoon. Their work considered only the binary classification
97 problem.

98 In such applicative and scientific contexts, we report here a method for
99 structural and symbolic knowledge extraction from land use/cover maps and
100 complementary geographic information layers, combined with a multi-class
101 classification approach. Our work does not deal with the delimitation of re-
102 gions (or segments) from images, but with the labeling of previously defined
103 image regions. Methods intended to image region delimitation, including
104 segmentation methods, are therefore beyond the scope of this study. In this
105 study we chose the Inductive Logic Programming framework (ILP) (Mug-

106 gleton, 1991) for the learning task, and a multi-class classification procedure
107 developed by Abudawood and Flach (2011) within the ILP framework, i.e.,
108 the Multi-class Rule Set Intersection (MRSI). This methodology was tested
109 to update land cover/use maps of the French Guiana coastline, and the re-
110 sulting classification system was thoroughly evaluated from qualitative and
111 quantitative points of view through a ten-fold cross-validation.

112 Our paper is organized as follows: the general methodology is explained,
113 by presenting the ILP approach, the geographic information extraction and
114 coding, the multi-class classification technique and the evaluation procedures.
115 Then, the application to land/use maps updating is described, by detailing
116 the exploited dataset and the adaptation of the general methodology. The
117 next section presents the results by qualifying the induced rules and provid-
118 ing prediction quantitative scores. We then discuss our results and a general
119 conclusion is given about the proposed approach.

120 **2. Materials and Methods**

121 *2.1. Inductive Logic Programming*

122 Inductive Logic Programming (ILP) (Muggleton, 1991) is a search field
123 that combines machine learning and logic programming. It is a technique for
124 learning a general theory H from a background knowledge B and examples
125 E within a framework provided by clausal logic.

126 ILP can model complex problems and has been used in several fields such
127 as chemistry (Blockeel et al., 2004), biology, physics, medicine (Luu et al.,
128 2012; Fromont et al., 2005), ecology and bio-informatics (Santos et al., 2012;
129 Lavrac and Dzeroski, 1994; Srinivasan et al., 1996). It has, also, been applied

130 to chess (Goodacre, 1996) and to test the quality of river water (Cordier,
131 2005). Very few studies have applied this method to geographical data, as
132 already discussed in the introduction (Malerba et al., 2003; Vaz et al., 2007,
133 2010; Chelghoum et al., 2006).

134 ILP is defined as follows (Lavrac and Dzeroski, 1994):

135 Given:

- 136 • A description language L .
- 137 • Background knowledge B , expressed under Horn clauses (a subset of
138 general first order logic formula, expressed using L , describing the ex-
139 isting knowledge and constraints on the target concept, *i.e.*, in our case,
140 the allocation to a given land cover/use class;
- 141 • A set of examples E , divided into two subsets, E^+ and E^- , which
142 represent the sets of positive and negative examples, respectively;

143 Find a "theory" H , *i.e.*, a set of formula using the description language
144 L that covers positive examples E^+ , but does not cover (or in a controlled
145 way) the negative examples E^- .

146 We chose the ILP engine Aleph (Srinivasan, 2007). It is an open source
147 ILP system, written in Prolog, using top-down search and based on inverse
148 entailment (Muggleton, 1995).

149 2.2. Geographic information extraction and coding

150 Each patch of land use/cover map is referred to as *object* and defines
151 the elementary geographical entity to which the reasoning will be applied.
152 Objects are used to define the examples for the learning and test phases.

153 Objects are described using predicates characterizing their intrinsic (class,
 154 area, fractal dimension, compactness, perimeter) and relational features (ad-
 155 jacency, inclusion, relative positions in latitudinal and longitudinal direc-
 156 tions) (*cf.* Table 1). The choice of such predicates is essentially based on *a*
 157 *priori* knowledge of the authors on the discriminating features of the spatial
 158 objects constituting land cover/use maps.

159 Inductive Logic Programming being adapted to symbolic information, dis-
 160 cretization of the numeric variables is performed, and the information recoded
 161 as follows: for any numeric variable V , the 10th, 20th, ..., 90th percentiles of
 162 the empirical distribution of V , denoted p_k ($k \in [1, 9]$), are computed. Then,
 163 for every p_k , two predicates were defined to indicate if an observed value X
 164 for V is lower or higher than p_k . For instance, the observed numeric value
 165 X , corresponding to the area of the object O , is recoded, for p_k , as follows:

$$\text{area_symb}(O, I_k) :- \text{area_num}(O, X), \quad X \leq p_k.$$

$$\text{or } \text{area_symb}(O, S_k) :- \text{area_num}(O, X), \quad X > p_k.$$

166 with I_k and S_k as the intervals $[-\text{inf}, p_k]$ and $]p_k, +\text{inf}]$, respectively.

167 Eventually, the latitude and longitude values were used to characterize the
 168 relative positions of the object pairs (*cf.* Table 1).

Table 1: Predicates used for object characterization. Asterisk indicates that the predicate is not used in the rule premises.

Predicates	Description
$\text{object}(O)$	Declaration of the object O
$\text{class}(O, \text{class_label})$	The object O belongs to the class class_label
$\text{adjacent}(O1, O2)$	$O1$ and $O2$ are two adjacent objects
$\text{included}(O1, O2)$	$O2$ is included in $O1$
$\text{contains}(O, E)$	O contains the entity E (<i>e.g.</i> $E \in \{\text{River}, \text{Road}, \text{Building}, \dots\}$)
$\text{area_num}(O, X)*$	X is the area (m^2), the compactness value, the fractal dimension and the perimeter (m) of the object O , respectively, with ($X \in \mathfrak{R}$)
$\text{compactness_num}(O, X)*$	
$\text{fract_dim_num}(O, X)*$	
$\text{perimeter_num}(O, X)*$	
$\text{area_symb}(O, I_k^{\text{area}} \text{ or } S_k^{\text{area}})$	Recoding of the numeric variables according to the percentiles (see text for details)
$\text{compactness_symb}(O, I_k^{\text{comp}} \text{ or } S_k^{\text{comp}})$	
$\text{fract_dim_symb}(O, I_k^{\text{df}} \text{ or } S_k^{\text{df}})$	
$\text{perimeter_symb}(O, I_k^{\text{per}} \text{ or } S_k^{\text{per}})$	
$\text{lat}(O, X)*$	X is the latitude and longitude of O , respectively, ($X \in \mathfrak{R}$)
$\text{long}(O, X)*$	
$\text{north}(O1, O2) :-$ $\text{lat}(O1, A), \text{lat}(O2, B), A > B.$	$O1$ is located north, south, east and west of $O2$, respectively.
$\text{south}(O1, O2) :-$ $\text{lat}(O1, A), \text{lat}(O2, B), A \leq B.$	
$\text{east}(O1, O2) :-$ $\text{long}(O1, A), \text{long}(O2, B), A > B.$	
$\text{west}(O1, O2) :-$ $\text{long}(O1, A), \text{long}(O2, B), A \leq B.$	

169 *2.3. Rule induction: one-vs-rest approach*

170 Once the information is extracted and coded according to the above
171 method, the classification rules are induced by the inductive system Aleph.
172 When applying ILP within the multi-class framework, i.e., in the case of more
173 than two classes (each object belonging to only one class), the *one-vs-rest*
174 approach is a commonly used approach (Abudawood and Flach, 2011). Such
175 method consists in generating as many classifiers as classes, by defining the
176 positive and negative example sets for each class c as follows:

$$\begin{cases} E^+ = \{O/\text{classe}(O, c)\} \\ E^- = \{O/\text{classe}(O, \bar{c})\} \end{cases}$$

177 and by running Aleph with such example sets, for each class c .

178 *2.4. Multi-class framework*

179 Considering the previously described one-vs-rest approach results in in-
180 ducing as many classifiers as classes. Considering the classifiers indepen-
181 dently of one another, one or several classes can be predicted when a new
182 object is to be classified. Abudawood and Flach (2011) proposed several
183 solutions to handle multi-class problems for ILP. Among them, the Multi-
184 class Rule Set Intersection (MRSI) method gave the highest accuracies and
185 Areas Under the ROC Curve (AUC) when taking multi-class data sets into
186 account (Abudawood and Flach, 2011). The principle of the MRSI method
187 is: i) the theories induced for each class are gathered in an unique rule set;
188 ii) for each rule i , the set of covered examples by the rule, C_i , is stored; iii) a
189 default rule is formed that concludes to the majority class of the uncovered
190 examples; iv) for an unseen object O , the intersection of the sets of examples

191 covered by the fired rules is computed ($I = \cap C_i | r_i \text{ is fired}$) and, finally; v)
 192 the predicted class \hat{c} is the majority class in the set I , i.e., the more probable
 193 class given to the new object O , with an empirical probability $p(c|O)$.

194 2.5. Prediction evaluation

195 Overall accuracy, sensitivity, specificity and *Kappa* index are computed
 196 based on a 10-fold stratified cross-validation procedure.

197 For each class C_i ($i \in [1, n]$), the set of positive examples E_i is randomly
 198 divided in ten subsets $E_{i,f}$ ($f \in [1, 10]$). If a class j is associated with p
 199 positive examples, with $p < 10$, then $E_{i,f>p} = \emptyset$. Then the f^{th} learning set
 200 for the i^{th} class is defined as follows:

$$\begin{cases} E_{i,f}^+ = \cup_{l=1, \dots, 10; l \neq f} E_{i,l} \\ E_{i,f}^- = \cup_{j=1, \dots, n; j \neq i} \{ \cup_{l=1, \dots, 10} E_{j,l} \} \end{cases}$$

201 In the multi-class classification framework, one test set T_f has to be de-
 202 fined for each fold f . Such test set is consequently defined as follows:

$$T_f = \cup_{i=1, \dots, n} E_{i,f}$$

203 Overall accuracy, sensitivity, specificity and *Kappa* index values are com-
 204 puted for each test set, then averaged. The formulas of these measures are
 205 given hereafter.

206 The multi-class classification procedure previously described permits to com-
 207 pute the multi-class contingency table (see Table 2) for each test set, and to
 208 obtain the overall accuracy as follows (Abudawood and Flach, 2011):

$$Overall\ Accuracy = \sum_{i=1}^n \frac{TP^{(i)}}{E} \quad (1)$$

209 where n is the number of classes, $TP^{(i)}$ the number of *true positives* for
 210 the class i , and E the total number of test examples.

Table 2: Contingency table with notations (TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative) for the class i only. (Adapted from Abudawood and Flach (2011))

		Predicted							Total
		C_1	...	C_{i-1}	C_i	C_{i+1}	...	C_n	
Actual	C_1	$TN_1^{(i)}$	$FP_1^{(i)}$	E_1

	$TN_{i-1}^{(i)}$	$FP_{i-1}^{(i)}$	E_{i-1}
	C_i	$FN_1^{(i)}$...	$FN_{i-1}^{(i)}$	$TP^{(i)}$	$FN_{i+1}^{(i)}$...	$FN_n^{(i)}$	E_i
	$FP_{i+1}^{(i)}$	$TN_{i+1}^{(i)}$	E_{i+1}

	C_n	$FP_n^{(i)}$	$TN_n^{(i)}$	E_n
	Total	\hat{E}_1	...	\hat{E}_{i-1}	\hat{E}_i	\hat{E}_{i+1}	...	\hat{E}_n	E

211 For each class i , the sensitivity, i.e. the ability of the classifier to success-
 212 fully classified positive examples, is computed as:

$$Sensitivity^{(i)} = \frac{TP^{(i)}}{TP^{(i)} + \sum_{j=1, j \neq i}^n FN_j^{(i)}} = \frac{TP^{(i)}}{E_i} \quad (2)$$

213 where $FN_j^{(i)}$ is the number of *false negatives* for the class i wrongly as-
 214 sociated to the class j .

215 The specificity, i.e. the ability of the classifier to successfully classified
 216 negative examples, is computed as:

$$Specificity^{(i)} = \frac{\sum_{j=1, j \neq i}^n TN_j^{(i)}}{\sum_{j=1, j \neq i}^n TN_j^{(i)} + \sum_{j=1, j \neq i}^n FP_j^{(i)}} \quad (3)$$

217 where $TN_j^{(i)}$ is the number of *true negatives* for the class i successfully
 218 attributed to the class j and $FP_j^{(i)}$ the number of *false positives* for the class
 219 i that actually belong to the class j .

220

221 Finally, the *Kappa* index is computed for each test set. *Cohen's Kappa*
 222 (Cohen, 1960) provides a statistical measure of inter-agreement for quali-
 223 tative items. In the framework of classification, it measures the degree of
 224 agreement between predicted and actual classes. *Kappa* index is defined as
 225 follows:

$$kappa = \frac{P(A) - P(H)}{1 - P(H)} \quad (4)$$

226 With $P(A)$ corresponding to the observed proportion of agreement be-
 227 tween two classifications, and $P(H)$ the estimated proportion of agreement
 228 expected by chance.

229 **3. Application to the update of the land cover/use maps of the** 230 **French Guiana coastline**

231 The concepts and methods previously defined were applied to an actual
 232 geographic situation. The French Guiana territory is subject to intense an-
 233 thropogenic and natural dynamics (Anthony et al., 2010): cyclic coastal
 234 erosion and accretion, notably due to the transport of sediments from the
 235 Amazon River by oceanic currents; and expansion of urban, peri-urban, agri-

236 cultural areas. In this context, it is essential to develop automated methods
237 for monitoring the land cover/use of the French Guiana territory. In partic-
238 ular, the large amount of available aerial photographs and satellite images is
239 a critical source of materials that should be better exploited. If the delim-
240 itation of the geographical objects of interest does not require a high level
241 of expertise and can be performed by operators, allocating these objects to
242 land cover/use classes appears far more complex and subjective. In fact, de-
243 spite efforts made to formalize and standardize the classification procedures,
244 such allocating task requires a deep knowledge of the different types of land
245 cover/use, both in the imaging and applicative domains. Consequently, the
246 learning and classification methods previously presented were applied to au-
247 tomatically perform the labeling task and update the land cover/use maps
248 of the French Guiana coastline.

249 *3.1. Dataset*

250 We took advantage of a series of three land cover/use maps of the French
251 Guiana coastline for 2001, 2005 and 2008. The classification nomenclature is
252 based on the CORINE Land Cover (CLC) European nomenclature, which is
253 adapted to the Amazonian context by the addition of 15 classes, 9 of them
254 corresponding to different types of forests, and consists of three nested levels
255 where the most detailed (level III) is composed of 39 classes.

256 The maps were produced by the French National Office of Forests (Of-
257 fice National des Forêts; ONF) by photo-interpretation of the BD-Ortho[®]
258 aerial photographs of the French National Geographic Institute (Institut Géo-
259 graphique National: IGN) for 2001 and 2005. Air photographs had a 50-cm
260 spatial resolution. The land cover/use map for 2008 was updated using 2.5-

261 meter spatial resolution satellite images acquired by the SPOT 5 satellite
262 and obtained through the SEAS-Guyane ¹ project.

¹<https://www.seas-guyane.org>

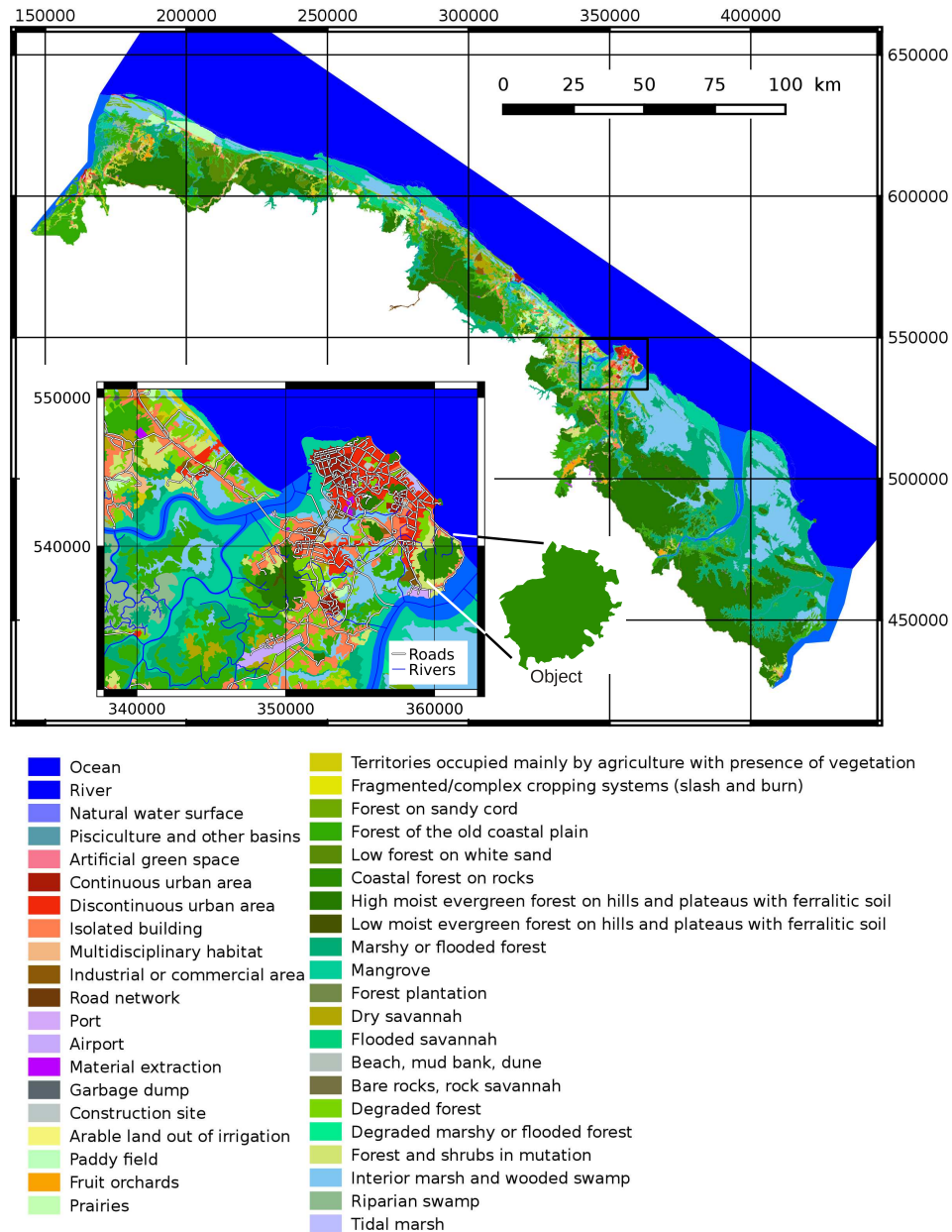


Figure 1: Land cover/use map and complementary geographic information layers (inset) used in this article (geographic coordinate system: WGS84 / UTM zone 22N). Sources: French National Office of Forests (Office National des Forêts; ONF); French National Geographic Institute (Institut Géographique National: IGN); French Ministry in charge of the environment; Regional Direction of the Environment (DIREN) of French Guiana; French National Agency for Water and Aquatic Environments (ONEMA). See text for details.

263 Two complementary geographic information layers were used (see Figure
264 1): the road network, provided by the BD-Carto® database of the IGN, and
265 the river network provided by the BD-Carthage® database of the French
266 Ministry in charge of the Environment and of the IGN, produced in 2009
267 for French Guiana by the Regional Direction of the Environment (DIREN)
268 of French Guiana and the French National Agency for Water and Aquatic
269 Environments (ONEMA).

270 3.2. Data pre-processing: definition of the map objects

271 Firstly, we completed the initial land cover/use classification by adding
272 three more classes: *Ocean*, *River* and *Unknown*. The first two classes con-
273 tribute significantly to the structure of the environment in the French Guiana
274 territory, and the *Unknown* class explicitly takes into account the fact that
275 information was not available for some areas in 2001 and/or 2005. However,
276 we did not induce any rules to predict membership to these three classes.
277 Finally, the class *Paddy field* was not considered as it was under-represented
278 in the maps (only 2 positive examples). Thus 38 land cover/use classes were
279 considered (see Tables 3, 4 and 5 for the class list).

280 In this study, we follow the land cover/use class of the objects in time. We do
281 not explicitly follow the object delimitations, which is a much more complex
282 task. In fact, by taking into account the information provided by three origi-
283 nal maps, object boundaries can change in time: an object can be splitted
284 into two or more objects belonging to different classes (see for instance object
285 s_{13} in figure 2), creating new object(s); an object can result from the merg-
286 ing of several objects, making one or several objects disappear. We handled
287 such situations by generating objects with invariant boundaries in time and

288 related to an unique class for each year. Practically, we produced a synthetic
 289 map by concatenating the information contained in the three original maps,
 290 by means of the "union" GIS operator, as schematically shown in Figure 2.
 291 The elementary geographical entities of the resulting map are referred to as
 292 *objects* thereafter, and contribute to define the *examples* in the ILP process.

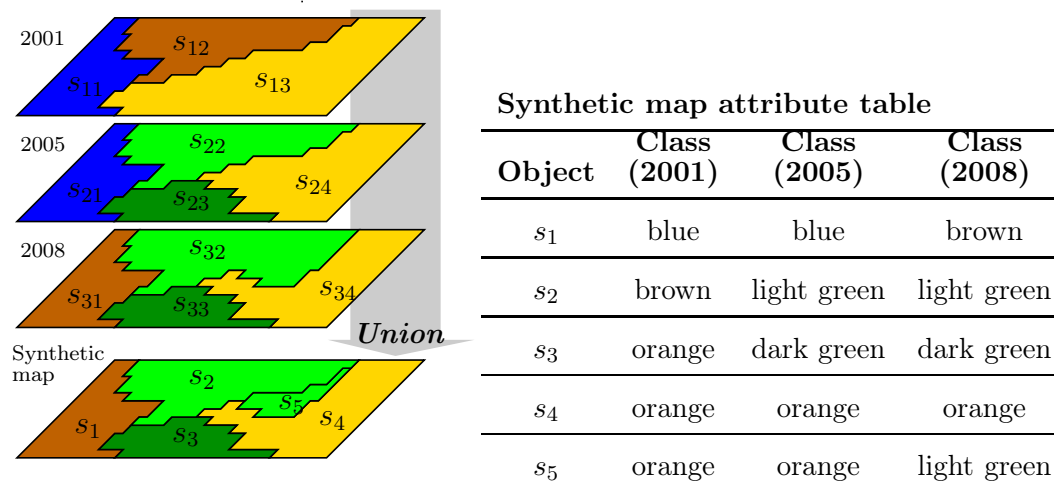


Figure 2: Illustrative example explaining the definition of a synthetic map that combines the information from the three initial maps.

293 3.3. Information coding

294 Target predicates (i.e., concepts to be learned) were defined as the land
 295 cover/use classes to which the objects of the synthetic map belonged in 2008,
 296 considered as the reference year y_0 .

297 Given the diachronic characteristics of the data, 3 predicates were defined to
 298 indicate the class of an object as a function of the time: $\text{class}_{y_0}(O, \text{class_name})$,
 299 $\text{class}_{y_{-3}}(O, \text{class_name})$ and $\text{class}_{y_{-6}}(O, \text{class_name})$, indicating the
 300 land cover/use class of the object O for the years y_0 , y_{-3} and y_{-6} , respec-

301 tively, *i.e.*, for 2008, 2005 and 2001. It is worth noting that from a relative
302 point of view, the year 2001, seven years prior 2008, is assumed to actually
303 correspond to the sixth year before the reference year y_0 . In fact, we can
304 assume marginal changes between 2001 and 2002. However, this assumption
305 has also a practical justification as it permits to consider the updating of the
306 land cover/use information every three years based on the maps established
307 three and six years before.

308 Given the complementary information layers used in our test, the predicate
309 $\text{contain}(O, X)$ referred to rivers and roads ($X \in \{\text{river}, \text{road}\}$) (see Table
310 1).

311 All object features were extracted using the free and open source GRASS
312 Geographic Information System (GRASS Development Team, 1999-2012).

313 *3.4. Rule induction: Aleph parametrization*

314 In Aleph, the accuracy of the candidate clauses was set to 0.7, considered
315 as a good compromise between precision and generalization requirements.
316 Such accuracy is defined as $p/(p+n)$, where p and n are the numbers of pos-
317 itive and negative examples, respectively, which are covered by the clause.
318 Consequently, it differs from the overall accuracy defined in section 2.5, which
319 evaluates the global prediction accuracy of the classification system, based
320 on the whole induced rule set.

321 The maximum premise length was set to 5 literals, such number of conditions
322 in a conjunction being practically considered as the limit for easy compre-
323 hension (Michalski, 1983).

324 4. Results

325 4.1. Set of induced rules

326 The induction process returned 158 classification rules for the 38 land
327 cover/use classes. However, the distribution among land cover/use classes is
328 not homogeneous (see Tables 3 to 5). For instance, we obtained 23 rules for
329 the class *Forest* of the *old coastal plain* whilst we had just one rule for the
330 *Riparian swamp* class. Rules cover from 2 to 692 positive examples while the
331 number of covered negative examples varied from 0 to 99.

332 Three examples of induced rules are shown below, with the number of positive
333 (Pos cover) and negative (Neg cover) examples covered by the rule, and the
334 total number of positive examples for the considered target predicate (Total
335 pos. ex.) in brackets.

336 (1) (Pos cover = 472; Neg cover = 88; Total pos. ex. = 552)
337 `class_y0(A, Multidisciplinary habitat) :- area_symb(A, ≤165567),`
338 `adjacent(A, B), class_y-3(B, Multidisciplinary habitat).`

339 (2) (Pos cover = 2 Neg cover = 0 Total pos. ex. = 40)
340 `class_y0(A, Industrial or commercial area) :- adjacent(A, B),`
341 `class_y-6(B, Construction sites), area_symb(A, ≤10831).`

342 (3) (Pos cover = 3 Neg cover = 0 Total pos. ex. = 166)
343 `class_y0(A, Discontinuous urban area) :- class_y-6(A, Construction`
344 `sites), area_symb(A, ≤76202), area_symb(A, >10831).`

345 Rule (1) covers 472 positive examples for a total of 552 objects actually
346 belonging to the class of interest (85.5%) and 88 negative examples. It in-
347 dicates that an object will belong to the *Multidisciplinary habitat* class if

348 its area is less than or equal to 165 567 m² and is adjacent to an object
349 belonging to the same class three years before. Rule (2) indicates that an
350 object will belong to the *Industrial or commercial area* class if its area is
351 less than or equal to 10 831 m² and is adjacent to an object belonging to
352 the class *Construction sites* 6 years before. Rule (3) indicates that an object
353 will belong to the *Discontinuous urban area* class if its area, in m², belongs
354 to the interval]10831, 76202] and if it belonged to the class *Construction*
355 *sites* 6 years before. By considering such rules for the characterization of the
356 territory dynamics, the first rule illustrates the extension dynamics of the
357 natural areas whereas the second and the third rules describe the extension
358 dynamics of the anthropogenic areas.

359 4.2. Prediction evaluation

360 Tables 3 to 5 report the sensitivity results for each land cover/use class in
361 the one-vs-rest framework by considering each classifier independently, and
362 correspond to sensitivity values that fall in the intervals]0%, 50%],]50%, 80%]
363 and]80%, 100%], respectively. Among the 38 land cover/use classes, only 5
364 classes (13.1%) were associated with sensitivity values under 50%. Twelve
365 classes (31.6%) had sensitivity values between 50% and 80%, and 21 classes
366 (55.3%) had the highest sensitivity values (greater than 80%).

367 All classifiers were 100% specific, except for one related to the class *Forest*
368 *and shrubs in mutation*, which had a specificity of 83.1%.

Table 3: Averaged sensitivities obtained with 10-fold cross validation, for land cover/use classes associated with "low" sensitivity values (lower than 50%), total number of positive examples and number of induced rules for each class, by taking into account the whole dataset as learning set. (The nomenclature is based on the CORINE Land Cover (CLC) European Nomenclature with three nested levels. We applied our method to the most detailed level (level III). The nomenclature levels I and II are indicated for facilitate results interpretation only.)

Class (level I)	Class (level II)	Class (level III)	Sensitivity	Total number of positive examples	Number of rules
Forest and semi-natural area	Open space with some/no vegetation	beach, mud bank, dune	5.0	15	1
	Forest	Moist evergreen forest of the main- land coastal plain	Low forest on white sand	41.7	24
Artificial Territories	Mine, garbage dump or construction sites	Garbage dump	25.0	15	1
		Construction sites	30.1	97	6
Agricultural Territories	Heterogeneous agricultural areas	Territories occupied mainly by agriculture with presence of vegetation	41.1	112	3

Table 4: Averaged sensitivities obtained with 10-fold cross validation, for land cover/use classes associated with "medium" sensitivity values (between 50% and 80%), total number of positive examples and number of induced rules for each class, by taking into account the whole dataset as learning set.

Class (level I)	Class (level II)	Class (level III)	Sensitivity	Total number of positive examples	Number of rules
Artificial Territories	Industrial zone	Industrial or commercial area	65.0	40	2
		Road network	56.9	84	3
		Port	80.0	5	1
	Mine, garbage dump or construction sites	Material extraction	63.5	137	5
	Artificial green space		75.0	8	1
Agricultural Territories	Prairies	Prairies	67.9	243	3
	Arable land	Arable land out of irrigation	70.0	12	1
Forest and semi-natural area	Degraded natural environment	Degraded forest	60.3	483	11
		Moist evergreen forest of the mainland coastal plain	70.0	14	3
	Forest	Coastal forest on rocks			
		Forest of the old coastal plain	79.9	543	23
		Moist evergreen forest on hills and plateaus with ferralitic soil	High forest	76.4	194
	Degraded natural environment	Degraded marshy or flooded forest	80.0	18	1

Table 5: Averaged sensitivities obtained with 10-fold cross validation, for land cover/use classes associated with "high" sensitivity values (greater than 80%), total number of positive examples and number of induced rules for each class, by taking into account the whole dataset as learning set.

Class (level I)	Class (level II)	Class (level III)	Sensitivity	Total number of positive examples	Number of rules	
Artificial Territories	Urbanized areas	Continuous urban area	93.0	42	3	
		Discontinuous urban area	87.9	166	5	
		Isolated building	95.3	1191	8	
		Multidisciplinary habitat	94.4	552	2	
Agricultural Territories	Industrial zone	Airport	100.0	12	1	
		Permanent cultivation	87.1	259	1	
	Heterogeneous agricultural areas	Fruit orchards				81.9
Forest and semi-natural area	Forest	Forest plantation	81.7	21	1	
		Moist evergreen forest of the mainland coastal plains	Forest on sandy cord	82.0	49	3
	Shrubby environment	Moist evergreen forest on hills and plateaus with ferralitic soil	Low forest	98.0	58	1
		Marshy or flooded forest	91.7	288	5	
	Open space with some/no vegetation	Mangrove	93.0	259	16	
		Dry savannah	93.9	164	1	
		Flooded savannah	92.0	98	3	
	Degraded natural environment	Bare rocks, Rock savannah	100.0	6	1	
		Forest and shrubs in mutation	100.0	602	18	
	Wet areas	Lower wet areas	Interior marshes and wooded swamps	92.6	163	4
Riparian swamp			100.0	38	1	
Water surface	Marin Wetland	Tidal marsh	88.9	9	1	
		Continental water	Pisciculture and other basins	85.0	18	1
	Natural water surface	100.0	4	1		

369 Table 6 summarizes the results for overall accuracy and *Kappa* Index.
 370 Overall accuracy values varied from 82.4% to 87.3% with an average of 84.6%.
 371 *Kappa* Index varied from 0.69 to 0.77 with an average value of 0.70.

Table 6: *Kappa* and overall accuracy values.

Test set	1	2	3	4	5	6	7	8	9	10
Kappa	0.69	0.67	0.74	0.71	0.75	0.68	0.69	0.73	0.60	0.77
	0.70 (average)									
Overall accuracy (%)	83.0	87.3	84.3	85.0	84.3	85.1	84.1	83.1	87.2	82.4
	84.6 (average)									

372 4.3. Map of prediction errors

373 By regrouping the results for the 10 test sets, it was possible to construct
 374 a prediction map for the year of interest (2008 in this case). Figure 3 is the
 375 spatial representation of such prediction errors, highlighting that the errors
 376 are not homogeneously distributed in space, two error clusters being present
 377 at the extreme west and at the center of the territory.

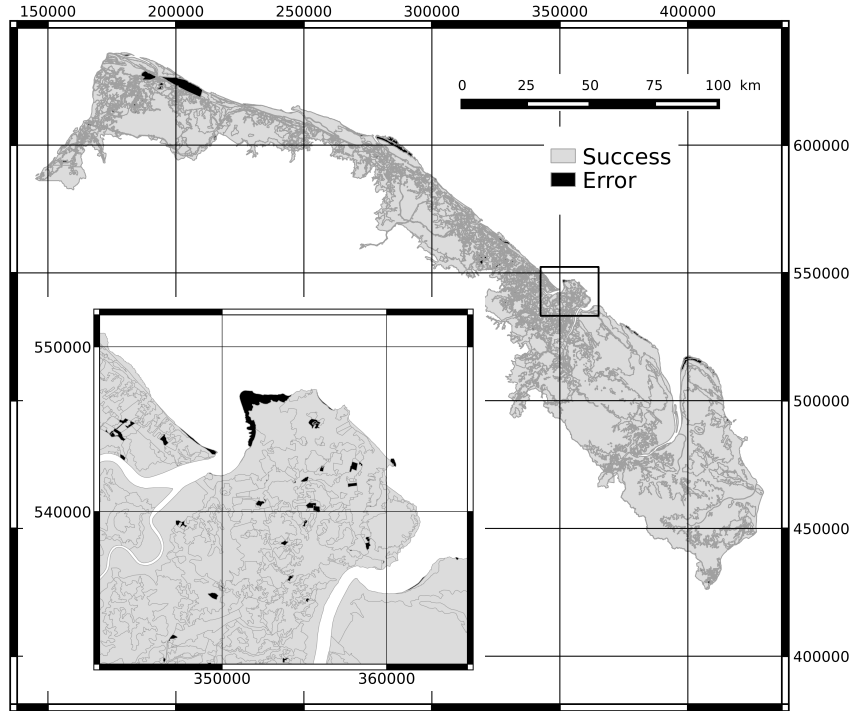


Figure 3: Map of prediction errors (geographic coordinate system: WGS84 / UTM zone 22N). Map at the top represents French Guiana coastline; Map in the inset zooms in on the "Cayenne Island".

378 **5. Discussion**

379 From a qualitative point of view, induced rules are consistent with the
 380 observed environmental features and dynamics of the study area. Moreover,
 381 they are provided in an expressive formalism, and are easily understandable
 382 and interpretable by non-experts, as they can be expressed in natural lan-
 383 guage. However, some rules covered very few (2 or 3) positive examples,
 384 whereas the total number of positive examples for the associated classes was

385 large (see rule (3) in paragraph 4.1 for example). Such rules were conse-
386 quently very specific and did not represent a significant knowledge within
387 the application domain.

388 The predicates *south*, *north*, *east* and *west* did not appear in the rules, show-
389 ing that such predicates were not pertinent for object discrimination, and
390 that characterization of the objects should make better use of expert knowl-
391 edge. In particular, domain ontologies could guide the learning process by
392 identifying the predicates and the learning constraints to use.

393 Whereas the maximum premise length was set to 5, induced rules comprised
394 at most 4 literals. For some classes, this can be explained by the fact that the
395 upper bound on the nodes to be explored when searching for an acceptable
396 clause (i.e., 5000, the default value) was reached and that Aleph stopped
397 before having scanned all the search space.

398 When considering the sensitivity values, we noticed that classes associated
399 with very high sensitivity (Table 5) underwent no or slow changes with time,
400 as the knowledge of the land cover type at one time in the past defined for
401 a large part the land cover type at present and in the future. It is the case
402 for very anthropogenic land use classes such as *Airport* and *Isolated build-*
403 *ings* or for very stable natural land cover types that cannot be exploited by
404 humans due to natural and/or legal constraints, such as *Bare rocks*, *Rock sa-*
405 *vannah*, *Riparian swamp*, or *Natural water bodies*. Instead, classes associated
406 with low sensitivity values (Table 3) seemed to correspond to continually and
407 rapidly shifting land cover/use types. It is more specifically the case for the
408 following classes: *Beach*, *mud bank or dune*, which is a class associated with
409 a highly dynamic environment (Anthony et al., 2010); *Construction sites* and

410 *Territories occupied mainly by agriculture with presence of vegetation*, which
411 is a complex class including traditional itinerant slash and burn activities
412 that consist in cultivating an area and then letting the natural vegetation
413 to regenerate. This seems to indicate that the information provided by the
414 land cover/use maps is insufficient in terms of anteriority and/or time resolu-
415 tion for these classes. However, prediction performances could be improved.
416 In fact, background knowledge can be enriched by adding predicates, pos-
417 sibly evaluated from complementary geographic information layers (digital
418 elevation model, soil map, *etc.*). As already mentioned, the choice of these
419 complementary object features can be guided by expert knowledge, notably
420 through domain ontologies. Better performances could also be obtained by
421 implementing different learning and classification strategies: in our case, *a*
422 *priori* known classes at year y_0 could be exploited to learn more efficient
423 rules. These classes should be the most stable in time and the easiest to
424 identify (e.g. *River, Continuous urban area, Airport, etc.*). An iterative
425 learning-classification strategy could also be implemented, by: i) first learn-
426 ing and classifying classes associated with high-performance predictions (e.g.
427 *Forest and shrubs in mutation*, see Table 5); ii) then using the prediction
428 to enrich the background knowledge of other classes; iii) learning-classifying
429 these classes; iv) repeating the procedure until all classes are predicted. How-
430 ever, the number of strategies is such that we must rely on objective criteria
431 and/or intensive simulations to determine the most appropriate one.
432 Nevertheless, our method gave good results globally. In fact, in addition to
433 the excellent sensitivity and specificity values returned by the procedure, the
434 *Kappa* Index and overall accuracy values were high. According to the *Kappa*

435 interpretation table by (Landis and Koch, 1977), these values denote "strong
436 agreement" between predicted and actual classes.

437 The spatial representation of the prediction errors highlighted that the errors
438 are not homogeneously distributed in space. Except for the errors already
439 discussed and associated with highly dynamic environmental processes, es-
440 sentially distributed along the ocean (e.g., *Beach, mud bank or dune*), two
441 error clusters were identified at the extreme west and at the center of the
442 territory. Understanding such errors will require further investigation, but
443 they may be explained by the presence of errors in the initial maps. Con-
444 sequently, we suggest that the present work can also be a tool to guide the
445 validation of the existing maps.

446 Inductive Logic Programming is devoted to symbolic data. The management
447 of numeric information by ILP constitutes a specific research field, which is
448 beyond the scope of this paper. However, several simple solutions exist in
449 order to code the numeric data into symbolic ones. In fact, the domain of
450 values of a numeric variables can be categorized by means of crisp or fuzzy
451 modalities. We propose here to code the numeric information by means of
452 inequalities taking into account quantiles of the numeric variable empirical
453 distribution. This enables Aleph to manage numeric information in a manner
454 comparable to the Confidence-based Concept Discovery (C²D) ILP system
455 (Kavurucu et al., 2011). This solution seems to offer a good compromise be-
456 tween information loss and generalization capacity, by allowing the system to
457 automatically discover significant value intervals (see rule (3) in the Results
458 section).

459 Finally, the method proposed here does not consider the image processing

460 step devoted to the delimitation of the regions of the image that define the ob-
461 jects. It only considers the labeling (or classification) of the regions. This im-
462 plies: that the partitioning of the image into regions is performed beforehand,
463 by means of any methods including fully manual ones (photo-interpretation)
464 or automatic image segmentation algorithms; that the new objects, which la-
465 bels have to be predicted, have been delimited by the method that produced
466 the objects used for the learning task of the classification rules.

467 **6. Conclusion**

468 This article describes an approach inducing classification rules to au-
469 tomatically label regions of remote sensing images in order to design land
470 cover/use maps. Automatic extraction of structural knowledge using Induc-
471 tive Logic Programming was implemented and new examples were classified
472 to a unique class by means of the Multi-class Rule Set Intersection method.
473 The proposed methodology was then applied to update the land cover/use
474 of the French Guiana coastline and evaluated thoroughly.

475 We show that the induced rules provide knowledge on structural aspects.
476 The quantitative evaluation of our method demonstrated promising results,
477 allowing to offer automatic updating of the land cover/use information in
478 the study region and significant support to the operators in charge of such
479 updating. In particular, our approach could provide valuable assistance to
480 operators using object-based image analysis. In fact, such image analysis ap-
481 proach allows integrating high level symbolic knowledge concerning spatial
482 relations in the classification process. However, to our knowledge, it does
483 not offer any support to the operators in order to define efficient and general

484 rules that take into account such knowledge.

485 Our future work should include guiding the learning process by specifying
486 background knowledge through domain ontologies (related to remote sensing,
487 images, environment, *etc.*). In return, the induced rules would contribute to
488 enrich the ontologies.

489 **Acknowledgements**

490 This work has been performed within the framework of the project CARTAM-
491 SAT (CARtographie du Territoire AMazonien: des Satellites aux AcTeurs
492 - Dynamic mapping of Amazonian Territories: from Satellites to Actors)
493 which is funded by the European Regional Development Funds (FEDER) for
494 French Guiana: agreement number 30492. The work was also supported by
495 the GEOSUD EQUIPEX Project.

496 **References**

497 Abudawood, T., Flach, P.A., 2011. Learning multi-class theories in ilp, in:
498 Proceedings of the 20th international conference on Inductive logic pro-
499 gramming, Springer-Verlag, Berlin, Heidelberg. pp. 6–13.

500 Andres, S., Arvor, D., Pierkot, C., 2012. Towards an ontological approach
501 for classifying remote sensing images, in: Signal Image Technology and
502 Internet Based Systems (SITIS), 2012 Eighth International Conference on,
503 pp. 825–832.

504 Anthony, E.J., Gardel, A., Gratiot, N., Proisy, C., Allison, M.A., Dolique, F.,
505 Fromard, F., 2010. The amazon-influenced muddy coast of south america:

- 506 A review of mud-bank-shoreline interactions. *Earth-Science Reviews* 103,
507 99–121.
- 508 Blaschke, T., 2010. Object based image analysis for remote sensing. *ISPRS*
509 *Journal of Photogrammetry and Remote Sensing* 65, 2–16.
- 510 Blockeel, H., Dzeroski, S., Kompare, B., Kramer, S., Pfahringer, B., Laer,
511 W.V., 2004. Experiments in predicting biodegradability. *Applied Artificial*
512 *Intelligence* 18(2), 157–181.
- 513 Chelghoum, N., Zeitouni, K., Laugier, T., Fiandrino, A., Loubersac, L.,
514 2006. Fouille de donnees spatiales - approche basee sur la programma-
515 tion logique inductive, in: 6emes Journées d’Extraction et de Gestion des
516 Connaissances, Edition CEPADUES. pp. 529–540.
- 517 Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational*
518 *and Psychological Measurement* 20, 37–46.
- 519 Cordier, M.O., 2005. Sacadeau: A decision-aid system to improve stream-
520 water quality. *ERCIM News* 61, 37–38.
- 521 Durand, N., Derivaux, S., Forestier, G., Wemmert, C., Gançarski, P., Bous-
522 said, O., Puissant, A., 2007. Ontology-based object recognition for remote
523 sensing image interpretation, in: *Proceedings of the 19th IEEE Interna-*
524 *tional Conference on Tools with Artificial Intelligence - Volume 01, IEEE*
525 *Computer Society, Washington, DC, USA.* pp. 472–479.
- 526 Fonseca, N.A., Silva, F., Camacho, R., 2006. April - an inductive logic
527 programming system, in: *JELIA*, pp. 481–484.

- 528 Forestier, G., Puissant, A., Wemmert, C., Gançarski, P., 2012. Knowledge-
529 based region labeling for remote sensing image interpretation. *Computers,*
530 *Environment and Urban Systems* 36, 470 – 480.
- 531 Fromont, E., Cordier, M.O., Quiniou, R., 2005. Extraction de connaissances
532 provenant de données multisources pour la caractérisation d'arythmies car-
533 diaques, in: *Fouille de données complexes. Cepaduès. volume RNTI-E-4 of*
534 *Revue des Nouvelles Technologies de l'Information*, pp. 25–45.
- 535 Goodacre, J., 1996. *Inductive Learning of Chess Rules Using Progol.* Oxford
536 University.
- 537 Hudelot, C., Atif, J., Bloch, I., 2008. Fuzzy spatial relation ontology for
538 image interpretation. *Fuzzy Sets and Systems* 159, 1929–1951.
- 539 Kavurucu, Y., Senkul, P., Toroslu, I., 2011. A comparative study on ilp-
540 based concept discovery systems. *Expert Systems with Applications* 38,
541 11598 – 11607.
- 542 Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for
543 categorical data. *Biometrics* 33, pp. 159–174.
- 544 Lavrac, N., Dzeroski, S., 1994. *Inductive Logic Programming: Techniques*
545 *and Applications.* Ellis Horwood.
- 546 Luu, T.D., Rusu, A., Walter, V., Linard, B., Poidevin, L., Ripp, R., Muller,
547 L.M.J., Raffelsberger, W., Wicker, N., Lecompte, O., Thompson, J.D.,
548 Poch, O., Nguyen, H., 2012. Kd4v: Comprehensible knowledge discovery
549 system for missense variant. *Nucleic Acids Research* 40, W71–W75.

550 Malerba, D., Esposito, F., Lanza, A., Lisi, F.A., Appice, A., 2003. Em-
551 powering a gis with inductive learning capabilities: the case of ingens.
552 Computers, Environment and Urban Systems 27, 265 – 281.

553 GRASS Development Team, 1999-2012. Welcome to grass gis.
554 <http://grass.fbk.eu/>.

555 Michalski, R.S., 1983. Machine learning: An artificial Intelligence Approach.
556 TIOGA Publishing Co.. chapter a theory and methodology of inductive
557 learning. pp. 110–161.

558 Muggleton, S., 1991. Inductive logic programming. New Generation Com-
559 puting 8, 295–318.

560 Muggleton, S., 1995. Inverse entailment and prolog. New Generation Com-
561 puting 13, 245–286.

562 Santos, J., Nassif, H., Page, D., Muggleton, S., Sternberg, M., 2012. Auto-
563 mated identification of protein-ligand interaction features using inductive
564 logic programming: A hexose binding case study. BMC Bioinformatics 13,
565 162.

566 Srinivasan, A., 2007. The aleph man-
567 ual. [http://www.cs.ox.ac.uk/activities/
568 machlearn/Aleph/aleph.html](http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html).

569 Srinivasan, A., Muggleton, S., Sternberg, M.J.E., King, R.D., 1996. Theo-
570 ries for mutagenicity: A study in first-order and feature-based induction.
571 Artificial Intelligence 85, 277–299.

- 572 Suzuki, H., Matsakis, P., Andrefouet, S., Desachy, J., 2001. Satellite image
573 classification using expert structural knowledge: A method based on fuzzy
574 partition computation and simulated annealing, in: IAMG 2001, pp. 251
575 – 268.
- 576 Vaz, D., Costa, V.S., Ferreira, M., 2010. Fire! firing inductive rules from
577 economic geography for fire risk detection, in: ILP, pp. 238–252.
- 578 Vaz, D., Ferreira, M., Lopes, R., 2007. Spatial-yap: a logic-based geographic
579 information system, in: Proceedings of the 23rd international conference
580 on Logic programming, Springer-Verlag, Berlin, Heidelberg. pp. 195–208.
- 581 Zaiche, L., Smith, A., 2011. 50% de satellites
582 en plus à lancer sur les dix prochaines années.
583 [http://www.perspectives-spatiales.com/sites/perspectives-spatiales.com](http://www.perspectives-spatiales.com/sites/perspectives-spatiales.com/files/50%25%20de%20satellites%20en%20plus%20sur%20la%20prochaine%20decennie.pdf)
584 [/files/50%25 de satellites en plus sur la prochaine](http://www.perspectives-spatiales.com/sites/perspectives-spatiales.com/files/50%25%20de%20satellites%20en%20plus%20sur%20la%20prochaine%20decennie.pdf)
585 [décennie.pdf](http://www.perspectives-spatiales.com/sites/perspectives-spatiales.com/files/50%25%20de%20satellites%20en%20plus%20sur%20la%20prochaine%20decennie.pdf).