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, 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

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

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

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

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

1. Introduction

The availability of remotely sensed Earth observation data, taken from aircrafts (including drones) and satellites, is constantly increasing. This obviously comes from the increasing number of Earth observation satellites and sensors. In fact, a recent report (Zaiche and Smith, 2011) estimates that the number of satellite launches will be 50% higher during the next ten years, when compared to the last decade. In particular, 200 governmental Earth observation satellites will be launched during that period. At the same time, as an increasing number of countries and/or organizations distribute remotely sensed data for free, the evolution in data distribution and use policies contributes to the use of huge volumes of data. Thus, data processing and interpretation have become a serious challenge for engineers and researchers. Therefore, classical procedures cannot continue to be used, and new approaches are needed to automatically update the land cover/use maps that provide essential information to decision makers.

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

A complementary approach to expert knowledge formalization is knowledge extraction from data. Such approach is utilized by all existing supervised image classification procedures, which first require a learning phase with delimitation and labeling (allocation to a class) of regions in the image. However, most methods consider only pixel information within such regions to separate and characterize the different classes. Structural aspects, i.e., information arrangement in space, are essentially taken into account by computing textural indexes within the same regions. To our knowledge, there is no operative tool that provides general and efficient classification rules exploiting structural knowledge at a higher semantic level, particularly at the object level within the object-oriented image analysis (Blaschke, 2010), when such knowledge is more robust and expressive than at the pixel level.

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

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

Malerba et al. (2003) implemented INductive GEographic iNformation System (INGENS) to assist with topographic map interpretation. INGENS consists of a prototypical extended Geographic Information System (GIS) with inductive learning capabilities. GIS classical functionalities are used to extract relevant concepts and features from spatial database, and the integrated inductive system allows finding rules to automatically recognize complex geographical contexts that are defined by the presence of specific geographical objects and their spatial arrangement in predefined spatial windows (cells). It is devoted to support map interpretation and geographical information retrieval by enriching geographical queries, but not to automatic classification in the context of large datasets. In fact, such automatic procedures require a quantitative evaluation that has not been performed with INGENS.

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

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

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

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

In such applicative and scientific contexts, we report here a method for structural and symbolic knowledge extraction from land use/cover maps and complementary geographic information layers, combined with a multi-class classification approach. Our work does not deal with the delimitation of regions (or segments) from images, but with the labeling of previously defined image regions. Methods intended to image region delimitation, including segmentation methods, are therefore beyond the scope of this study. In this study we chose the Inductive Logic Programming framework (ILP) (Muggleton, 1991) for the learning task, and a multi-class classification procedure developed by Abudawood and Flach (2011) within the ILP framework, i.e., the Multi-class Rule Set Intersection (MRSI). This methodology was tested to update land cover/use maps of the French Guiana coastline, and the resulting classification system was thoroughly evaluated from qualitative and quantitative points of view through a ten-fold cross-validation.

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

2. Materials and Methods

2.1. Inductive Logic Programming

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

ILP can model complex problems and has been used in several fields such as chemistry (Blockeel et al., 2004), biology, physics, medicine (Luu et al., 2012; Fromont et al., 2005), ecology and bio-informatics (Santos et al., 2012; Lavrac and Dzeroski, 1994; Srinivasan et al., 1996). It has, also, been applied to chess (Goodacre, 1996) and to test the quality of river water (Cordier, 2005). Very few studies have applied this method to geographical data, as already discussed in the introduction (Malerba et al., 2003; Vaz et al., 2007, 2010; Chelghoum et al., 2006).

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

Given:

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

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

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

2.2. Geographic information extraction and coding

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

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

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

$$\texttt{area\_symb}(O, I_k)\texttt{:-}\ \texttt{area\_num}(O, X), \quad X \leq p_k.$$
$$\text{or } \texttt{area\_symb}(O, S_k)\texttt{:-}\ \texttt{area\_num}(O, X), \quad X > p_k.$$

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

Eventually, the latitude and longitude values were used to characterize the 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**
`object(`O`)`	Declaration of the object O
`class(`O`,class_label)`	The object O belongs to the class `class_label`
`adjacent(`$O1$`,`$O2$`)`	$O1$ and $O2$ are two adjacent objects
`included(`$O1$`,`$O2$`)`	$O2$ is included in $O1$
`contains(`O`,`E`)`	O contains the entity E (*e.g.* $E \in \{River, Road, Building, ...\}$)
`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 ($X \in \Re$)
`compactness_num(`O`,`X`)*`	
`fract_dim_num(`O`,`X`)*`	
`perimeter_num(`O`,`X`)*`	
`area_symb(`O`,`I_k^{area} `or` S_k^{area}`)`	Recoding of the numeric variables according to the percentiles (see text for details)
`compactness_symb(`O`,`I_k^{comp} `or` S_k^{comp}`)`	
`fract_dim_symb(`O`,`I_k^{df} `or` S_k^{df}`)`	
`perimeter_symb(`O`,`I_k^{per} `or` S_k^{per}`)`	
`lat(`O`,`X`)*`	X is the latitude and longitude of O, respectively, ($X \in \Re$)
`long(`O`,`X`)*`	
`north(`$O1$`,`$O2$`):- lat(`$O1$`,A),lat(`$O2$`,B),A>B.`	$O1$ is located north, south, east and west of $O2$, respectively.
`south(`$O1$`,`$O2$`):- lat(`$O1$`,A),lat(`$O2$`,B),A`$\leq$`B.`	
`east(`$O1$`,`$O2$`):- long(`$O1$`,A),long(`$O2$`,B),A>B.`	
`west(`$O1$`,`$O2$`):- long(`$O1$`,A),long(`$O2$`,B),A`$\leq$`B.`	

2.3. Rule induction: one-vs-rest approach

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

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

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

2.4. Multi-class framework

Considering the previously described one-vs-rest approach results in inducing as many classifiers as classes. Considering the classifiers independently of one another, one or several classes can be predicted when a new object is to be classified. Abudawood and Flach (2011) proposed several solutions to handle multi-class problems for ILP. Among them, the Multi-class Rule Set Intersection (MRSI) method gave the highest accuracies and Areas Under the ROC Curve (AUC) when taking multi-class data sets into account (Abudawood and Flach, 2011). The principle of the MRSI method is: i) the theories induced for each class are gathered in an unique rule set; ii) for each rule i, the set of covered examples by the rule, C_i, is stored; iii) a default rule is formed that concludes to the majority class of the uncovered examples; iv) for an unseen object O, the intersection of the sets of examples covered by the fired rules is computed ($I = \cap C_i | r_i \; is \; fired$) and, finally; v) the predicted class $\hat{c}$ is the majority class in the set I, i.e., the more probable class given to the new object O, with an empirical probability $p(c|O)$.

2.5. Prediction evaluation

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

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

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

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

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

Overall accuracy, sensitivity, specificity and *Kappa* index values are computed for each test set, then averaged. The formulas of these measures are given hereafter.

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

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

where n is the number of classes, $TP^{(i)}$ the number of *true positives* for 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							
		C_1	...	C_{i-1}	C_i	C_{i+1}	...	C_n	Total
	C_1	$TN_1^{(i)}$	...	...	$FP_1^{(i)}$	...	...	...	E_1
	...	...	...	...	...	...	...	...	...
	...	...	...	$TN_{i-1}^{(i)}$	$FP_{i-1}^{(i)}$	...	...	...	E_{i-1}
Actual	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

For each class i, the sensitivity, i.e. the ability of the classifier to successfully classified positive examples, is computed as:

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

where $FN_j^{(i)}$ is the number of *false negatives* for the class i wrongly associated to the class j.

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

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

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

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

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

With $P(A)$ corresponding to the observed proportion of agreement between two classifications, and $P(H)$ the estimated proportion of agreement expected by chance.

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

The concepts and methods previously defined were applied to an actual geographic situation. The French Guiana territory is subject to intense anthropogenic and natural dynamics (Anthony et al., 2010): cyclic coastal erosion and accretion, notably due to the transport of sediments from the Amazon River by oceanic currents; and expansion of urban, peri-urban, agricultural areas. In this context, it is essential to develop automated methods for monitoring the land cover/use of the French Guiana territory. In particular, the large amount of available aerial photographs and satellite images is a critical source of materials that should be better exploited. If the delimitation of the geographical objects of interest does not require a high level of expertise and can be performed by operators, allocating these objects to land cover/use classes appears far more complex and subjective. In fact, despite efforts made to formalize and standardize the classification procedures, such allocating task requires a deep knowledge of the different types of land cover/use, both in the imaging and applicative domains. Consequently, the learning and classification methods previously presented were applied to automatically perform the labeling task and update the land cover/use maps of the French Guiana coastline.

3.1. Dataset

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

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

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

[1] https://www.seas-guyane.org

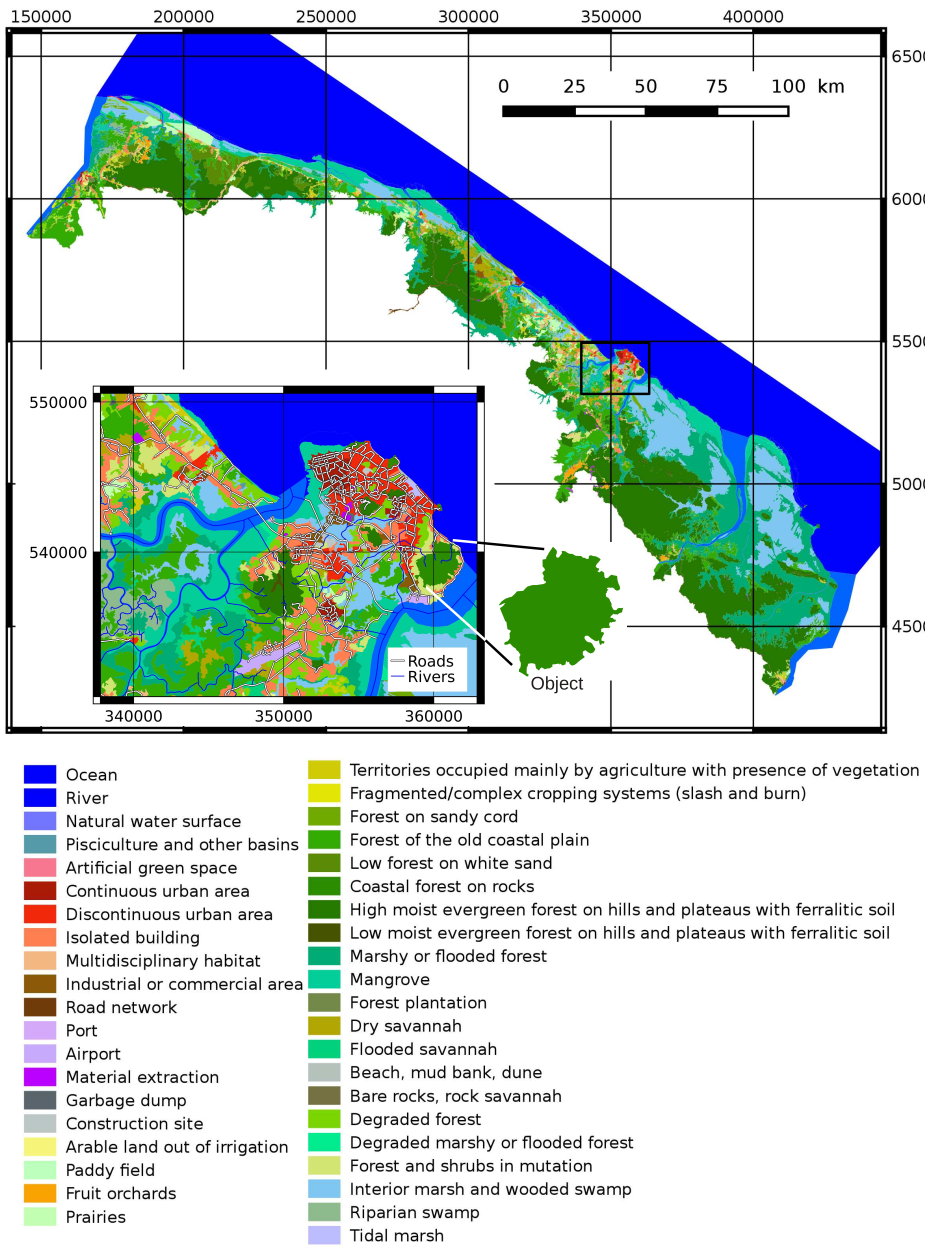


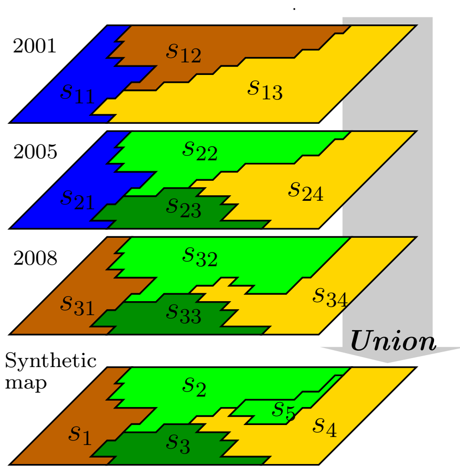
Figure 1: Land cover/use map and complementary geographic information layers (inset) used in this article (geographic coordinate system: WGS84 / UTM zone 22N). Sources: French National Office of Forests (Office National des Forêts; ONF); French National Geographic Institute (Institut Géographique National: 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.

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

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

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

In this study, we follow the land cover/use class of the objects in time. We do not explicitly follow the object delimitations, which is a much more complex task. In fact, by taking into account the information provided by three original maps, object boundaries can change in time: an object can be splitted into two or more objects belonging to different classes (see for instance object s_{13} in figure 2), creating new object(s); an object can result from the merging of several objects, making one or several objects disappear. We handled such situations by generating objects with invariant boundaries in time and related to an unique class for each year. Practically, we produced a synthetic map by concatenating the information contained in the three original maps, by means of the "union" GIS operator, as schematically shown in Figure 2. The elementary geographical entities of the resulting map are referred to as *objects* thereafter, and contribute to define the *examples* in the ILP process.



Synthetic map attribute table

Object	Class (2001)	Class (2005)	Class (2008)
s_1	blue	blue	brown
s_2	brown	light green	light green
s_3	orange	dark green	dark green
s_4	orange	orange	orange
s_5	orange	orange	light green

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

3.3. Information coding

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

Given the diachronic characteristics of the data, 3 predicates were defined to indicate the class of an object as a function of the time: `class_`y_0`(`O`,class_name)`, `class_`y_{-3}`(`O`,class_name)` and `class_`y_{-6}`(`O`,class_name)`, indicating the land cover/use class of the object O for the years y_0, y_{-3} and y_{-6}, respectively, *i.e*, for 2008, 2005 and 2001. It is worth noting that from a relative point of view, the year 2001, seven years prior 2008, is assumed to actually correspond to the sixth year before the reference year y_0. In fact, we can assume marginal changes between 2001 and 2002. However, this assumption has also a practical justification as it permits to consider the updating of the land cover/use information every three years based on the maps established three and six years before.

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

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

3.4. Rule induction: Aleph parametrization

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

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

4. Results

4.1. Set of induced rules

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

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

(1) (Pos cover = 472; Neg cover = 88; Total pos. ex. = 552)
class_y_0(A, Multidisciplinary habitat) :- area_symb(A, $\leq$165567), adjacent(A, B), class_y_{-3}(B, Multidisciplinary habitat).

(2) (Pos cover = 2 Neg cover = 0 Total pos. ex. = 40)
class_y_0(A,Industrial or commercial area) :- adjacent(A, B), class_y_{-6}(B, Construction sites), area_symb(A, $\leq$10831).

(3) (Pos cover = 3 Neg cover = 0 Total pos. ex. = 166)
class_y_0(A, Discontinuous urban area) :- class_y_{-6}(A, Construction sites), area_symb(A, $\leq$76202), area_symb(A, >10831).

Rule (1) covers 472 positive examples for a total of 552 objects actually belonging to the class of interest (85.5%) and 88 negative examples. It indicates that an object will belong to the *Multidisciplinary habitat* class if its area is less than or equal to 165 567 m^2 and is adjacent to an object belonging to the same class three years before. Rule (2) indicates that an object will belong to the *Industrial or commercial area* class if its area is less than or equal to 10 831 m^2 and is adjacent to an object belonging to the class *Construction sites* 6 years before. Rule (3) indicates that an object will belong to the *Discontinuous urban area* class if its area, in m^2, belongs to the interval $]10831, 76202]$ and if it belonged to the class *Construction sites* 6 years before. By considering such rules for the characterization of the territory dynamics, the first rule illustrates the extension dynamics of the natural areas whereas the second and the third rules describe the extension dynamics of the anthropogenic areas.

4.2. Prediction evaluation

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

All classifiers were 100% specific, except for one related to the class *Forest 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 and semi-natural area	Forest	Moist evergreen forest of the main-land coastal plain	Low forest on white sand	41.7	24	1
Artificial Territories	Mine, garbage dump or construction sites	Garbage dump		25.0	15	1
Artificial Territories	Mine, garbage dump or construction sites	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
	Forest	Moist evergreen forest of the mainland coastal plain	Coastal forest on rocks	70.0	14	3
			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	10
	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
	Industrial zone	Airport		100.0	12	1
Agricultural Territories	Permanent cultivation	Fruit orchards		87.1	259	1
	Heterogeneous agricultural areas	Fragmented/complex cropping systems (slash & burn)		81.9	814	6
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
		Moist evergreen forest on hills and plateaus with ferralitic soil	Low forest	98.0	58	1
		Marshy or flooded forest		91.7	288	5
		Mangrove		93.0	259	16
	Shrubby environment	Dry savannah		93.9	164	1
		Flooded savannah		92.0	98	3
	Open space with some/no vegetation	Bare rocks, Rock savannah		100.0	6	1
	Degraded natural environment	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
	Marin Wetland	Tidal marsh		88.9	9	1
Water surface	Continental water	Pisciculture and other basins		85.0	18	1
		Natural water surface		100.0	4	1

Table 6 summarizes the results for overall accuracy and *Kappa* Index. Overall accuracy values varied from 82.4% to 87.3% with an average of 84.6%. *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)									

4.3. Map of prediction errors

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

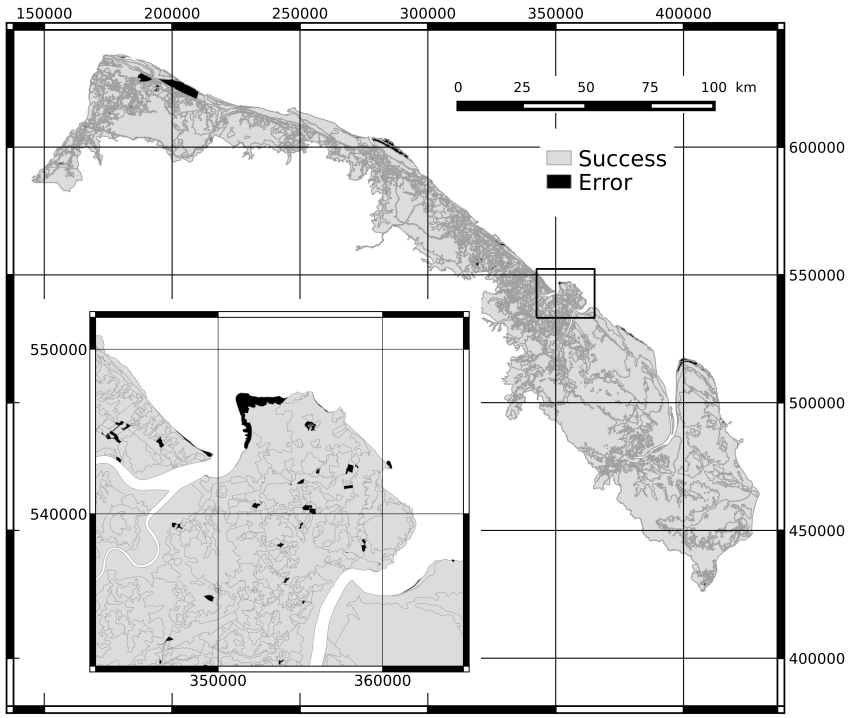


Figure 3: Map of prediction errors (geographic coordinate system: WGS84 / UTM zone 22N). Map at the top represents French Guiana coastline; Map in the inset zooms in on the "Cayenne Island".

5. Discussion

From a qualitative point of view, induced rules are consistent with the observed environmental features and dynamics of the study area. Moreover, they are provided in an expressive formalism, and are easily understandable and interpretable by non-experts, as they can be expressed in natural language. However, some rules covered very few (2 or 3) positive examples, whereas the total number of positive examples for the associated classes was large (see rule (3) in paragraph 4.1 for example). Such rules were consequently very specific and did not represent a significant knowledge within the application domain.

The predicates *south*, *north*, *east* and *west* did not appear in the rules, showing that such predicates were not pertinent for object discrimination, and that characterization of the objects should make better use of expert knowledge. In particular, domain ontologies could guide the learning process by identifying the predicates and the learning constraints to use.

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

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

Territories occupied mainly by agriculture with presence of vegetation, which is a complex class including traditional itinerant slash and burn activities that consist in cultivating an area and then letting the natural vegetation to regenerate. This seems to indicate that the information provided by the land cover/use maps is insufficient in terms of anteriority and/or time resolution for these classes. However, prediction performances could be improved. In fact, background knowledge can be enriched by adding predicates, possibly evaluated from complementary geographic information layers (digital elevation model, soil map, *etc.*). As already mentioned, the choice of these complementary object features can be guided by expert knowledge, notably through domain ontologies. Better performances could also be obtained by implementing different learning and classification strategies: in our case, *a priori* known classes at year y_0 could be exploited to learn more efficient rules. These classes should be the most stable in time and the easiest to identify (e.g. *River*, *Continuous urban area*, *Airport*, *etc.*). An iterative learning-classification strategy could also be implemented, by: i) first learning and classifying classes associated with high-performance predictions (e.g. *Forest and shrubs in mutation*, see Table 5); ii) then using the prediction to enrich the background knowledge of other classes; iii) learning-classifying these classes; iv) repeating the procedure until all classes are predicted. However, the number of strategies is such that we must rely on objective criteria and/or intensive simulations to determine the most appropriate one.

Nevertheless, our method gave good results globally. In fact, in addition to the excellent sensitivity and specificity values returned by the procedure, the *Kappa* Index and overall accuracy values were high. According to the Kappa interpretation table by (Landis and Koch, 1977), these values denote "strong agreement" between predicted and actual classes.

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

Inductive Logic Programming is devoted to symbolic data. The management of numeric information by ILP constitutes a specific research field, which is beyond the scope of this paper. However, several simple solutions exist in order to code the numeric data into symbolic ones. In fact, the domain of values of a numeric variables can be categorized by means of crisp or fuzzy modalities. We propose here to code the numeric information by means of inequalities taking into account quantiles of the numeric variable empirical distribution. This enables Aleph to manage numeric information in a manner comparable to the Confidence-based Concept Discovery (C^2D) ILP system (Kavurucu et al., 2011). This solution seems to offer a good compromise between information loss and generalization capacity, by allowing the system to automatically discover significant value intervals (see rule (3) in the Results section).

Finally, the method proposed here does not consider the image processing step devoted to the delimitation of the regions of the image that define the objects. It only considers the labeling (or classification) of the regions. This implies: that the partitioning of the image into regions is performed beforehand, by means of any methods including fully manual ones (photo-interpretation) or automatic image segmentation algorithms; that the new objects, which labels have to be predicted, have been delimited by the method that produced the objects used for the learning task of the classification rules.

6. Conclusion

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

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

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

Acknowledgements

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

References

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

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

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

A review of mud-bank-shoreline interactions. Earth-Science Reviews 103, 99–121.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, 2–16.

Blockeel, H., Dzeroski, S., Kompare, B., Kramer, S., Pfahringer, B., Laer, W.V., 2004. Experiments in predicting biodegradability. Applied Artificial Intelligence 18(2), 157–181.

Chelghoum, N., Zeitouni, K., Laugier, T., Fiandrino, A., Loubersac, L., 2006. Fouille de donnees spatiales - approche basee sur la programmation logique inductive, in: 6emes Journées d'Extraction et de Gestion des Connaissances, Edition CEPADUES. pp. 529–540.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46.

Cordier, M.O., 2005. Sacadeau: A decision-aid system to improve stream-water quality. ERCIM News 61, 37–38.

Durand, N., Derivaux, S., Forestier, G., Wemmert, C., Gançarski, P., Boussaid, O., Puissant, A., 2007. Ontology-based object recognition for remote sensing image interpretation, in: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 01, IEEE Computer Society, Washington, DC, USA. pp. 472–479.

Fonseca, N.A., Silva, F., Camacho, R., 2006. April - an inductive logic programming system, in: JELIA, pp. 481–484.

Forestier, G., Puissant, A., Wemmert, C., Gançarski, P., 2012. Knowledge-based region labeling for remote sensing image interpretation. Computers, Environment and Urban Systems 36, 470 – 480.

Fromont, E., Cordier, M.O., Quiniou, R., 2005. Extraction de connaissances provenant de données multisources pour la caractérisation d'arythmies cardiaques, in: Fouille de données complexes. Cepaduès. volume RNTI-E-4 of *Revue des Nouvelles Technologies de l'Information*, pp. 25–45.

Goodacre, J., 1996. Inductive Learning of Chess Rules Using Progol. Oxford University.

Hudelot, C., Atif, J., Bloch, I., 2008. Fuzzy spatial relation ontology for image interpretation. Fuzzy Sets and Systems 159, 1929–1951.

Kavurucu, Y., Senkul, P., Toroslu, I., 2011. A comparative study on ilp-based concept discovery systems. Expert Systems with Applications 38, 11598 – 11607.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, pp. 159–174.

Lavrac, N., Dzeroski, S., 1994. Inductive Logic Programming: Techniques and Applications. Ellis Horwood.

Luu, T.D., Rusu, A., Walter, V., Linard, B., Poidevin, L., Ripp, R., Muller, L.M.J., Raffelsberger, W., Wicker, N., Lecompte, O., Thompson, J.D., Poch, O., Nguyen, H., 2012. Kd4v: Comprehensible knowledge discovery system for missense variant. Nucleic Acids Research 40, W71–W75.

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

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

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

Muggleton, S., 1991. Inductive logic programming. New Generation Computing 8, 295–318.

Muggleton, S., 1995. Inverse entailment and progol. New Generation Computing 13, 245–286.

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

Srinivasan, A., 2007. The aleph manual. http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html.

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

Suzuki, H., Matsakis, P., Andrefouet, S., Desachy, J., 2001. Satellite image classification using expert structural knowledge: A method based on fuzzy partition computation and simulated annealing, in: IAMG 2001, pp. 251 – 268.

Vaz, D., Costa, V.S., Ferreira, M., 2010. Fire! firing inductive rules from economic geography for fire risk detection, in: ILP, pp. 238–252.

Vaz, D., Ferreira, M., Lopes, R., 2007. Spatial-yap: a logic-based geographic information system, in: Proceedings of the 23rd international conference on Logic programming, Springer-Verlag, Berlin, Heidelberg. pp. 195–208.

Zaiche, L., Smith, A., 2011. 50% de satellites en plus à lancer sur les dix prochaines années. http://www.perspectives-spatiales.com/sites/perspectives-spatiales.com/files/50%25 de satellites en plus sur la prochaine décennie.pdf.