Assessment of classification methods for soil erosion risks

[1]Gay, D., [2,5]Rouet I., [4]Mangeas M., [3,4]Dumas P., [1]Selmaoui N.

[1]ERIM, [2]EA3325, [3] CNEP-ETAPP,
University of New-Caledonia, BP R4 F-98851 NOUMEA Cedex, New-Caledonia

[4]ESPACE 140, [5]UMR 161 - CEREGE,
Research Institut for Development, BP A5 F-98848 NOUMEA Cedex, New-Caledonia

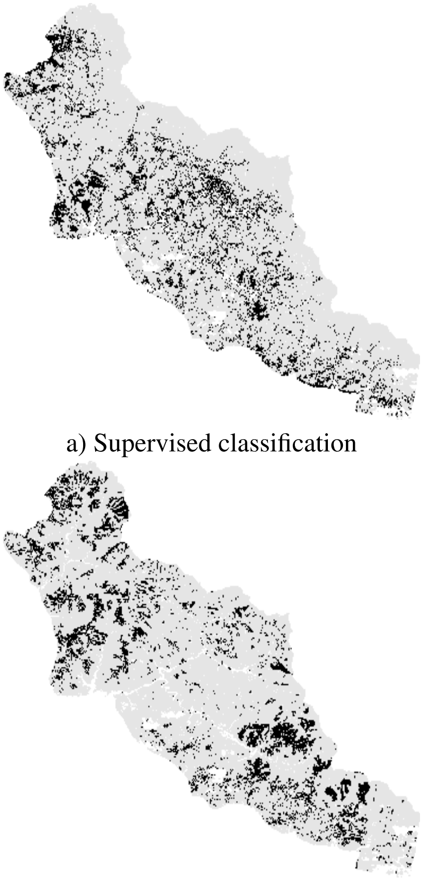
Keywords: *Soil erosion mapping, knowledge discovery, classification*

EXTENDED ABSTRACT

Soil erosion is a serious problem in New-Caledonia mainly due to the cyclonic tropical weather, bush fires and human activities (openpit mining). Runoff on stripped soils causes degradations for anthropic laying out on high ultramafic terranes and pollution (mobilized soil particles) which modifies the coastal region and degrades the coral reefs by hypersedimentation processes. All the human activities chain that depend on natural resources regularly suffer the effects or the consequences of these phenomena. Geologists and geographers need to identify and rank erosion parameters and define sensitivity maps, particularly for decision-makers. For a peculiar region of new-caledonia, eight identified factors such as relief parameters, land use, geological substrate type and precipitations are available and the area of interest has been already studied and classified by experts from very sensitive to very robust at soil erosion.

We propose in this paper to study classification methods in order to set soil erosion sensitivity maps. The first step consists in data mining work for knowledge discovery. A systematic cartography of actually stripped soils is used in this work as learning data. We have highlighted that used methods for data analysis are really suitable tools for experts because results are in closed correlation with their terrain observations when results are average for only mathematical analysis. These methods need very important exchanges between geologists/geographers experts and computer scientists to identify the real gain that knowledge discovery methods can produce.

Several supervised classification methods such the bayesian classifier, decision trees are then used to predict erosion hazard on the basis of knowledge discovery results. Predicting results are described and compared to a linear combination of the factors based on expert knowledge (cf. Fig. 1). The output models (the soil erosion sensitivity maps) are then examined and commented from a geologic point of view. Once more we identify an improvement in the results, especially in the spatial repartition of erosion sensitive areas which are more similar with terrain observations than actuals experts predictive models. With these statistical approaches we are able to propose improved prediction tools for soil erosion sensitivity and on the top of that experts can access to attribute ranking and association rules with a confidence indicator, thing for which classical expert models give reliable results with difficulty. Knowledge discovery and prediction model construction constitute very interesting new methods of analysis for many specialists as in environmental and natural sciences as in fundamental or applied research.



a) Supervised classification

(b) Expert approach

Figure 1. Comparison between the two approaches. In very black: the potentially erodible surfaces

1 INTRODUCTION

1.1 Soil erosion

The ”Grande Terre” of New Caledonia is a vast high island where many areas are made up of deeply weathered geological materials. Soils develop on this interface and are therefore the first target of erosion in a cyclonic regime of precipitations. Particularly, the effect of runoff increases when degradations of protective plant cover appear after bush fires or nickel openpit mining, the two most common events on the high terranes of this land. This disappearance of plant cover causes all the more problems that many endemic vegetables – this country is also a hot spot of biodiversity – grow again very slowly on specific soils, due to high levels of natural phytotoxic elements like nickel, cobalt or manganese. Runoff mobilized soils particles from high terranes pass quickly through creeks and rivers, stay in the cultural plain when inundations occur and stop their way in the closed lagoon after just fewer 20 to 30 kilometers since soil removal. Then coral reef and lagoon life is disrupt too by sporadic episodes of hypersedimentation. All human activities chain that depend on natural resources are regularly affected by soil erosion and consequences of this phenomenon.
Experts are very often consulted about the relationships between soil erosion and mining activities but if few events are good examples of a direct relation, many others cases of erosion are more difficult to be interpreted and many others prove indisputably a natural origin of erosion. These different ways of erosion evolve at different timescale too and human degradations very intensify erosive processes that naturally occurs at geological scale. Erosion issues in New Caledonia actually need fundamental studies to understand how soil loss impacts environment and landscape evolution and the very quickly development of the country and conversely what is the part of human responsibility in erosion phenomena.

Up to now, general experts knowledge and local studies are available but no regional approach have been designed. We present in this paper a methodology that can be useful for processing the potential soil erosion mapping at a regional scale.

1.2 Knowledge discovery and prediction for soil erosion

Knowledge Discovery in Databases (KDD) is a complex interactive and iterative process. Past studies enlightened several Data Mining tasks. In our context, we will focus on association rules mining task and prediction model construction. Note that our model does not take into account spatial dependance in data but provides preliminary results using well-known data mining and machine learning techniques.

2 DATA AND METHODS

2.1 Physic parameters

Experts have determined 8 physic parameters [6] like erosion factors in New Caledonia: 4 relief parameters, geological and vegetation factor, trails and precipitations.
Relief parameters have been calculated with numeric topographical map 1/10000 [2]. The first step consists in processing pseudo-continuous spatial data (Digital Elevation Model) that give regular values of elevation. Slope, planform and profile curvatures are then computed with morphometric formulas on DEM for each location. Flow accumulation is another parameter derived from DEM, calculated with application of a 3x3 kernel to know how precipitations are drained on the topographical surface.

Geological factor is based on lithology information obtained from the geological map 1/50000 [8]. Vegetation data have been designed by aerial photo-interpretation and remote sensing analysis ([2], [3]).

Finally, trails are made using the topographic map 1/10000 [2] and precipitations data are simulated by Meteo France during the 1991-2000 period.

All of these raster data take place into a Geographical Information System (GIS) in order to easily cross, analyse and visualize the various information.

2.2 Field data

In order to proceed to the classification, we need training data able to specify the repartition of breaked out erosion phenomena.

At this time, only one type of data is available and gives a real indicator of soil erosion for each location of the study area The New Caledonia stake holders have order an inventory of mining impact on soil denudation in ultramafic terranes of the ”Grande Terre” with satellite SPOT 5 data. These particular rocks represent one of the biggest global nickel ore reservoir. It covers around 1/3 of the island and is being intensively mined. A process has been defined ([7], [10]) to extract an indicator of soil denudation and a systematic aerial photo-interpretation proved that results give a good indicator of erosion locations for any origin of the phenomenon. This information stands for a good erosion indicator which represents reality of soil erosion in our area of interest.

All data (physic parameters and field data) are processed to appear as grids of 30 x 30 meters homogeneous cells. Each cell can then be classified as an object described by attributes.

2.3 Knowledge discovery

The different steps of our KDD-based methodology for erosion data analysis are detailed in Figure 2. In this subsection, we give further details on each step of the process.

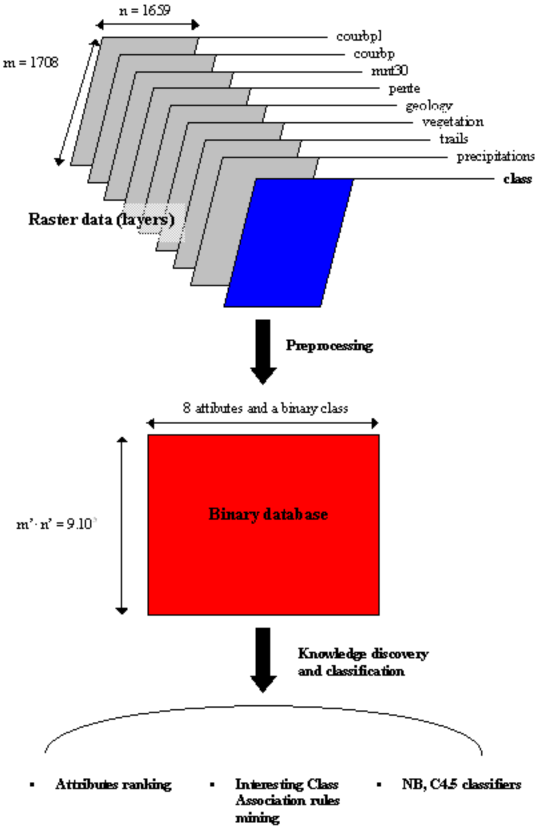


Figure 2. Knowledge discovery process

2.3.1 Pre-processing data

Raster data consist in 1659 × 1708 points. When focusing on working data (removing data concerning sea, clouds and missing values) and translating the nine layers into table data, we obtain about 9.10^5 objects (lines) described by eight attributes (columns) and a binary class (erosion/no-erosion). A brief descriptive summary of data is given in Table 1. Data are clearly unbalanced: there are about 870 thousand points ($\sim 97\%$) considered as not erodible and only 30 thousand erodible points ($\sim 3\%$).

Attributes	Range of values
precipitations	$[829; 3487]$
Profile Curvature	$[-35.26; 8.179]$
Planform curvature	$[-26.22; 20.088]$
geology	nominal (28 values)
DEM30	$[5; 1604.99]$
vegetation	nominal (9 values)
slope	$[0; 349.137]$
trails	binary
class	binary

Table 1. Erosion table data description.

2.3.2 Discretization

Erosion data contains both continuous and nominal attributes. Relief parameters and precipitations attributes are continuous and others (vegetation and geological factors and trails attributes) are nominal ones. While nominal attributes are trivially translated into binary ones, numerical attributes are first discretized with Fayyad & Irani entropy-based method [4] then binarized.
To discretize a continuous attribute A, Fayyad & Irani recursively choose the best cut point in range values of A w.r.t. class entropy criterion (i.e. the one that minimizes class entropy) until information gain goes under a certain threshold (see [4] for more details).

Definition 1 (Entropy, class Entropy)**.** *For a table data M, an attribute A and a cut point T in range values of A,* entropy *function and* class entropy *are defined as follows :*

$$E(M) = -\sum_{j=1}^{j=nbc} P(j|M) \times \log_2(P(j|M))$$
$$cE(A,T,M) = \frac{|M_1|}{|M|}E(M_1) + \frac{|M_2|}{|M|}E(M_2)$$

where $P(j|M)$ is the proportion of objects of M classified as j and nbc the number of classes. Intuitively, entropy is minimal when all objects of M are in the same class and is maximal when objects of M are equi-partitioned in classes.

2.3.3 Attributes ranking

We are interested in confirming experts opinions about following questions :

- Which ones of the attributes have an important role in the erosion process ?
- Which one is the most relevant ?

As a preliminary task, attributes ranking w.r.t. objective measures gives elements for answering. We used Information Gain (IG), Gain Ratio (GR) as useful objective measures for attributes utility.

Definition 2 (Information gain). *Information Gain (IG) is an entropy-based objective measure for attributes defined as follows :*

$$IG(M, I) = E(M) - \sum_{k=1}^{n} (P_k \times E(M_k))$$

where M is the data matrix, I a k-values attribute, P_k the proportion of objects of M satisfying k^{th} value of I and M_k the submatrix of objects of M satisfying k^{th} value of I.

Intuitively, higher values of $IG(M, I)$ mean that M_k is well partitioned into classes and I shall be considered as an interesting attribute.
GR is based on IG. GR is weighted by Split Information function that gives more importance to attributes with less different values (see [9] for more details).

2.3.4 Association rules mining.

From a binary database, we can mine frequent patterns to discover trends in databases. We can also mine frequent associations between different sets of attributes of a database. In this paper, we are interested in frequent associations between a set of attributes and a class label. This will help us to answer questions such as:

- What are the most significant features of an erodible area ?
- Are there typical erodible/non-erodible areas ?

Definition 3 (Association rule). *An* association rule *π is an implication of the form $X \Rightarrow Y$ where X (condition) and Y (consequence) are different sets of attributes. When Y is a class attribute, π is a class association rule.*

The set of extracted association rules could be huge and among them, a few are interesting. Association rules interest can be estimated with two measures : *Support* (the frequency in data) and *confidence* (frequency of π over frequency of X in data). Intuitively, $\pi : X \Rightarrow c$ with support 10% means that 10% of objects match the rule, and with a confidence 0.7 indicates that when X is verified, c is verified with confidence degree 0.7 (at most 30% of violations of the rule). Since [1], given support and confidence thresholds, extraction of such rules can be managed efficiently with A-PRIORI algorithm.

2.4 Prediction model construction

Given a labeled training database M, the goal in prediction task is to build a classifier that classify well novel unseen objects. In this paper, we focus on two well-known methods : (1) Naïve Bayes classification rule (NB [5]), (2) C4.5 decision tree [9]. Prediction task will help expert to build new maps for study and will allow them to compare learning model to their own expert model.

Naïve Bayes classifier. NB classification rule is defined as follows :

$$\mathcal{C}_{Bayes}(t) = \underset{j=1,\ldots,nbc}{\text{argmax}} \prod_{i=1,\ldots,m} \hat{P}(t_i \mid j) \times \hat{P}(j)$$

where $t = \{t_1, \ldots, t_m\}$ is an unseen object, nbc is the number of classes, m the number of attributes. $\hat{P}(j)$ is the proportion of objects of class j in M. Assuming independence of attributes, $\hat{P}(t_i \mid j)$, the proportion of objects of class j having t_i value for i^{th} attribute, is an estimation of probability $P(t_i \mid j)$.

Decision tree. Decision tree is made of leaves and nodes. A leaf indicates a class label and a node (test node) is a fork with a threshold applied to an attribute value that defines the two possible branches. To classify an unseen object, one starts from the root and moves down (with respect to the test nodes) until reaching a leaf. The class label of the leaf is the prediction.

Decision tree construction process is recursive: at each step the most class-discriminant attribute A is chosen with respect to a certain measure (GR for C4.5), then data is split with respect to the values of A into as many groups as values of A. Next and recursively, for each group, a new relevant attribute is chosen to further separation until a certain stop condition is reached (depending on algorithms and options).

3 EXPERIMENTS

We used WEKA platform [11] to perform our experiments. Erosion database is made of a set of numerous data ($\sim 9.10^5$ described by 8 attributes). When all attributes binarized, we obtain about 139 binary attributes. The attributes ranking are performed on the whole database and the results can be seen in Table 2 (original group). It emerges that attributes vegetation, slope and geology are more important with respect to both IG and GR measure.

But some data mining tasks (such as association rules mining) become hard on such a database. Let D_n be the set of objects of the major class (non-erodible) and D_e the set of objects of minor class (erodible). $\mathcal{D} = D_n \cup D_e$. We know that $\mathcal{D}$ is huge and unbalanced ($D_n \gg D_e$. To face up

	IG	GR
Original	vegetation (0.05995)	vegetation (0.03129)
	slope (0.03705)	trails (0.02821)
	geology (0.03506)	geology (0.02227)
	DEM30 (0.02617)	slope (0.01286)
	rainfall (0.0158)	DEM30 (< 0.01)
	Profile curvature (< 0.01)	rainfall (< 0.01)
	trails (< 0.01)	Planform curvature (< 0.01)
	Planform curvature (< 0.01)	Profile curvature (< 0.01)
with slope > 15	vegetation (0.039883)	trails (0.02922)
	geology (0.01641)	vegetation (0.02131)
	rainfall (0.00859)	geology (0.01116)
	slope (0.00829)	slope (0.00317)
	trails (0.00707)	rainfall (0.00299)
	DEM30 (0.00402)	DEM30 (< 0.001)
	Profile curvature (< 0.001)	Planform curvature (< 0.001)
	Planform curvature (< 0.001)	Profile curvature (< 0.001)

Table 2. Attributes ranking results.

to the database size, we built 10 smaller databases $\mathcal{D}_i, i = \{1, 2, \dots, 10\}$ such that $\mathcal{D}_i = D_{n,i} \cup D_e$. Now, $|\mathcal{D}_i| \simeq 60000$ and data mining tasks can be efficiently performed. Class association rules extraction are then processed on each D_i (in Table 3 where only non-redundant rules appearing in ten extractions are reported with average confidence). It shows an unexpected association : $\pi : slope[0; 7] \Rightarrow e$ which is not intuitive.

$X \Rightarrow$ e(rodible)	$X \Rightarrow$ n(on erodible)
$veg = 14 \Rightarrow e(0.97)$	$veg = 1 \Rightarrow n(0.85)$
$geol = 20 \Rightarrow e(0.96)$	$veg = 2 \Rightarrow n(0.77)$
$slope[0; 7] \Rightarrow e(0.96)$	$geol = 19 \Rightarrow n(0.69)$
$\dots$	$\dots$

Table 3. Frequent (10%) and confident (0.5) class association rules. The term veg denotes $vegetation$ and $geol$ denotes $geology$

Many points of learning data have very gentle slopes, due to remote sensing method which do not make valley floors different from stripped grounds, because objective was to consider all erosion phenomena (removal, transport and deposit). The main attribute values which characterize $X \Rightarrow e$ in our first experiments are controlled by valley floors in this landscape: any vegetation ($vegetation = 14$), alluvial areas ($geology = 20$) and very gentle slopes. In this paper, we essentially focus on potential removal occurrences.

In a second phase, only objects with slopes > 15 are considered. Attributes ranking results are reported in Table 2 (group $slope > 15$). In this case, main parameters in $X \Rightarrow e$ change (see Table 4). If lack of vegetation still first characterize erodible areas, the two others are different: slope seems not be as important as in first experiments and alluvial areas turn into laterites ($geology = 2$) even if geology remains as the second main parameter. The third main attribute value becomes presence of trails ($trails = 1$). It has to be noted that if slopes values have an important impact on erodible characterization ($slope < 15$ or > 15 for $X \Rightarrow e$), there is any change for $X \Rightarrow n$ (non erodible areas): presence of vegetation (dense forest = 1, maquis forest =1) is the most important parameter to prevent erosion, especially in the case of harzburgites outcrop ($geology = 19$).

$X \Rightarrow$ e(rodible)	$X \Rightarrow$ n(onerodible)
$veg = 14 \Rightarrow e(0.97)$	$veg = 1 \Rightarrow n(0.86)$
$geol = 2 \Rightarrow e(0.73)$	$veg = 2 \Rightarrow n(0.76)$
$trails = 1 \Rightarrow e(0.86)$	$geol = 19 \Rightarrow n(0.67)$
$\dots$	$\dots$

Table 4. Frequent (10%) and confident (0.5) class association rules.

For prediction task, we used 10 folds cross-validation on each D_i and reported average estimated accuracy over D_i and per class average accuracies (see Table 5). The no-erodible class is particularly sensitive. For example, an estimated error of 18.89% for the class means that 151120 erodible pixels (18.89% of 800000 pixels) are misclassified.

4 COMPARISON WITH AN EXPERT APPROACH

The expert model used to compare results of classification methods is recent model proposed for neo-caledonian ultramafic terranes. It is based on

NB	Real no-erodible	Real erodible	All
Predicted no-erodible	**81.11**%	18.89%	**83.49**
Predicted erodible	14.14%	**85.86**%	

C4.5	Real no-erodible	Real erodible	All
Predicted no-erodible	**84.91**%	15.09%	**86.44**
Predicted erodible	12.05%	**87.95**%	

Table 5. Confusion matrix for NB and C4.5 classifiers.

expert segmentation for each parameter which is considered as relevant for erosion occurrence by geographers, geologists and hydrologists. Parameters ranking comes from knowledge of this experts. Physic attributes are weighted in relation with their ranking: 1 for the main parameters (slope and rainfall), 1/2 (flow accumulation, geology, vegetation), 1/5 (profile and horizontal curvatures) or 1/10 (trails) [6].

Classification methods used in this frame can predict potential erosion and provide erosion hazard results. First, they identify physical configurations for erosion occurrences (and non erosion too) and then highlight this sites into the entire region. So resulting models contain more points in erodible class than the learning data which only give effective erosion, not potential erosion sites.

NB	Expert ne	Expert e	All
classif ne	**86.72**%	70.18%	**77.59**
classif e	13.28%	**29.82**%	

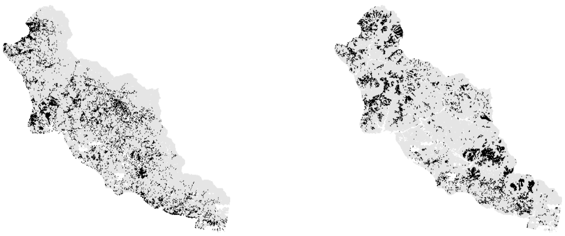
Table 6. Confusion matrix for the Classification approach versus the Expert approach. The term *e* (resp. *ne*) denotes *erodible* (resp. *non erodible*)

From the comparative assessment (Table 6) between classification methods and expert model, it emerges that it exists a poor overlapping for the erodible class (30% of common surface).

We may logically announce that we are very disappointed by this results. Yet many signs let us suppose that classification methods can be used for erosion prediction. Attribute ranking and main association rules show that vegetation always control occurrence or lack of erosion (only 1/2 for expert model) whereas slope $>$ 15 (1 for experts) appears just at the 4^{th} rank (IG and GR) and do not comes up as a main attribute in the association rules. Moreover, when the precipitations factor is weighted 1 by the experts, the classifiers moderate its position (3th). Finally, the geology attribute is ranked to a similar place by the expert approach and the classification methods and their values highlighted by association rules correspond to the ground measurements. Similarly, the same kind of relationship for curvatures can be deduced, even if it is considered for both methods as a lesser important attribute.

The case of trails is also interesting. The main association of factor for erodible surfaces appears to be the stripped ground / laterite / trail which is perfectly known by terrain workers as the critic configuration for erosion occurrence. So despite the fact that the trail by itself does not emerge as a important factor, it appears in a association well identified. Conversely, the association rules for non erodible surface: unweathered hard rocks under dense or maquis forest never eroded, seems also very pertinent.

In order to understand the spatial repartition of success and failures, four types of surfaces have been studied: (a) points are classified as erodible for all methods, (b) points are estimated as non erodible for all methods, (c) points are considered as erodible for classifications but not for the expert model, (d) points are considered as erodible for expert model but not for classifications (see Figure 3). This spatial analysis allows to give some hints about the observed differences: results (a) mainly correspond to known eroded points, (c) points are mostly isolated while (d) points lie in homogeneous areas.



(a) Supervised classification (b) Expert approach

Figure 3. Comparison between the two approaches. In very black: the potentially erodible surfaces

5 CONCLUSION AND FUTURE WORK

These work present new opportunities for experts to improve their knowledge of erosion occurrence by non cognitive methods. Indeed, the apparent poor global results not illustrate the benefits that experts can gain from the classification methods. We have seen how association rules results are pertinent and cartographic representation of comparison results can suggest over/underestimations of parameters in the expert model. Many points of comparison allow us to think that using classification methods for erosion prediction is an interesting way to contribute to the improvement of our knowledge on erosion phenomenon in specific land, particularly when only general knowledge exists for understanding erosive processes.

This first prospective results would help experts to adjust their erosion hazard model (weighting and decision rules) to subsequently propose optimized risk maps for decision-makers.

As future work, we plan to investigate spatial dependence in erosion data and build new models with spatial data mining techniques. Another way for future work is to mixe statical approach and expert approach in order to converge toward a more robust model..

6 ACKNOWLEDGMENTS

Authors want to thank P. Maurizot for its precious help in data preparation. This material is partially based on work under grant of *Coral Reef Initiative in South Pacific*.

7 REFERENCES

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings 20th International Conference on Very Large Data Bases VLDB'94*, pages 487–499. Morgan Kaufmann, 1994.

DITTT. Carte topographique de la nouvelle-calédonie au 1/10000., 1998-2004.

Pascal Dumas. *Caractérisation des littoraux insulaires : approche géographique par télédétection et SIG pour une gestion intégrée. Application en Nouvelle-Calédonie.* PhD thesis, Université d'Orléans, 2004.

Usama M. Fayyad and K. B. Irani. Multi-interval discretization of continous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

Gaëlle Luneau. La spatialisation de l'aléa érosion en nouvelle-calédonie : méthodologie définie sur les communes de dumbéa, païta et bouloupari. Master's thesis, ENSAT – SIGMA, 2006.

Pierre Maurizot and Isabelle Rouet. Aide à la mise au point d'une méthodologie de traitement de la couverture territoriale spot 5 pour le calcul des superficies dégradées par lancienne activité minière en nouvelle-calédonie. Technical Report RP 54 787, BRGM, 2006.

Pierre Maurizot, C. Schmitt, and Myriam Vendé-Leclerc. Harmonisation de la couverture cartographique géologique numérique de la nouvelle-calédonie. Technical Report RP 54 117, BRGM, 2005.

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

Isabelle Rouet and Philipson Bani. Participation l'inventaire des sites dégradés par l'activité minière par traitement semi-automatique d'images spot 5: validation méthodologique des traitements par télédétection. Technical report, EA 3325, Université de la Nouvelle-Calédonie, 2006.

Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques (2nd edition).* Morgan Kaufmann Publishers Inc., San Francisco, USA, 2005.