Mining simulation data by rule induction to determine critical source areas of stream water pollution by herbicides

Ronan Trépos, Véronique Masson, Marie-Odile Cordier, Chantal Gascuel, Jordy Salmon-Monviola

Computers and Electronics in Agriculture 86 (2012) 75–88

Contents lists available at SciVerse ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Mining simulation data by rule induction to determine critical source areas of stream water pollution by herbicides

Ronan Trépos [a,b,c], Véronique Masson [b], Marie-Odile Cordier [b], Chantal Gascuel-Odoux [c,d,*], Jordy Salmon-Monviola [c,d]

[a] INRA, Unité BIA, BP 27, 31326 Castanet-Tolosan Cedex, France
[b] Université de Rennes 1, UMR 6074, IRISA, F-35000 Rennes, France
[c] INRA, UMR 1069, Sol Agro et hydrosystème Spatialisation, F-35000 Rennes, France
[d] Agrocampus Ouest, UMR 1069, Sol Agro et hydrosystème Spatialisation, F-35000 Rennes, France

ARTICLE INFO

Article history:
Received 14 January 2011
Received in revised form 16 December 2011
Accepted 8 January 2012

Keywords:
Symbolic learning
Decision support
Tree structure
Modelling
Catchment
Pesticide

ABSTRACT

Modelling in environmental sciences is becoming increasingly complex because ever-increasing numbers of processes are combined, thus making model-based decision aids both more relevant but more difficult to develop. Our approach focused on water quality and aimed to identify spatial tree patterns that represent surface flow and pollutant pathways from plot to plot involved in water pollution by herbicides. First, a simulation model predicted herbicide transfer rate, the proportion of applied herbicide that reaches water courses, based on the spatial and temporal distribution of weed-control activities. These predictions were used as a set of learning examples for symbolic learning techniques to induce rules based on qualitative and quantitative attributes and explain two classes of transfer rate. In this study we compared two automatic symbolic learning techniques applied to a set of simulations: (1) a relational-inductive method using the inductive logic programming (ILP) approach to induce spatial tree patterns; and (2) an attribute-value method to induce aggregated attributes of the trees. Twenty-eight and thirty-three rules were learnt by the ILP and attribute-value approaches which explained 81% and 88% of the examples, respectively. The ILP approach provided relevant indicators of plot-to-plot connectivity. The integrated attribute-value approach is simpler and quicker, but the ILP patterns are more useful for stakeholders.

1. Introduction

Modelling in environmental sciences is useful for predicting the effect of human activities or climate changes on anthropogenic environmental systems, motivating intense research activities in this field (Durand et al., 2002). Models are becoming increasingly complex because they simulate an increasing number of interacting processes that are distributed over time and space. Consequently, this allows investigation of more scenarios combining constraints related to climate, physical environment or human activities and decisions. Modelling faces additional difficulties, however, when simulating large numbers of scenarios, in which the qualitative responses required for management, such as ranking risks according to various situations, cannot be assessed easily. Moreover, models often are used to predict effects of scenarios; however, a reverse approach which identifies temporal drivers or spatial origins of pollution would be useful in environmental management. This is particularly a problem at the mesoscale (from a few km^2 to a few tenths of km^2), where environments and activities are highly heterogeneous and regulated by larger-scale processes, but decisions to improve landscape quality and human management activities require impact assessment based on mesoscale environmental and socio-economic characteristics.

The use of data mining and other artificial-intelligence techniques has increased along with the recognition of their potential to analyse environmental data (see Chen et al., 2008; Gibert et al., 2008, for a review). When analysing environmental data, many systems propose statistical data analysis or filtering (Kawano et al., 2005; Painter et al., 2006) and visualisation tools to provide insight (e.g., Chertov et al., 2005), while decision support tasks often rely on rules employed as assessment tools or link representations between causes and effects (Brown, 2006; Ivanovska et al., 2008; Janssen et al., 2005; Poch et al., 2004; Tan, 2005).

Automatic symbolic learning provides tools to acquire these general rules from data (e.g., Gibert et al., 2008). So far, these tools mainly have been applied to observed data. This is a critical limitation because of the difficulty and cost of collecting large sets of

* Corresponding author. Address: INRA, UMR 1069, Sol Agro et hydrosystème Spatialisation, 65 Route de Saint Brieuc, CS 84215, 35042 Rennes Cedex, France. Tel.: +33 2 23 48 52 27; fax: +33 2 23 48 54 30.
E-mail address: chantal.gascuel@rennes.inra.fr (C. Gascuel-Odoux).

0168-1699/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.compag.2012.01.006

environmental observations. Our approach uses symbolic learning techniques on simulated data to synthesise complex processes and help in decision making.

Relational decision tree induction has been used to analyse the results of simulation models (Ivanovska et al., 2008) and to obtain decision rules by generalising data (Kohavi and John, 1997; Gibert et al., 2006; Poch et al., 2004). Decision trees predict the value of a dependent variable, called a class, from the values of independent variables, called attributes, by partitioning the space of attributes. The class of a data example is predicted by one branch of a decision tree, which represents a rule explaining the example's membership in the class according to its attributes. Our goal was to develop a decision-support system that can provide a variety of explanations by identifying possibly more than one rule to explain the class of an example.

Because decision tree induction does not suit this purpose, we applied and compared two symbolic learning techniques to detect spatial rules. We conducted tests to analyse their ability to mine spatial rules in a spatio-temporal process, which is a general issue in environmental systems. We focused on spatial rules that (1) describe flow and pollutant pathways in the form of tree patterns and (2) identify areas at risk for stream water pollution. Two automatic symbolic learning techniques were applied to a set of simulations from which possibly redundant classification rules (called characterisation rules in Zhang et al. (2002)) were induced. The two techniques were (1) a relational inductive method, using the inductive logic programming (ILP) approach (Muggleton and De Raedt, 1994) to induce spatial tree patterns, and (2) an attribute-value method (Clark and Boswell, 1991) to induce aggregated attributes globally describing trees. A few studies have been performed on mining tree or graph structures. For instance, in computational biology, some studies have examined the problem of finding relevant substructures in chemical components (Inokuchi et al., 2000; Dehaspe et al., 1998). However, these methods do not handle general constraints on attributes. Learning relational rules as proposed here by ILP is a relatively complex and highly computationally intensive task, which explains our testing another approach, one based on learning attribute-value rules.

The environmental issue we focus on is the effect of agricultural activities on water quality at the landscape scale, specifically, the identification of critical source areas which contribute to stream water pollution by herbicides applied on maize cultivated fields. This issue is unsolved for three fundamental reasons: (1) paucity of observed data, partly due to the high cost of data acquisition both for sampling and analysis; (2) high diversity of chemicals and conditions for their application; and (3) large spatial and temporal variation in water pollution. These variations are related mainly to the spatio-temporal distribution of treatments within the catchment and characteristics of the field and stream margins (Colin et al., 2000; Leu et al., 2004a,b; Louchart et al., 2001); however, these relationships cannot be analysed easily from observation data due to a low ratio of input–output mass (on the order of a few percentage, Voltz et al., 2003; Clement et al., 1999) because of pesticide retention in the soil. The observation dataset is generally not large and various enough to aid in identifying critical source areas. At a large scale such as Europe, different meta-modelling approaches have been developed recently to improve pesticide-risk assessment. All are based on one-dimensional stochastic modelling (Tiktak et al., 2006; Auteri et al., 2007; Stenemo et al., 2007) constrained by pedo-transfer functions to restrict the range of variations in parameter values (Centofantia et al., 2008). But at the landscape scale, ephemeral and highly variable lateral flow and retention processes occur along hillslopes and influence water pollution dynamics (Dubus et al., 2002, 2003). Consequently, questions regarding critical treatment periods and the location of critical source areas which contribute to stream water pollution remain a scientific and management challenge.

This is a relevant application for comparing the two rule-induction techniques and analysing their ability to help identify critical source areas of stream water pollution by herbicides. Scenarios were simulated using the SACADEAU model, which represents farmers' decision processes (Salmon-Monviola et al., 2011) and transfer processes of water and chemicals within a catchment (Gascuel-Odoux et al., 2009). The catchment is represented by means of a nodal network, often used in water resource allocation and management models (Letcher et al., 2007), that consists of a set of elementary plot outlet trees (Fig. 1), each plot outlet tree describing flow pathways from plot to plot that finally lead to the stream (Aurousseau et al., 2009). The two automatic learning techniques then were used to identify patterns in the form of plot outlet trees (with ILP) and global characteristics of plot outlet trees (with the attribute-value approach).

The objective of this paper is to investigate whether automatic learning techniques on simulated data can help identifying spatial rules: here, these are the critical source areas involved in stream water pollution. First, we describe the simulator and the procedures used in generating and analysing the simulations. Next, we present and compare the results obtained from the two automatic learning tools and the way in which they have been applied to the case study, and describe the graphical interface tool used to analyse the results in a real decision making process.

2. Methods

Fig. 2 presents data flows from the simulator to the visualisation tool in a decision support system. Three groups of tools were involved: a model to generate a set of learning examples, symbolic techniques to extract rules from simulation data, and decision-oriented visualisation tools to identify the learnt rules involved in any given situation of the studied catchment (Fig. 3).

2.1. Generate a set of learning examples with a simulator

2.1.1. The simulator

The SACADEAU model used as the simulator predicts stream water pollution (Gascuel-Odoux et al., 2009) with three sub-models that represent farmers' decision processes (SACADEAU-Deci, Salmon-Monviola et al., 2011), crop production and transfer processes of water and chemicals within a catchment (SACADEAU-transf, Gascuel-Odoux et al., 2009) (Fig. 3). The models' input is several series of weather data, as well as technical and environmental data such as a digital elevation model (DEM) (O'callaghan and Mark, 1984; Fairfield and Leymarie, 1991) and a plot map of the catchment, as well land uses, farm boundaries and strategies used for weed control within the catchment.

To predict the date, location and amount of herbicide applied within a catchment, the decision sub-model represents farmers' weed-control activities, including environmental and technical constraints at plot and catchment scales. Concretely, it spatially and temporally distributes herbicides on maize crops within the catchment during the spring by means of decision rules. A technical operation is performed when technical and environmental conditions are satisfied that relate to soil workability (climatic conditions, slope and position of the plot) and application efficiency (machine availability, farmers' working time). The spatial schema takes into account constraints at the plot, farm, farm-group and catchment levels. The outputs of the decision sub-model used as inputs for the transfer sub-model are sowing and weeding dates and herbicide amounts applied to maize plots in the catchment.

The transfer sub-model represents water and herbicide transfer by surface and subsurface flow up to the catchment outlet (Gascuel-Odoux et al., 2009). The transfer sub-model simulates the retention, degradation and transfer processes of herbicides at

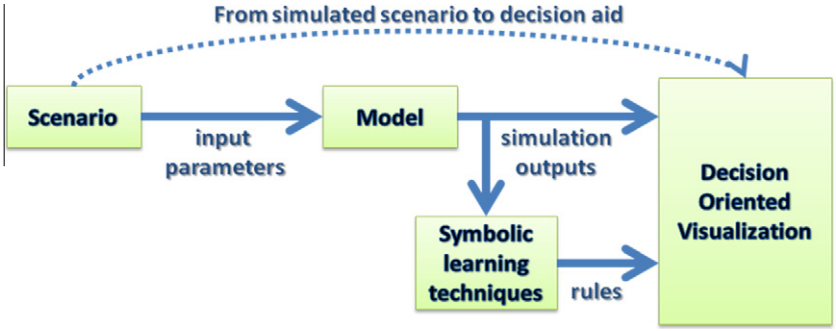

Fig. 1. A plot outlet tree is a sub-catchment where each vertex ("outlet") is part of a plot. Attributes given for illustration are determined for each outlet.

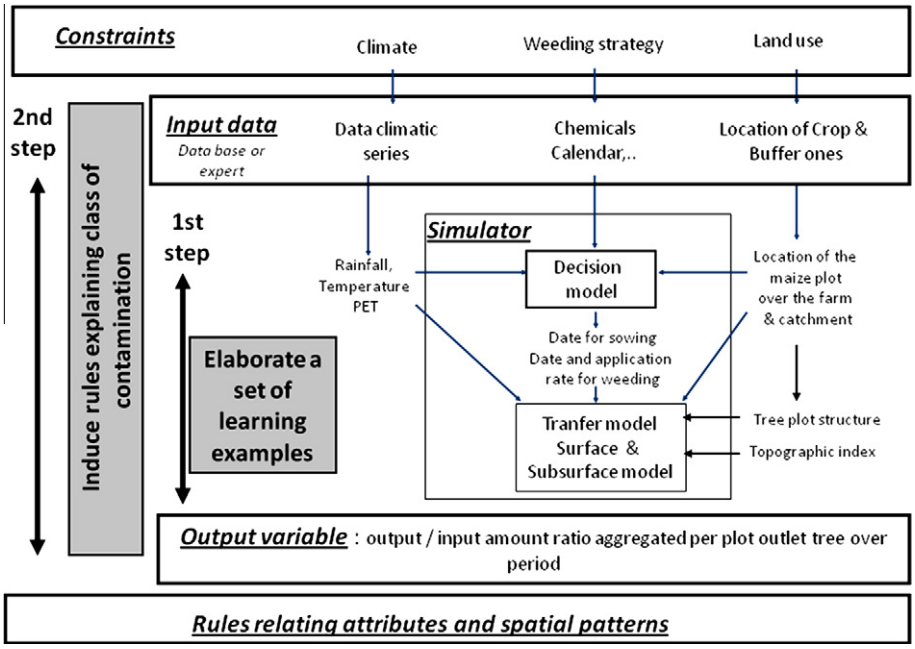

Fig. 2. Integration of a simulation model into a decision support system.

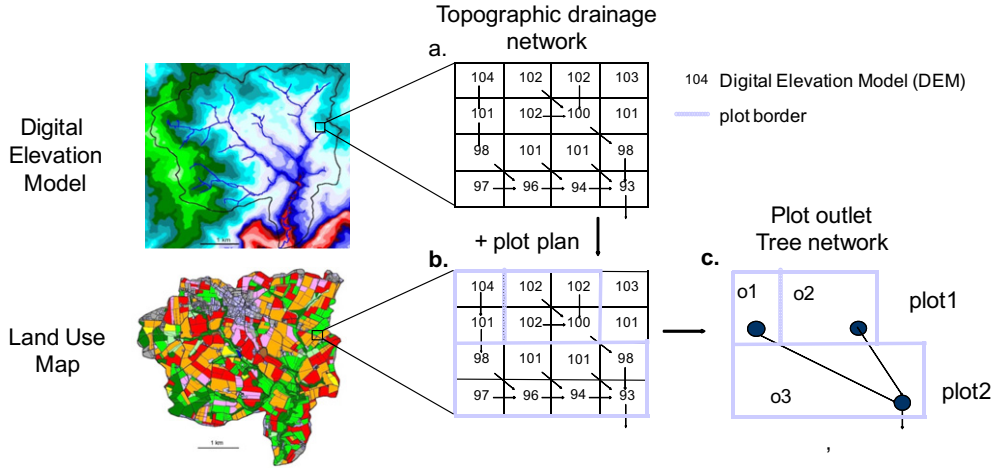

Fig. 3. SACADEAU framework, including two steps: (1) generate a set of learning examples using a simulator, and (2) induce rules to explain classes of pollution.

the plot and catchment scales. This sub-model combines other sub-models that consist of expert rules and mechanistic models: (1) the model TOPMODEL (Beven and Kirkby, 1979) to predict discharge and water-table depth; (2) simple functions or models, such as a crop model, an herbicide retention and degradation function, and an herbicide exchange function between soil and water; and (3) expert rules to predict soil surface conditions and soil infiltrability related to crop development and weather conditions, and

therefore herbicide concentrations in surface runoff (Cerdan et al., 2001) and groundwater recharge. The spatial schema of this model is original (Aurousseau et al., 2009): surface flow connectivity from plot to plot within the catchment is represented by a drainage network overlaid on the land-use map (Fig. 4a and b), so that a plot outlet is defined as the set of cells that contribute to a specific outlet of the plot (plot outlets o1 and o2, Fig. 4c). Buffer areas such as hedgerows or grass strips can redirect or interrupt surface flow pathways. The catchment's drainage network for surface flow routing is represented by a set of plot outlet trees, each of which feeds the stream independently.

Finally, daily fluxes in herbicides during the cropping period are aggregated in time and space over all plot outlets to obtain a prediction of the proportion of input herbicide mass transferred to the stream, called the transfer rate and defined as follows:

$$transfer\_rate(t) = \frac{\sum_{w \in W} \frac{transferred(w,t)}{applied(w,t)}}{|W|} \quad (1)$$

where t is a plot outlet tree, and W is a set of real weather data series. The function *transferred* (w,t) stands for the total amount of herbicides (in grams) transferred from plot outlet tree t to the stream water for weather data series w. The function *applied*(w,t) stands for the total amount of herbicides applied to plot outlet tree t for weather data series w. To compensate for the effect of variability in the weather data series, the output variable is averaged over a set of real weather years (W). This variable is computed for each plot outlet tree that feeds the stream water and defines its buffer capacity (i.e., the resistance of the landscape to pollutant transfer; Viaud et al., 2004). Therefore, it is well-suited to test the effect of landscape structure and agricultural practices on stream water pollution. The interest of this output variable has been discussed in the previous paper which describes the transfer model (Gascuel-odoux et al., 2009).

2.1.2. Building the set of simulations

Usually, when rules are induced from observations, there is no problem in selecting learning examples since all observations are used. But observed data do not necessarily represent all possible situations, and studies on the induction of classification rules from simulation data often address the problem of generating representative learning examples (Mladenic et al., 1994; Huber and Berthold, 1997; Zhang et al., 2002). If model inputs are discrete and few in number, then all combinations of input values can be simulated and become learning examples (Mladenic et al., 1994). On the contrary, if inputs are continuous or drawn from infinite domains, then some inputs have to be considered constants while others vary within their domains. One approach to deal with such inputs is to attribute random values to the most relevant variables to analyse their effects on outputs, while other variables are fixed (Huber and Bertold, 1997; Zhang et al., 2002). We did not rely on classical techniques for designing computer experiments (e.g., Fang et al., 2005), since they are not adapted to manage relational input data such as plot outlet trees.

To carry out the exploration of the model, we focused on scenarios based on a given study site (described later). Constraints on the simulation inputs are defined so that some observed data on this study site are preserved, mainly the landscape and farm structure: (1) topography, plot map and number of farmers; (2) total maize crop area; (3) inter-farm distances of maize crops; and (4) weeding strategies for maize crops.

To generate one simulation input configuration that fulfils these constraints, we applied the following strategy:

(1) The DEM of the study site is used (see Section 3).
(2) A land use map is generated so that the total maize crop area is preserved. The constraint programming system CHOCO (Laburthe, 2000) is used to allocate the maize plots amongst the plots of the catchment.
(3) Maize plots are clustered with "k-means" algorithm (MacQueen, 1967) to reflect observed inter-farm distances. As a result, a cluster of maize plots represents a farm. The parameter k is the number of farms and is set to 36 which is the observed number of farms on the study site.
(4) From observed data, we conclude that each farmer applies at most two weeding strategies. These strategies are randomly chosen amongst 50 weeding strategies previously collected on the study site.

Once the land-use map is built the site's DEM was overlaid to extract all plot outlet trees that contained at least one plot outlet with a maize crop (Fig. 4). A set of simulations was built according to these constraints and then run to predict transfer rates for these plot outlet trees. Finally, each plot outlet tree was labelled by a transfer rate and sorted at into different classes according to the value of the transfer rate.

2.2. Induce rules using symbolic learning

2.2.1. Criteria for selecting rules

The set of simulated and labelled plot outlet trees constituted the learning examples used to automatically induce classification rules of the following form:

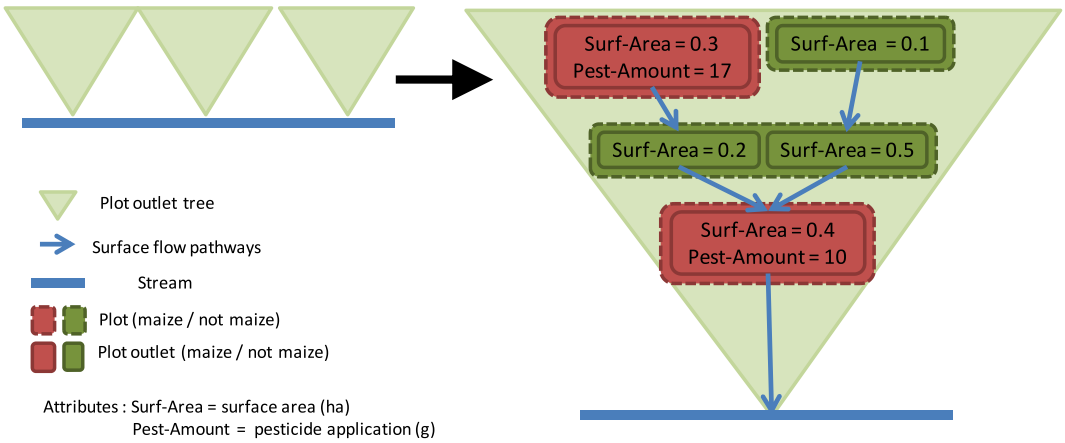

Fig. 4. Overland flow pathways: plot outlets (here o1, o2 and o3) are determined in three steps (a, b and c) from a digital elevation map and a plot map of the catchment.

"If a plot outlet tree t satisfies conditions C, then the class of t is L" (2)

where L is a label standing for a transfer-rate class. A plot outlet tree t is "covered" by a rule R if t satisfies the conditions C, and the set of plot outlet trees covered by rule R is $covers(R)$.

In rule induction, syntactic bias refers to the language used to express the conditional part of the rules, which defines the search space (i.e., the set of "admissible" rules). An induction process explores the search space to discover the best rules. To compare rules, the user also should provide a rule quality criterion, called semantic bias, which measures how well a rule fits the learning examples. Let us define:

(1) n^L: The number of examples of class L.
(2) $n^{\bar{L}}$: The number of examples of class different from L.
(3) n^L_R: The number of examples of class L covered by R.
(4) $n^{\bar{L}}_R$: The number of examples of class different from L covered by R.

The support of a rule is the proportion of learning examples covered by a rule: $support(R) = (n^L_R + n^{\bar{L}}_R)/(n^L + n^{\bar{L}})$. "Rule accuracy" is the ratio of examples of class L to all examples covered by rule R : $accuracy(R, L) = n^L_R/(n^L_R + n^{\bar{L}}_R)$. The closer accuracy is to 1, the safer a classification becomes when based only on R. It is recognised, however, that a measure of rule quality should be a trade-off between accuracy and generalisation. Allowing a few misclassifications is acceptable for a rule if it covers a large subset of learning examples. Moreover, a rule also should take into account the *a priori* distributions of labels amongst examples. Thus, we chose to use the m-estimate measure (Cestnik, 1990), which is used widely in rule induction studies, as a rule-quality criterion parameterised by m:

$$m\text{-}estimate(R, L) = \frac{n^L_R + m \times n^L/(n^L + n^{\bar{L}})}{n^L_R + n^{\bar{L}}_R + m} \quad (3)$$

A rule R that causes misclassifications ($n^{\bar{L}}_R > 0$) is preferable to a rule that causes no misclassification if it covers "enough" examples of class L (n^L_R). Note that if $m = 0$, then the *m-estimate* equals the accuracy.

Moreover, as it is computationally too expensive to browse the entire search space, heuristics were developed to explore it. First, a relation of generality between rules induced a lattice structure on the search space to facilitate the search. If rule R is more general than rule R', then $covers(R') \subseteq covers(R)$. To browse the search space, a "general to specific" strategy was then adopted which began by evaluating a general rule R and then iteratively building and evaluating the rules R' that are more specific than R. A specialisation operator built more specific rules R' from rule R, taking into account the syntactic bias to build only admissible rules. To limit search space exploration, a "beam search" strategy (e.g., Clark and Boswell, 1991) was used at each step to focus on the few best rules according to the *m-estimate* value. Finally, the best rule generated during the process was returned and the process iterated until all examples were covered by at least one rule or until no more admissible rules could be built. The minimal thresholds for the *m-estimate* value and the number of examples covered by a rule were given as parameters by the user.

A key point was to define well-adapted syntactic biases which expressed admissible rules for the application. Two syntactic biases were proposed and experimented with to describe the plot outlet trees (Fig. 1). In the first approach, the conditional part of the rules directly referred to the tree structure of the plot outlets and thus, tree-structured rules were learnt (the ILP mining method). In the second approach, the conditional part of the rules referred to a synthetic description of the outlet trees, through a set of aggregate attributes, and thus, attribute-value rules were learnt (the attribute-value rule mining method).

2.2.2. ILP mining method

ILP (Muggleton and De Raedt, 1994) is well-suited to induce rules from data with relational properties such as spatial ones. The ILP method learnt rules with tree-structured conditions (Eq. (2)) from plot outlet trees that constituted the set of examples, called learning examples. We computed tree-structured patterns, the conditional parts of the rules to be induced, that generalised plot outlet trees.

Given a tree, many patterns, or sub-trees, may be inferred (Chi et al., 2005) so that induction must take the application into account. We remind that the plot outlet trees describe the flow pathways from plot to plot that leads to the stream. As we regard plot outlets closer to the stream as more important for herbicide transfer than plot outlets at the top of the catchment, the root of a plot outlet tree represents the plot outlet feeding the stream. ILP computed a pattern that generalised examples by taking the root of a plot outlet tree as the root of the pattern. Then, repeatedly, either an outlet was added to the pattern (from the stream up to the catchment) or a constraint on the attribute value of outlets was introduced. At every stage, computed patterns were checked on learning examples (plot outlet trees).

As mentioned, ILP specialisation operators explored the search space with a "general to specific" strategy. These specialisation operators either added an outlet (specialisation operator 1) to the inferred pattern or added an attribute-value pair to the last outlet added (specialisation operator 2). The order in which outlets were added is meaningless regarding the run-off process and thus the induction of unordered sub-trees was adapted. The ILP system used was Aleph (Srinivasan, 2003), which inferred unordered rooted sub-trees with these two specialisation operators.

Thus, from a sub-tree pattern P, a set of more specific sub-tree patterns P' can be generated. For example, four sub-tree patterns are generated from a pattern g with two outlets: s1 and s2 are specialised by adding a third outlet, and s3 and s4 by introducing a constraint on the last outlet added (Fig. 5).

When exploring the search space the main difficulty is avoiding generating the same pattern twice, even if it can be reached by different paths in the lattice structure of the search space. This redundancy problem is a well-known issue in ILP that can involve extra computational costs. When generating plot outlet tree patterns, the redundancy problem mainly arises from the trees' lack of ordering. To solve this problem, a specialisation step that adds an outlet to the pattern (specialisation operator 1) numbers the unordered trees (Nakano and Uno, 2003). A naïve strategy for this specialisation step could be achieved by adding a vertice child to a randomly chosen vertice into the current tree. On contrary, the algorithm proposed by Nakano and Uno selects, by studying the tree structure, the vertices into the current tree for which one can add a child, in order to ensure that only unordered rooted trees are produced. For comparison, the naïve strategy produces 40,320 trees of size 9, while the Nakano and Uno algorithm produces 286 trees of size 9, which is the total number of unordered trees of size 9.

Moreover, when adding an attribute-value pair to the last added outlet (specialisation operator 2), the value is determined (Srinivasan and Camacho, 1999), using entropy measures to find a value relevant to a given class. The examples are projected onto the axis of real values taken by the attribute under consideration, and the value that best separates the different classes is chosen.

2.2.3. Attribute-value rule mining method

The attribute-value rule mining method selects a set of aggregate attributes that describe structured data, in our case, plot

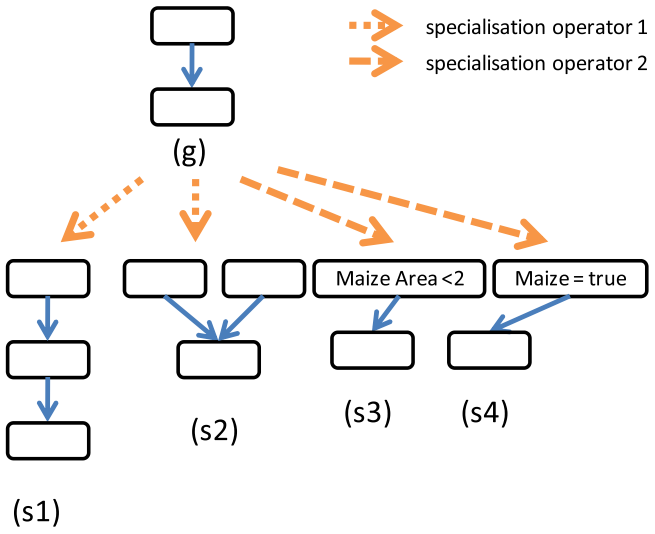

Fig. 5. Specialisation of a plot outlet tree pattern (g) with two specialisation operators. Tree patterns (s1), (s2), (s3) and (s4) are more specific than (g).

outlet trees. Given such a set of attributes $\{a_1, \ldots, a_n\}$, an attribute-value rule is written as:

$$\text{“}If\ (t \cdot a_{k_1} \#_1 v_1) \wedge \cdots \wedge (t.a_{k_r} \#_r v_r)\ then\ class(t) = L\text{”} \quad (4)$$

where t is a plot outlet tree, L is a class label, $\{k_1, \ldots, k_r\} \subseteq \{1, \ldots, n\}$, $t \cdot a_{k_i}$ is the value of the attribute a_{k_i} in t, v_i is a value of the domain of a_{k_i} and $\#_i$ is a relational operator among $\leqslant, =, \geqslant$. An attribute-value rule R covers t if each attribute-value constraint $t \cdot a_{k_i} \#_i v_i$ is satisfied by t.

Many efficient methods and systems have been devised for mining attribute-value rules. The search space is explored in a “general to specific” way by applying the classical specialisation operator that adds an attribute-value pair to the conditional part of the rules. When faced with tree-structured data in our study, each attribute aggregated the values of attributes on the tree nodes (i.e., the plot outlet tree nodes). The attribute-value description of a tree was then related to its root (i.e., the outlet directly connected to the stream). The main difficulty with this approach is building a set of relevant attributes which depends on the application. Selecting these attributes is especially difficult when mining relational data, as they must synthesise a complex data structure sufficiently and simply.

2.2.4. Selecting and classifying attributes

Two sets of attributes were selected to identify the relevant sub-trees in the ILP approach and the attributes per tree in the attribute-value approach. For the two approaches, the attributes were chosen to represent the entire set of factors that explained the transfer rate. In the ILP approach, the attributes depicted each plot outlet of a tree. They were selected among variables used in the model to describe outlets. In the attribute-value approach, the attributes depicted a plot outlet tree and thus aggregated the attribute on tree nodes; therefore, the attributes are not directly used in the model. They are defined and selected informally and iteratively. The final set of attributes is one that yielded the most relevant rules.

The attributes are classified as Boolean, qualitative or quantitative. Searching for spatial patterns in the form of plot outlet trees is really a qualitative way to express spatial patterns. These patterns constitute a reality that the stakeholders can observe in the field. Consequently, such spatial patterns, can be considered to be site- or stakeholder-dependent. The attributes are also classified according their physical significance. We distinguished source factors from transfer factors, which is a common distinction in risk assessment (Heathwaite et al., 2000, 2005). Some attributes can be considered source factors: for example, related to pollutant pressure, which can be direct (e.g., herbicide amount) or an indirect (e.g., surface area of maize). Other attributes can be considered transfer factors: for example, related to the ability of the landscape to regulate a pollutant pressure, which can be due directly to the presence of a buffer area or indirectly due to slope gradient or topographic position.

2.3. Analysis and comparison of approaches

Approaches were analysed and compared according to their percentages of correct and incorrect classifications (in a confusion matrix) and computation time. An *a posteriori* classifier used the induced rules to predict the class of testing examples, which were not used during the induction process (Clark and Boswell, 1991; Van Laer, 2002). It assigned a class to an example when rules conflicted by tagging each rule with the distribution of covered learning examples among classes and then summing these distributions to find the most probable class to assign to the testing example.

Correlations between attributes were computed to identify and analyse their interactions in the rules. Given two attributes, a and b, we computed the correlation of occurrences in the induced rules. More precisely, let E_a be the event that attribute a occurs in a rule and $\overline{E_a}$ be the event that attribute a does not occur in a rule. We relied on estimating the probabilities that attributes $P(a)$ and $P(b)$ occurred, or did not occur in rules, and on the joint probabilities that they occurred together $P(a,b)$. We were interested in “mutual information” about the occurrences of a pair (a,b) of attributes in rules (Jakulin and Bratko, 2004):

$$I(a,b) = \sum_{e_1 \in \{E_a, \overline{E_a}\}} \sum_{e_2 \in \{E_b, \overline{E_b}\}} P(e_1, e_2) \log_2 \left(\frac{P(e_1, e_2)}{P(e_1) \times P(e_2)} \right) \quad (5)$$

If $P(e_1) = 0$ or $P(e_2) = 0$ then $1/(P(e_1) \times P(e_2))$ is set to 0. By construction, $I(a,b) \geqslant 0$. When $I(a,b) = 0$, then the occurrences of a and b in rules are independent. The larger $I(a,b)$, the more they are correlated. Actually, for tree patterns, we did not compute the correlation of occurrences of two attributes in a rule but rather in an outlet of a pattern: two attributes on different outlets do not refer to the same object.

3. Application

3.1. Study site

The study site chosen to test these methods was the Frémeur catchment (14 km^2 surface area), located in western France. The DEM was extracted from an elevation database with a 20-m grid size produced by stereo plotting of panchromatic SPOT images to a resolution of 10 m. The cultivated area represented 72% of the catchment area and the maize crops 21%. The catchment contained 1419 plots, from which 416 plots were used for maize, temporary grassland or cereals. These 416 plots were considered usable only for maize in simulations. The catchment contained 5312 plot outlets. A plot outlet tree had a mean of 7.6 plot outlets, and the catchment contained 692 plot outlet trees. A detailed survey taken in 2005 identified land cover and weeding strategies within the catchment (Tortrat, 2005). Two types of weeding strategies were observed on maize plots: (1) pre-post-emergence, with one application after sowing and another at the 5-leaf stage; and (2) post-emergence, with one application at the 3-leaf stage and another at the 5- to 7-leaf stage. Altogether, about 50 different weeding

programs and a large number of herbicides are inventoried on maize due to numerous commercial channels even at local scale.

3.2. Simulation outputs

Based on the programming and clustering techniques previously described, we generated 20 land-use maps and their associated farm layout. By overlaying these maps on the single DEM of the study site, we extracted 3431 plot outlet trees, each containing at least one plot outlet with maize crops. We can consider that these 3431 plot outlet trees cover a high diversity of spatial patterns, and therefore, is representative to deal with the identification of herbicide source areas. A mean transfer rate (Eq. (1)) was predicted for each plot outlet tree based on observed nine weather years, considered representative of weather variability, as previously studied (Salmon-Monviola et al., 2011). The set of simulated plot outlet trees was divided into three roughly equal-sized transfer classes according to the value of the transfer rate (Fig. 6). We have discarded the intermediate class while our objective was to identify rules differentiating the two extreme classes (high- and low-transfer), considered as "acceptable" and "unacceptable", respectively (Fig. 6). Firstly, from an environmental point of view, this is relevant to identify no or poor risk situations (low transfer) from high risk situations (high transfer). Secondly, we claim that errors when labelling intermediate values (as high- or low-transfer) would occur more probably than errors when labelling extreme values.

The experiments used the same set of examples: 557 and 561 examples were used as learning examples for the high and low-transfer class, respectively, while 587 and 582 examples were used as testing examples for the high and the low-transfer class, respectively.

During ILP search, the size of the beam was fixed at 50, rules were discarded if they covered fewer than 15 learning examples or if their accuracy was less than 0.6, and plot outlet sub-trees were discarded if they contained more than nine outlets. The plot outlet sub-tree mining method was implemented in the Aleph ILP system (Srinivasan, 2003), while the CN2 system (Clark and Boswell, 1991) induced attribute-value rules. Experiments were run on an Intel® Xeon 3.4 GHz processor.

4. Results

ILP patterns and attribute-value rules learning methods are applied on a dataset composed of 1143 learning examples: 561 of low transfer and 557 of high transfer (Fig. 6). Classification performance is computed on a test set of 1144 test examples: 582 of low transfer and 587 of high transfer (Table 5).

4.1. Selected attributes

Seven attributes were selected for describing a plot outlet in the ILP approach (Table 1) amongst around 10 attributes. For example, the Top-Index attribute is a function of the local slope of the plot outlet and the drained surface (Beven and Kirkby, 1979); replacing these two last attributes by the Top-Index leads to better classification results for the learning process. Ten attributes for the attribute-value approach (Table 2) amongst around 30 were selected. In this approach, considered attributes are naturally more numerous since one can consider different functions that synthesize data (count, sum, mean, etc.). Quantitative variables were less numerous in the ILP approach (3 of 7) than in the attribute-value approach (8 of 10). In the attribute-value approach, synthesized data relies, by nature, on quantitative attributes.

4.2. ILP patterns

ILP learning produced 28 induced rules (14 high-transfer class and 14 low-transfer class; Table 3). These 28 rules correctly classified 88% of the testing examples. Six main plot outlet sub-tree patterns were learnt (Fig. 7). One high-transfer-rate sub-tree pattern described a situation in which at least two plot outlets are connected to the root outlet, and thus, adjacent to the stream (Fig. 7a). On one, a large quantity of herbicide was applied (at least 17 g for the whole outlet). The other was a maize-cultivated outlet itself connected by at least two outlets, of which one was a surface area greater than 0.32 ha. This sub-tree pattern covered 19% of plot outlet trees belonging to the high-transfer class. One low-transfer-rate tree pattern described a situation where the outlet was just upstream of the root and had a high slope and a low topographic index (less than 3.12 ha; Fig. 7e). This pattern covered plot outlet trees located uphill in the catchment, where the low topographic index indicated a deep water table. This sub-tree pattern covered 8% of plot outlet trees belonging to the low-transfer class.

High-transfer-class sub-tree patterns had a total of 37 attributes (26 quantitative and 11 qualitative) while, those for the low-transfer class had 34 attributes (22 quantitative and 12 qualitative). To analyse the rules, we registered the number of occurrences *nbocc* of attributes in the ILP patterns (Table 3). The most frequently selected attributes in the high-transfer class were amount of herbicide applied in the sub-tree (Herb-Amount, nbocc = 9), surface area of the contribution area to the outlet (Surf-area, nbocc = 9), topographic

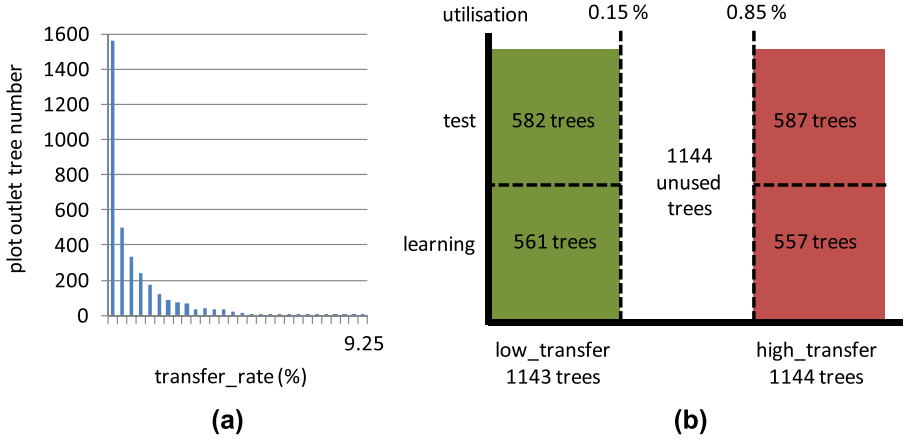

Fig. 6. Simulated outputs: (a) histogram of simulated plot outlet trees according to the transfer rate; (b) distribution of plot outlet trees in the example set.

Table 1
Attributes selected for the inductive logic programming (ILP) approach.

Attribute	Type	Significance	Description
Surf-area	Real	Source	Surface area contributing to the outlet
Maize	Boolean	Source	True if plot is cultivated in maize
Buffer	Boolean	Transfer	True if there is a buffer zone (hedge row or grass strip)
Slope	Qualitative	Transfer	Slope gradient: flat (f), low (l) or high (h)
Top-Index	Real	Transfer	Index that reflects water-table depth (Beven and Kirkby, 1979), and therefore slope position
Pre	Boolean	Source	True if there is a pre-emergence weeding strategy on the plot
Herb-Amount	Real	Source	Herbicide amount transferred from the contributing area to the outlet

Table 2
Aggregate attributes for the attribute-value approach.

Attribute	Type	Significance	Description
Herb-Amount	Real	Source	Total amount of pesticides applied on the tree (in grams)
HighRisk LowRisk	Real	Transfer	Proportion of chemicals at low or high risk, defined by sorption (Koc) and half-life (%)
Pre	Real	Source	Proportion of pre-emergence application on maize (%)
Surf-area	Real	Source	Surface area of the plot outlet (ha)
Maize	Real	Source	Percentage of surface area cultivated in maize over the sub-tree (%)
Max-Maize	Real	Source	Surface area of the largest plot outlet cultivated in maize (ha)
Buffer	Real	Transfer	Ratio of surface used as buffer zone (%)
Tree-Depth	Integer	Transfer	Depth of the plot outlet tree (number of outlets)
Tree-Shape	Qualitative	Transfer	Form of the tree regarding number and position of the plot outlets depth: I (isolate: only one outlet per tree), l (linear: only one outlet at any depth), u (many outlets at depth 1); v (intermediate shape between *i* and *u*)
Tree-Top	Qualitative	Transfer	Spatial distribution of slopes over the tree: s (steep); f (flat); cav (concave: steep at the top of catchment); vex (steep close to the stream)

index (Top-Index, nbocc = 8) and presence of maize (Maize, nbocc = 7). The most frequently selected attributes in the low-transfer class were topographic index (nbocc = 13), slope (Slope, nbocc = 8) and surface area of the contribution area to the outlet (nbocc = 6). Therefore, the surface area and the amount of herbicide mainly explained the high-transfer class, while the topographic index mainly explained the low-transfer class. Conversely, the presence of a buffer area and the type of weeding strategy were not used.

Among the 71 attributes involved in the 28 rules of the ILP approach (Table 3), 25 in the high-transfer class and 11 in the low-transfer class could be considered source factors and 12 in the high-transfer class and 23 in the low-transfer class could be considered transfer factors (Table 1). The most frequent source factors involved in the high-transfer class were the surface area and the amount of herbicide, and the most involved transfer factor in the low-transfer rate was the topographic index, indicating the water table depth, and therefore, the position of the plot on the hill slope.

The positions of the attributes within the tree patterns varied. For the most frequent variables in the high-transfer class, the surface area attribute is located at the sub-tree root (nbocc = 1), at depth 1 (nbocc = 4), at depth 2 (nbocc = 3) at depth 3 (nbocc = 1), and the amount of herbicide was located at the outlet (nbocc = 1), at depth 1 (nbocc = 6) or at depth 3 or 4 (nbocc = 1 each), where tree-depth is the distance to sub-tree root. In the low-transfer class, the topographic index attribute was located at the sub-tree root (nbocc = 1), at the tree depth 1 (nbocc = 8), at depth 2 and depth 3 (nbocc = 2 each). Therefore, the main attributes of the high-transfer class were not preferentially located, while they were at depth 1 for the low-transfer class.

4.3. Attribute-value rules

The learning process resulted in 33 induced attribute-value rules (16 high-transfer class and 17 low-transfer class; Table 4). These 33 induced rules correctly classified 88% of the testing examples. Five of the most important attribute-value rules were the following:

Rule 1. IF Max-Maize > 1.54 ha

AND HighRisk > 50%
THEN class = high_transfer [support = 22% and accuracy = 1]
Rule 2. IF Buffer < 16.5%
AND Maize > 53.5%
AND Pest-Amount > 61.5 g
AND Tree-Top = flat
THEN class = high_transfer [support = 17% and accuracy = 1]
Rule 3. IF 10% < Maize < 39.5%

AND Tree-Shape = I
AND Surf-area > 0.18 ha
THEN class = low_transfer [support = 10.4% and accuracy = 1]
Rule 4. IF 4.5 < Tree-Depth < 7.5

AND 0.62 ha < Surf-area < 1.58 ha
THEN class = low_transfer [support = 14% and accuracy = 1]

Rule 5. IF Pre > 65.5%
AND 35.5% < Maize < 67.5%
AND Tree-Depth < 3.5
THEN class = low_transfer [support = 5% and accuracy = 0.74]

Table 3
Attributes of the 28 rules selected for the inductive logic programming (ILP) approach and their characteristics.

(C,NC)	High-transfer rate
(152,1)	Slope (0) P1, Herb-Amount (lt) P1, Surf-area (gt) P2
(109,0)	Maize (T) P1, Surf-area (gt) P1, Buffer (F) P1
(104,0)	Maize (T) P1, Herb-Amount (gt) P1, Surf-area (gt) P2
(74,0)	Herb-Amount (lt) P0, Surf-area (gt) P0
(70,6)	Maize (t) P2, Herb-Amount (gt) P3
(61,0)	Top-Index (gt) P1, Herb-Amount (gt) P4
(60,2)	Slope (0) P1, Herb-Amount (lt) P1, Top-Index (gt) P1, Surf-area (lt) P1
(57,0)	Top-Index (gt) P1, Herb-Amount (lt) P1
(56,0)	Surf-area (lt) P1, Top-Index (gt) P2, Maize (T) P2
(52,1)	Slope (0) P2, Surf-area (gt) P3, Top-Index (gt) P3
(51,0)	Surf-area (gt) P2, Herb-Amount (gt) P4
(46,0)	Top-Index (gt) P1, Maize (T) P2
(41,0)	Maize (T) P1, Top-Index (gt) P1, Surf-area (lt) P1
(30,0)	Herb-Amount (lt) P1, Top-Index (gt) P1, Maize (F) P2
	Low-transfer rate
(66,5)	Top-Index (lt) P2, Surf-area (gt) P2
(56,11)	Surf-area (lt) P1, Top-Index (lt) P3
(55,5)	Top-Index (lt) P2
(46,0)	Surf-area (lt) P1, Top-Index (lt) P1, Slope (f) P5
(46,0)	Top-Index (lt) P1, Slope (f) P1, Maize (t) P2
(34,0)	Top-Index (lt) P3, Surf-area (gt) P4
(33,0)	Top-Index (lt) P1, Surf-area (gt) P1, Slope (f) P2
(32,8)	Top-Index (lt) P1, Herb-Amount (lt) P1
(24,0)	Top-Index (lt) P1, Slope (f) P1, Surf-area (gt) P1, Maize (F) P1, Buffer (F) P2
(23,0)	Top-Index (lt) P1, Surf-area (gt) P2
(19,0)	Herb-Amount (lt) P1, Top-Index (lt) P1
(18,3)	Top-Index (lt) P1, Herb-Amount (gt) P1, Slope (f) P1
(17,2)	Top-Index (lt) P0, Slope (0) P0
(15,0)	Buffer (T) P1, Slope (0) P1

(C,NC): (number of covered examples of the same class, number of covered examples of the other class); lt: lower than; gt: greater than; F: false; T: true; 0: flat. For gt and lt, the values of the thresholds are not indicated. P#: depth of the plot within the corresponding sub-tree (0 is the root outlet).

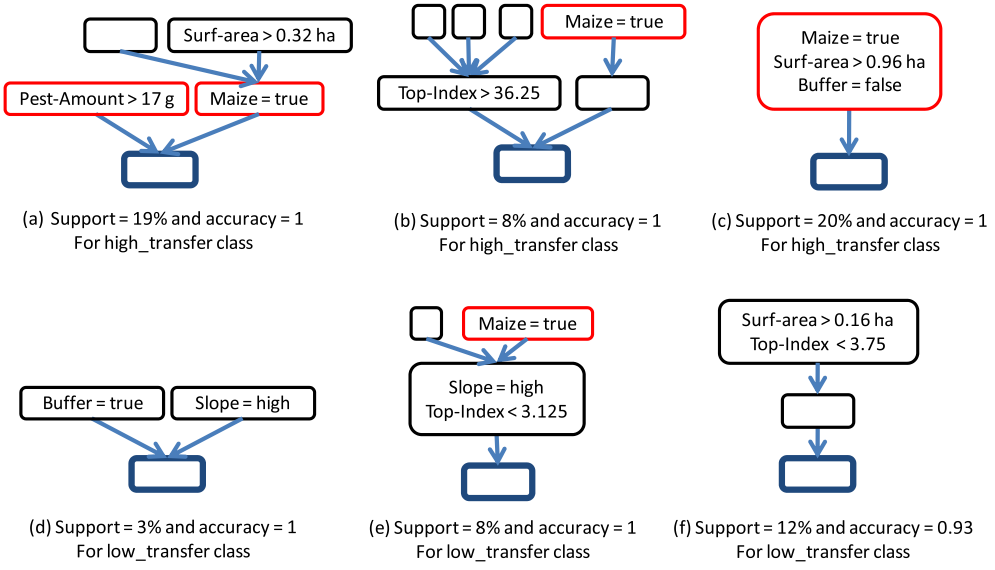

Fig. 7. Six induced tree structure patterns.

Rule 1, belonging to the high-transfer class, covered plot outlet trees with a large amount of area in maize (more than 1.54 ha) on which at least 50% of the applied quantity were high risk herbicides. Rule 3, belonging to the low-transfer class, covered plot outlet trees with a moderately-sized maize cultivated area (from 10% to 39.5% of sub-tree surface area), a sub-tree surface area higher than 0.18 ha, and a linear shape (I = only one outlet at any depth).

High-transfer rules had a total of 60 attributes (52 quantitative and 8 qualitative), while low-transfer rules had 70 attributes (62 quantitative and 8 qualitative). Similarly than previously, we noticed the occurrence *nbocc* of a given attribute (Table 4). The most frequently selected attributes in the high-transfer class were surface area of a sub-tree (Surf-area, nbocc = 13), surface area with maize in a sub-tree (Maize, nbocc = 10) and surface area of the

Table 4
Attributes of the 34 rules selected for the attribute-value approach and their characteristics.

(C,NC)	High transfer rate
(125,0)	HighRisk (gt), Max-Maize (gt)
(105,0)	Herb-Amount(gt), Surf-area (gt), Maize(gt), Max-Maize (gt), Buffer (lt), Tree-Depth (lt)
(92,0)	Herb-Amount (gt), Maize (gt), Buffer (lt), Tree-Top (f)
(91,1)	Surf-area (gt), Maize (gt), Max-Maize (gt)
(49,0)	Surf-area (lt), Maize (lt), Max-Maize (gt)
(44,0)	Surf-area (lt), Max-Maize (gt), Buffer (lt)
(39,0)	Surf-area (bet), Max-Maize (lt), Tree-Depth (gt), Tree-Shape (u)
(34,4)	Surf-area (bet), Tree-Depth (lt), Tree-Shape (v)
(22,0)	Surf-area (lt), Maize (lt), Max-Maize (lt), Tree-Shape (v)
(22,0)	HighRisk (gt), Surf-area (gt), Maize (gt), Max-Maize (lt), Tree-Depth (lt)
(19,6)	Herb-Amount (lt), Surf-area (bet), Max-Maize (bet), Tree-Depth (gt)
(17,6)	Surf-area (bet), Maize (lt), Buffer (lt), Tree-Depth (gt), Tree-Top (f)
(17,12)	Pre (pre lt), Surf-area (bet), Maize (bet)
(16,0)	Maize (lt), Tree-Depth (lt), Tree-Top (cav)
(15,0)	Surf-area (lt), Max-Maize (lt), Tree-Depth (gt), Tree-Shape (u)
(15,1)	Surf-area (lt), Maize (lt), Tree-Depth (lt), Tree-Shape (v)
	Low transfer rate
(80,0)	Surf-area (bet), Tree-Depth (bet)
(74,1)	HighRisk (lt), Surf-area (lt), Maize (lt), Max-Maize (lt), Buffer (lt)
(57,0)	Surf-area (bet), Maize (gt), surf-max-maïs(gt), Tree-Depth (lt), Tree-Top (f)
(54,0)	Surf-area (gt), Maize (bet), Tree-Shape (i)
(42,0)	Herb-Amount (lt), Surf-area (bet), Max-Maize (lt), Tree-Depth (lt)
(40,0)	Herb-Amount (gt), Surf-area (bet), Maize(lt), Tree-Depth (lt), Tree-Top (f), Tree-Shape (u)
(38,0)	Surf-area (lt), Max-Maize (lt), Tree-Depth (gt), Tree-Top (vex)
(36,0)	Surf-area (bet), Max-Maize (lt), Tree-Top (cav)
(33,2)	Herb-Amount (lt), HighRisk (lt), Maize (lt), Max-Maize (lt), Tree-Top (f)
(33,4)	Herb-Amount (lt), Surf-area (bet), Maize (lt), Max-Maize (bet), Tree-Depth (gt)
(31,0)	Herb-Amount (gt), type (lt), Surf-area (gt), Maize (lt), Max-Maize (lt), Tree-Depth (lt)
(26,0)	Herb-Amount (gt), Max-Maize (lt), Buffer (gt)
(25,9)	Pre (gt), Maize (lt), Tree-Depth (lt)
(21,0)	Surf-area (lt), Maize (lt), Max-Maize (bet), Tree-Top (f)
(20,0)	Herb-Amount (lt), Surf-area (gt), Maize (bet), Max-Maize (lt)
(20,0)	Herb-Amount (bet), HighRisk (lt), Surf-area (gt), Max-Maize (lt), Tree-Depth (lt)
(16,4)	Herb-Amount (lt), Surf-area (bet), Maize (lt), Max-Maize (lt)

(C, NC): (number of covered examples of the same class, number of covered examples of the other class); lt: lower than; gt: greater than; F: false; T: true; 0: flat. For gt and lt, the values of the thresholds are not indicated.

largest maize plot (Max-Maize, nbocc = 10). The most frequently selected attributes in the low-transfer class were the same (nbocc = 14, 12 and 11, respectively). The attributes related to the shape of the sub-trees were used infrequently in the rules.

Among the 130 attributes involved in the 34 rules of the attribute-value approach (Table 4), 37 in the high-transfer class and the 48 in low-transfer class could be considered source factors and 23 in the high-transfer class and 22 in the low-transfer class could be considered transfer factors (Table 2).

4.4. Visualisation tool

A visualisation tool was developed to allow users of the decision-aid system to see relationships between the plot outlet trees of the catchment and the learnt rules (Fig. 8). The user can open maps of the catchment, each representing a simulation (e.g., spatial distribution of crops and quantified herbicide treatments) for all the plot outlet trees of the catchment. The user also can open a file of the learnt rules (attribute-value rules or tree-structured patterns) and then ask for the relations between rules and plot outlet trees (examples) generalised by rules. It can be done "automatically" or by a query language.

In an automatic mode, the user has to select rules among the list of rules. Plot outlet trees generalised by these rules become red on the map if their predicted transfer rates are high or green if they are low. Note that the user can choose to request colouring on the map only of the examples used for learning or the examples used for testing, or both. As previously mentioned, rules are learnt only from high and low-transfer examples. "Medium-transfer-rate" plot outlet trees are not used but the user can request to see them if they can be generalised by the selected rules. These plot outlet trees are then orange-coloured. Fig. 8 shows six selected attribute-value rules (three high-transfer and three low-transfer) that generalise 98 learning examples (from 20 simulations). For the displayed simulation (opened map), there are 16 high-transfer examples, 14 low-transfer examples and 4 medium-transfer-examples. Each rule is shown with its support (number of examples generalised by this rule among learning examples), its accuracy (ratio of examples of the same class as the rule to all the generalised examples) and the number of generalised examples by the rule among displayed examples of the map.

In language query mode, the set of shown examples (coloured plot outlet trees in the left panel) and the set of shown rules (in the right panel) are linked according to the user request. For example, if two rules *ru1* and *ru2* are shown, the results of the underneath request will be to show the plot outlet trees that are covered by the rule *ru1* and not covered by the rule *ru2*.

$$diff\_e(covered\_by\_all(\text{“}ru1\text{”}), covered\_by\_all(\text{“}ru2\text{”}))$$

In this language, selection of examples and rules based on the covering relation are available: *covered_by_some, covering_all, etc*. Union, intersection and difference operators between set of examples or rules allow the user to explain the chosen examples or the most powerful rules.

As a decision-support aid, the interface incorporates a recommendation tool (DAKAR, Trépos et al., 2005). Starting from an undesired situation, such as a high-transfer plot outlet tree, and given a set of already learnt attribute-value rules, the recommendation tool estimates actions (attribute values to change) that could

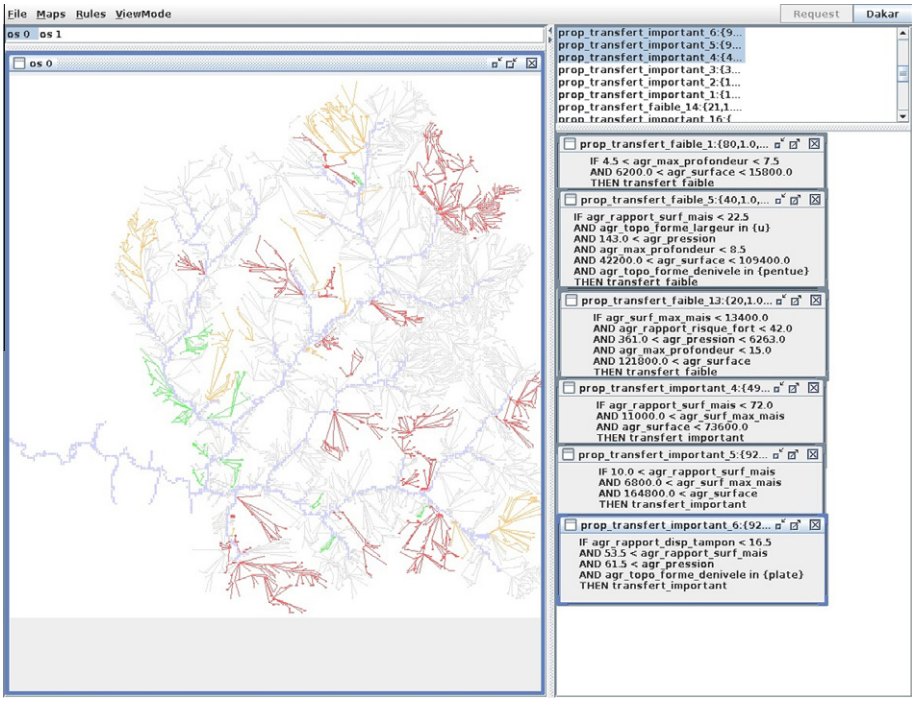

Fig. 8. Interface to visualise sub-trees. All sub-trees that correspond to a given rule are shown on a map. On the left, maps of the catchment, including the hydrological network and the plot outlet trees. On the right, learnt rules (attribute-value rules or tree-structured patterns). Plot outlet trees in red and green have high and low transfer rates, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

decrease the transfer rate. Finally, an advanced version of this tool allows users to launch simulations and to learn rules online. Users can choose, for example, the proportion of maize crops on the catchment area or modify the list of herbicides used on the catchment. Learnt rules have to be attribute-value, as the ILP mining method is not fast enough to be proposed for online use, i.e., with a direct interaction with the user.

4.5. Comparison of the two rule mining approaches

4.5.1. Efficiency

The CPU time necessary (especially to induce sub-trees) clearly favours the attribute-value approach, despite efforts to avoid redundancy in the generation of plot outlet sub-trees (Table 5). The number of correct classified testing examples favoured the attribute-value approach on the labelled-as-"high-transfer" examples (81%, vs. 67% for ILP), and was equal for the two approaches on the labelled-as-"low-transfer" examples (93% vs. 96% for ILP). Combining the two approaches improved classification results: grouping the attribute-value and ILP rules, 90% of classifications were correct, compared to 88% with only the attribute-value rules (averaging the two classes). Focusing on the labelled-as-"high-transfer" examples only, 88% of classifications were correct with combined rules, compared to 81% with only the attribute-value rules.

4.5.2. Correlation between attributes within each approach

For the ILP approach, the highest correlations were related to the topographic index (Top-Index), the attribute most frequently selected: 0.046 with Herb-Amount, 0.043 with Surf-area and 0.019 with Slope (Table 6). Other of the highest correlations included Surf-area with Maize (0.016), Maize with Herb-Amount (0.012) and Maize with Buffer (0.011). For the attribute-value approach as well, the most frequently selected attribute (Max-Maize) was most highly correlated with the other attributes (Table 7), especially with HighRisk (0.087). In the two methods, the presence of two attributes in a rule seemed generally independent; nonetheless, high correlation between two attributes in rules could be used to identify new attributes, especially in the attribute-value approach, where the selecting process was more difficult.

5. Discussion

Our approach analysed both qualitative and quantitative predictions from a model of an agro-environmental system and is an alternative to black-box approaches such as neural networks. As pointed out in Gibert and Sànchez-Marrè (2011), it is important to provide exclusive models to stakeholders.

5.1. Relationships between methods and attributes

For the ILP approach, source factors explained the high-transfer class better, while transfer factors explained the low-transfer class better. For the attribute-value approach, when considering the sub-tree as a whole object, the main factors controlling the transfer rate were source factors. This apparent contradiction between the results of the two methods can be explained by a scaling effect. When looking at the pattern of the sub-tree, the transfer factors that determine the flow connectivity from plot to plot were the most significant. When considering the plot outlet tree as an entity, the source factors dominated. Buffer-zone attributes were infrequently chosen in both methods, which can be explained by the low number of buffer zones in the studied catchment and not by a lack of influence. The attributes describing the transfer processes per sub-tree in the attribute-value approach (e.g., Tree-Depth, Tree-Shape and Tree-Top) were infrequent, suggesting that these attributes were insufficiently relevant to represent transfer processes by themselves. Due to the high frequency of the Top-Index attribute in the ILP approach, it would be interesting to test an attribute describing the proportion of the sub-tree surface area

Table 5
Results of the inductive logic programming (ILP) and attribute-value mining approaches computed on the set of test examples. In the confusion matrix, the top and bottom values on the left show the number of low-transfer examples (n = 582) correctly and incorrectly classified, respectively, while the top and bottom values on the right show the number of high-transfer examples (n = 587) incorrectly and correctly classified, respectively.

Set of rules	CPU time[a]	Number of rules	Correct classification	Confusion matrix	
ILP	≈18 h	28	81%	558	197
				24	390
Attribute-value	≈1 min	33	88%	544	108
				38	479

[a] Intel® Xeon 3.4 GHz processor.

Table 6
Mutual information (correlation coefficients) for occurrences of outlet attributes in induced inductive logic programming (ILP) rules.

	Maize	Buffer	Slope	Top-Index	Pre	Herb-Amount
Surf-area	0.016	0.005	0.001	0.043	0	0.001
Maize		0.011	0.001	0.009	0	0.012
Buffer			0.003	0.007	0	0.004
Slope				0.019	0	0.006
Top-Index					0	0.046
Pre						0

with an index value higher or lower than a threshold. In such a way, the ILP approach could be considered an initial step to create a list of relevant attributes for the sub-tree level. This process was easier for source factors; for example, although the attribute for presence of maize appeared frequently in the ILP approach, we introduced an attribute related to the largest surface area of maize in the tree, which then was used frequently in attribute-value rules.

Otherwise, herbicide amount, type of chemicals (High and Low-Risk) and application date (Pre) were not present in the rules, indicating that surface-area criteria were more influential than those related to chemical characteristics. When we averaged nine weather years, we may have blurred interactions between weather and herbicide applications, such as the time between application and conditions of soil-surface sealing or proximity of high-rainfall events and application. In so doing, the surface area involved may have become more important than the characteristics of applications. We focused on spatial patterns since we targeted stakeholder decisions regarding the interactions between the landscape structure and the weeding strategies. Learning spatio-temporal patterns would be technically more complex, but possible by considering different temporal dimensions. It might be achieved by aggregating the outputs on a shorter time step (week, month, etc.) then analyse the temporal variations of the spatial patterns vs. the weather characteristics during the corresponding period.

The number of rules can be considered as a limit of this approach. Identifying relevant patterns is time consuming. The cost in time is directly related to the number of rules. We are currently developing an application where simulated data are stored in a data warehouse. Interaction with the stakeholder is preferred for identifying spatial or temporal patterns (e.g., iterative SQL request).

5.2. From ILP to attribute-value approach

In the ILP approach, the identified spatial patterns provide an easily understandable explanation of the transfer, due to the description of the plot-to-plot connectivity. Nevertheless, some rules contained apparent errors, such as implausible thresholds (e.g., a maize surface area lower than a threshold) or irrelevant presence of attributes on upper outlets. For example, a low value for Top-Index, indicating an upslope position of a plot, is associated with completely different values of the surface area or herbicide amount. In such a case, it signifies that the latter attributes had no effect if they were related to a plot located uphill.

In the attribute-value approach, the same difficulty can occur. For example, a surface area or a herbicide amount greater than given threshold values were generally associated with higher maize or maximal maize areas, which could be considered as dominant factors. To avoid difficulties in interpreting the selection of attributes in a rule, we could select only uncorrelated attributes. Such a solution would only work for the attribute-value approach. Indeed, in the ILP approach developed at the subtree scale, all attributes can be meaningful since they described plot to plot connectivity into the context of a given subtree. Moreover, in the attribute-value approach, the thresholds themselves that are computed during the learning process are sometimes irrelevant (abnormal from a physical point of view) or non useful (too low). Finally, in order to combine the efficiency of attribute-value approach, i.e., fast learning and good classification results, and easy interpretation of ILP patterns, we could add Boolean attributes, based on the occurrence of spatial patterns identified in a learning pre-process step, i.e., spatial patterns present in the plot outlet tree.

6. Conclusions

This study aimed to test whether automatic learning techniques based on simulation data could help identify factors with strong

Table 7
Mutual information (correlation coefficients) for occurrences of aggregate attributes in induced attribute-value rules.

	LowRisk	HighRisk	Pre	Surf-area	Maize	Max-Maize	Buffer	Tree-Depth	Tree-Shape	Tree-Top
Herb-Amount	0	0.001	0	0.013	0.002	0.039	0.013	0.002	0.001	0.046
LowRisk		0	0	0	0	0	0	0	0	0
HighRisk			0.023	0.035	0.001	0.087	0	0.011	0.004	0.057
Pre				0.01	0.063	0.041	0.028	0.004	0.044	0.033
Surf-area					0.001	0.028	0.022	0.029	0.038	0.07
Maize						0.039	0.001	0.025	0.025	0.003
Max-Maize							0.001	0.031	0.025	0.062
Buffer								0.029	0.003	0.07
Tree-Depth									0	0.023
Tree-Shape										0.018

influence on herbicide stream-pollution, such as characteristics of source areas, to improve agricultural practices and management. The results are richer than simulation results at two levels: (1) at the catchment level, by identifying and analysing the reasons for stream water pollution in a given location; and (2) at the issue level (stream herbicide pollution), by inducing a set of rules that describe the factors controlling system function. We have shown the role of transfer factors, particularly topographic position, that appear to control low-transfer rates within a given sub-tree; the role of the surface area of the sub-tree; the importance of maize crops when considering the sub-tree as a whole; and the low influence of herbicide applications (type, date and application rate), when the transfer rates were averaged over several weather years.

This study's main contributions include:

- Model predictions are used rather than observations as a set of learning examples on which to apply automatic learning techniques. Simulations can produce a larger set of learning examples than observations because they can vary a larger number of factors over a wider range. The experimental design is richer; for example, the effect of crop location could not be tested by observations.
- A comparison of two approaches in rule learning from spatial data shows that: (1) ILP method describes well the plot-to-plot connectivity in a spatial way adapted to stakeholders (2) a fast and efficient attribute-value method that globally describes sub catchments but may appear more arbitrary for stakeholders. Therefore, an approach consisting in gradually introducing variables deduced from ILP approach within attribute-value approach could improve this.
- Output variables are rules (i.e., attributes included in relationships and spatial patterns). The identification of spatial patterns averaging temporal effects was approached by averaging results over a set of years considered to represent weather variability. The results show that spatial patterns can be identified regardless of the year's weather.
- The model identified spatial factors in a partially or completely qualitative approach. This approach is not specific to the studied application and could be applied to explore the functioning of any environmental system affected by spatial and temporal processes.

Acknowledgments

The authors thank the Chamber of Agriculture of Brittany (France) for providing the data. This work was financed under the project "Sacadeau" set up within the framework of INRA-CIRAD transverse action "Decision-making aid: how to link together knowledge and action in agriculture, agro-food industry and rural areas" and the project "APPEAU" in the French national research agency programme "Agriculture and sustainable development". The authors also thank Michelle and Michael Corson for their English and Scientific review.

References

Aurousseau, P., Gascuel-Odoux, C., Squividant, H., Tortrat, F., Cordier, M.O., 2009. A plot drainage network as a conceptual tool of spatial representation in agricultural catchments. Comput. Geosci. 35, 276–288.

Auteri, D., Azimonti, G., Galimberti, F., Ragni, P., 2007. Pesticide risk indicators: application of metaPEARL and ARI indicators to a pilot area in Northern Italy. In: 13th Symposium Pesticide, Chemical, pp. 633–641.

Beven, J.K., Kirkby, M.J., 1979. A physically based variable contributive area model of basin hydrology. Hydrol. Sci. Bull. 24, 43–69.

Brown, I., 2006. Modelling future landscape change on coastal floodplains using a rule-based GIS. Environ. Model. Softw. 21, 1479–1490.

Centofantia, T., Hollisb, J.M., Blenkinsopc, S., Fowlerc, H.J., Truckella, I., Dubus, I.G., Reichenbergere, S., 2008. Development of agro-environmental scenarios to support pesticide risk assessment in Europe. Sci. Tot. Environ. 407, 574–588.

Cerdan, O., Souchere, V., Lecomte, V., Couturier, A., Le Bissonnais, Y., 2001. Incorporating soil surface crusting processes in an expert-based runoff model: sealing and transfer by runoff and erosion related to agricultural management. Catena 46, 189–205.

Cestnik, B., 1990. Estimating probabilities: a crucial task in machine learning. In: Proceedings of the Ninth European Conference on Artificial-Intelligence (ECAI'90), pp. 147–149.

Chen, S.H., Jakeman, A.J., Norton, J.P., 2008. Artificial intelligence techniques: an introduction to their use for modelling environmental systems. Math. Comput. Simul. 78, 379–400.

Chertov, O., Komarov, A., Mikhailov, A., Andrienko, G., Andrienko, N., Gatalsky, P., 2005. Geovisualization of forest simulation modelling results: a case study of carbon sequestration and biodiversity. Comput. Electron. Agric. 49 (1), 175–191.

Chi, Y., Muntz, R.R., Nijssen, S., Kok, J.N., 2005. Frequent subtree mining – an overview. Fundam. Inform. 66, 161–198.

Clark, P., Boswell, R., 1991. Rule induction with CN2: some recent improvements. In: Proceedings of the Fifth European Working Session on Learning (EWSL'91), pp. 151–163.

Clement, M., Cann, C., Seux, R., Bordenave, P., 1999. Facteurs de transfert vers les eaux de surface de quelques phytosanitaires dans le contexte agricole breton. In: Pollutions diffuses: du bassin versant au littoral. Editions Ifremer, pp. 141–156.

Colin, F., Puech, C., de Marsily, G., 2000. Relations between triazine flux, catchment topography and distance between maize fields and the drainage network. J. Hydrol. 236, 139–152.

Dehaspe, L., Toivonen, H., King, R.D., 1998. Finding frequent substructures in chemical compounds. In: 4th International Conference on Knowledge Discovery and Data Mining, pp. 30–36.

Dubus, I., Beulke, S., Brown, C.D., 2002. Calibration of pesticide leaching models: critical review and guidance for retorting. Pest. Manage. Sci. 58, 745–758.

Dubus, I., Brown, C.D., Beulke, S., 2003. Sources on uncertainty in pesticide fate modelling. Sci. Tot. Environ. 317, 53–72.

Durand, P., Gascuel-Odoux, C., Cordier, M.O., 2002. Parameterisation of hydrological models: a review and lessons learned from studies of an agricultural catchment (Naizin, France). Agronomie 22, 217–228.

Fairfield, J., Leymarie, P., 1991. Drainage networks from grid elevation models. Water Resour. Res. 27, 709–717.

Fang, K.-T., Li, R., Sudjantio, A., 2005. Design and Modeling for Computer Experiments. Computer Science & Data Analysis Series. Chapman & Hall/CRC, 290 pp.

Gascuel-Odoux, C., Aurousseau, P., Cordier, M.O., Durand, P., Garcia, F., Masson, V., Salmon-Monviola, J., Tortrat, F., Trépos, R., 2009. A decision-oriented model to evaluate the effect of land use and agricultural management on herbicide contamination in stream water. Environ. Model. Softw. 24, 1433–1446.

Gibert, K., Sànchez-Marrè, M., Rodríguez-Roda, I., 2006. GESCONDA: an intelligent data analysis system for knowledge discovery and management in environmental databases. Environ. Model. Softw. 21, 115–120.

Gibert, K., Spate, J., Sànchez-Marrè, M., Athanasiadis, I., Comas, J., 2008. Data mining for environmental systems. In: Jackeman, A.J., Voinov, A., Rizzoli, A., Chen, S. (Eds.), Environmental Modeling, Software and Decision Support. State of the art and New Perspectives. IDEA Series v3. Elsevier NL, pp. 205–228.

Gibert, K., Sànchez-Marrè, M., 2011. Outcomes from the iEMSs data mining in the environmental sciences workshop series. Environ. Model. Softw. 26, 983–985.

Heathwaite, L., Sharpley, A., Gburek, W., 2000. A conceptual approach for integrating phosphorus and nitrogen management at watershed scales. J. Environ. Qual. 29, 158–166.

Heathwaite, A.L., Quinn, P.F., Hewett, C.J.M., 2005. Modelling and managing critical source areas of diffuse pollution from agricultural land using flow connectivity simulation. J. Hydrol. 304, 446–461.

Huber, K.P., Berthold, M.R., 1997. Simulation data analysis using fuzzy graphs. In: Proceedings of the Second International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data (IDA'97). Springer-Verlag, pp. 347–358.

Inokuchi, A., Washio, T., Motoda, H., 2000. An apriori-based algorithm for mining frequent substructures from graph data. In: Principles of Data Mining and Knowledge Discovery (PKDD'00), pp. 13–23.

Ivanovska, A., Vens, C., Colbach, N., Debeljak, M., Dzeroski, S., 2008. The feasibility of co-existence between conventional and genetically modified crops: using machine learning to analyse the output of simulation models. Ecol. Model. 215, 262–271.

Jakulin, A., Bratko, I., 2004. Quantifying and visualizing attribute interactions: an approach based on entropy. Computing Research Repository (CoRR), 308002, 30 pp.

Janssen, R., Goossen, H., Verhoeven, M., Verhoeven, J.T.A., Omtzigt, N., Maltby, E., 2005. Decision support for integrated wetland management. Environ. Model. Softw. 20, 215–229.

Kawano, S., Huynh, V.N., Ryoke, M., Nakamori, Y., 2005. A context-dependant knowledge model for evaluation of regional environment. Environ. Model. Softw. 20, 343–352.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. J. 97, 273–324 (Special issue on relevance).

Laburthe, F., 2000. Choco: implementing a cp kernel. In: Proceedings of Techniques for Implementing Constraint programming Systems (TRICS'00), pp. 118–133.

Letcher, R.A., Croke, B.F.W., Jakeman, A.J., 2007. Integrated assessment modelling for water resource allocation and management: a generalised conceptual framework. Environ. Model. Softw. 22, 733–742.

Leu, C.M., Singer, H., Stamm, Ch., Müller, S.R., Schawarzenbach, R.P., 2004a. Simultaneous assessment of sources, processes, and factors influencing herbicide losses to surface waters in small agricultural catchment. Environ. Sci. Technol. 38, 3827–3834.

Leu, C.M., Singer, H., Stamm, Ch., Müller, S.R., Schawarzenbach, R.P., 2004b. Variability of herbicide losses from 13 fields to surface waters within a small catchment after a controlled herbicide application. Environ. Sci. Technol. 38, 3835–3841.

Louchart, X., Voltz, M., Andrieux, P., Moussa, R., 2001. Herbicides runoff at field and watershed scales in a Mediterranean vineyard area. J. Environ. Qual. 30, 982–991.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297.

Mladenic, D., Bratko, I., Paul, R.J., Grobelnik, M., 1994. Using machine learning techniques to interpret results from discrete event simulation. In: Proceedings of the European Conference on Machine Learning (ECML'94). Springer-Verlag, pp. 399–402.

Muggleton, S., De Raedt, L., 1994. Inductive logic programming: theory and methods. J. Logic Program. 19/20, 629–679.

Nakano, S., Uno, T., 2003. Efficient Generation of Rooted Trees. National Institute of Informatics, Tokyo. Report NII-2003-005E, 4554.

O'Callaghan, J.F., Mark, D.M., 1984. The extraction of drainage networks from digital elevation data. Comput. Vis. Graph. Image Process. 28, 323–344.

Painter, M.K., Erraguntla, M., Hogg, G.L., Beachkofski, B., 2006. Using simulation, data mining, and knowledge discovery techniques for optimized aircraft engine fleet management. In: Perrone, L.F., Wieland, F.P., Liu, J., Lawson, B.G., Nicol, D.M., Fujimoto, R.M. (Eds.), Proceedings of the 2006 Winter Simulation Conference.

Poch, M., Comas, J., Rodríguez-Roda, I., Sànchez-Marré, M., Cortés, U., 2004. Designing and building real environmental decision support systems. Environ. Model. Softw. 19, 857–873.

Salmon-Monviola, J., Gascuel-Odoux, C., Garcia, F., Tortrat, F., Cordier, M.O., Masson, V., Trépos, R., 2011. Simulating the effect of technical and environmental constraints on the spatio-temporal distribution of herbicide applications and stream losses. Agric. Environ. Ecosyst. 140, 382–394.

Stenemo, F., Lindahl, A.M.L., Gärdenäs, A., Jarvis, N., 2007. Meta-modelling of the pesticide fate model MACRO for groundwater exposure assessments using artificial neural networks. J. Contam. Hydrol. 93, 270–283.

Srinivasan, A., Camacho, R.C., 1999. Numerical reasoning with an ILP program capable of lazy evaluation and customised search. J. Logic Program. 40, 185–214.

Srinivasan, A., 2003. Aleph manual, Version 4 and above. Available from: <http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph/aleph_toc.html/>.

Tan, R.R., 2005. Rule-based life cycle impact assessment using modified rough set induction methodology. Environ. Model. Softw. 20, 509–513.

Tiktak, A., Boesten, J.J.T.I., Van Der Linden, A.M.A., Vanclooster, D.M., 2006. Mapping ground water vulnerability to pesticide leaching with a process-based metamodel of EuroPEARL. J. Environ. Qual. 35, 1213–1226.

Tortrat, F., 2005. Modélisation orientée décision des processus de transfert par ruissellement et subsurface des herbicides dans les bassins versants agricoles. Thèse de l'Ecole Nationale Supérieure Agronomique de Rennes, UMR SAS INRA Agrocampus Rennes, 161 pp.

Trépos, R., Salleb, A., Cordier, M.O., Masson, V., Gascuel-Odoux, C., 2005. A distance approach for action recommendation. In: Proceedings of European Conference on Machine Learning, pp, 425–436.

Van Laer, W., 2002. From Propositional to first order logic in machine learning and data mining – induction of first order rules with ICL. Ph.D. thesis. Department of Computer Science, K.U. Leuven, Belgium, 239 pp.

Viaud, V., Merot, P., Baudry, J., 2004. Hydrochemical buffer assessment in agricultural landscapes from local to catchment scale. Environ. Manage. 34, 559–573.

Voltz, M., Louchart, X., Andrieux, P., Lennartz, B., 2003. Processes of pesticide dissipation and water transport in a Mediterranean farmed catchment. IAHS Publication 278.

Zhang, J., Bala, J., Barry, P.S., Meyer, T.E., Johnsson, S.K., 2002. Mining characteristic rules for understanding simulation data. In: Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02). IEEE Computer Society, p. 381.