A machine learning approach for evaluating the impact of land use and management practices on streamwater pollution by pesticides

M.O. Cordier[1], F. Garcia[2], C.Gascuel-Odoux[3], V. Masson[1], J. Salmon-Monviola[3], F. Tortrat[3], R. Trepos[1]

[1]IRISA/Univ. Rennes 1, Rennes [2]INRA UBIA, Toulouse, [3]INRA-ENSAR UMR SAS, Rennes, France

E-Mail: fgarcia@toulouse.inra.fr

Keywords: *stream water quality, modeling, simulation, machine learning.*

EXTENDED ABSTRACT

Streamwater pollution by pesticides is a critical environmental issue in farmed catchment areas. Many important factors are involved in this pollution phenomenon, like weather, area topology, land use and crop management practices, which all influence streamwater quality.

The purpose of the ongoing study presented in this paper is to evaluate the impact of land use and management practices on streamwater pollution. We use modelling, simulation and machine learning techniques for acquiring knowledge about this complex domain. Our main objective is to learn qualitative rules relating the pollution factors to the temporal distribution of the stream pesticide concentration. The study area is the farmed catchment of Fremeur (~17 km2), located in Brittany, France.

Our approach relies on a simulation model, called SACADEAU, which is based on two main components: a biophysical transfer model and a decision model. The biophysical transfer model simulates pesticide transfer through the catchment, from application sites on maize parcels, to the river. The decision model simulates farm management practices such as tillage, sowing, and pesticide application. The two other components of the SACADEAU model include a climate model which provides daily rainfall and temperature, and a spatial model which describes land use and catchment topology (Figure 1).

This simulation model is used for generating a large number of scenarios of the catchment system, considering different weather series or spatial distributions of land use and agricultural activities.

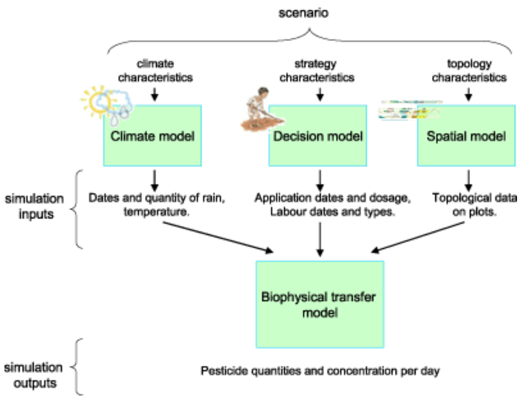

Figure 1: The SACADEAU model

Machine learning techniques were used to interpret the very complex and large set of results. ICL, an inductive logic programming software, generated a set of simple rules which described the factors influencing streamwater contamination. This demonstrated that soil characteristics, and in particular organic carbon content, are a key factor controlling contamination. Other important factors are: type of pesticide used, timing and quantity of rainfall, and topology of the catchment.

The Sacadeau model is not yet fully implemented and The first results have been obtained with a simplified model. We were able to check the coherence and the feasibility of our approach, and to build a first view of the role of some attributes in stream-water quality. When the SACADEAU model is fully operational, it should be possible to develop more specific rules that incorporate a greater level of details about spatial and temporal variations.

1. INTRODUCTION

Water quality is a critical and complex environmental issue. It is characterised by biological, physical and chemical components. Each component depends on numerous processes that are highly variable with time and space. Time variations are related rainfall events, intra-annual variations and inter-annual trends. Space variations may be due to local features such as human activities or management or global change such as crop system evolution. Since a large number of parameters are involved in pollution phenomena, modelling, simulation and machine learning appears to be useful techniques for acquiring knowledge in this complex domain.

The objective of the SACADEAU project is to build a decision-aid tool to help specialists in charge of catchment area management to preserve streamwater quality. It is focused on herbicides stream water pollution that may be rather high during the few months after applications, and dedicated to medium size catchment (5-100 km²). A qualitative transfer model, simulating herbicide transfer through the catchment area, is coupled with a management model, simulating farmers' decisions concerning weeding strategies and herbicide applications. The main objective of SACADEAU is to evaluate the impact of land use and management practices by simulating scenarios, and by discovering discriminating variables and acquiring knowledge in this complex domain through the use of machine learning techniques.

2. THE EXPERIMENTAL SITE

The experimental site, the Fremeur catchment area, is located in Brittany, France and covers about seventeen square kilometres (Figure 1).

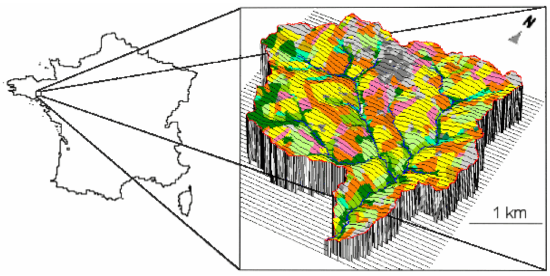

Figure 1: The Fremeur catchment area.

It was chosen because it is one of 43 catchments, from 10 to 100 km² in area, that together represent 60% of the regional drinkable water resource and that are specifically observed within the *Bretagne Eau Pure* Project: data on soil, climate, land cover, agricultural practices are thus numerous. This catchement is very close to Kervidy-Naizin catchment where hydrological processes were studied (Bruneau *et al*., 1995 ; Crave and Gascuel-Odoux, 1997 ; Molenat *et al*., 2002, 2005).

The Fremeur stream is 28 km long and presents 1,65 km/km² density. The average annual precipitation is about 900 mm. The soils are silty loam, with an organic matter content about 50g/kg. The weathered bedrock, 1 to 30 m deep, and fissured and fractured Brioveran Schist underlies the soil. A shallow and perennial groundwater mainly controls runoff processes. A survey made in 2001 shows that the agricultural land use (72%) is essentially intensive farming. The agricultural surface area was covered by maize (30%), wheat (30%), meadow (30%) and leguminous plant (10%). The remainder surface area was covered by forest, road or housing.

3. GENERAL FRAMEWORK OF THE SACADEAU MODEL

The Sacadeau model simulates streamwater contamination by herbicides spraying on maize crops. It models herbicide transfer through the catchment area and simulates the resulting mean daily river contamination. This phenomenon depends on numerous parameters, including human activities, climate, soil type and fields and catchment area topology. Since different topics are involved, we created three sub-models to describe them (see Figure 2):

- A decision model, which models farmers' strategies. This provides herbicide application characteristics (date, substance, quantity) and agricultural interventions (soil tillage, sowing and weeding dates) according to predefined farmers' strategies and weather conditions.
- A climate model, which provides daily weather data such as temperature and rainfall amount.
- A spatial model, which distributes in space the agricultural activities, according to the fields and catchment area topology.

Using the outputs of these three sub-models, a biophysical transfer model determines herbicide outflow, modelling transfer from application locations, through the catchment area, to the stream.

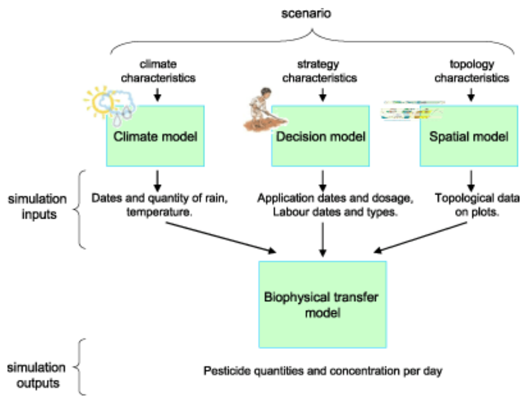

Figure 1: The SACADEAU model

At this step of the project, the spatial and the climate model have not been implemented. The experimental site was used as a framework for the spatial model. The catchment database was used to build scenario on soil type, locations of crop and herbicide application, extension of buffer zones. A database of thirty years of weather record was used in place of the climate model, from three sites (700, 900 and 1200 mm annual rainfall amount), with ten years data per site.

4. THE TRANSFER MODEL

Modelling pesticide transfer in agricultural catchments is very challenging because of difficulties to represent numerous hydrological processes and spatial structures in an adequate way (Moussa et al., 2002). We propose here an oriented decision making framework. The transfer model takes into consideration surface and subsurface flow (Tortrat, 2005). It runs on a short period of few months after herbicide applications, when transfer processes are dominant, with a daily time step because of the availability of climatic and hydrological data. It is based on an original spatial scheme. Common models take in consideration a regular squared grid to model transfer from pixel to pixel. We preferred to chose the parcels as the core model elements because of the uniformity of input data on agricultural practices and the interest of expressing outflow per field for decision making.

Surface flow is estimated from soil surface conditions. The concepts were developed in STREAM model, an expert-based runoff model (Cerdan et al., 2001). In this model, decision rules in the form of matching tables characterizing agricultural fields according to soil surface conditions (roughness, soil surface structure, crop cover) were used to determine infiltration capacity. Here, a daily mean infiltration capacity was considered and calibrated from regional data. It was assumed to be uniform per parcel. Surface flow is computed per parcel and aggregated on the whole catchment by using a tree structure linking field outlets and their contributive areas (Tortrat, 2005). The usual pixel drainage network was so replaced by a parcel outlet tree, and, finally, a parcel tree. For each parcel, it is necessary to define “parcel outlet” that may be unique or multiple. The flow pathways within the field were integrated as proposed by Souchère et al. (1998). The ditch and hedge networks that modify flow pathways at the bottom of the field were integrated in the field outlet tree (Tortrat, 2004).

The pesticide subsurface flow is calculated according to the depth of shallow groundwater from the soil surface. The temporal and spatial variations of the shallow water table depth were estimated by TOPMODEL (Beven and Kirkby, 1979), that allows good estimates for midslope and bottom domains (Molenat, 2005). Empirical relationships expressed in the forms of matching tables were used to estimate the contamination to the shallow ground water. It is equal to zero when the water table depth is deeper than 3m and exponentially increasing as depth is lower. A saturated surface flow was integrated. Each parcel is characterised by its topographic index distribution and relative surface area per class. The contribution to subsurface flow is computed per parcel and aggregated on the whole catchment in a single constant volume store with a constant drainage coefficient (Tortrat, 2005).

Water and pesticide transfer are coupled in a very simple way. The degradation is modelling by a standard one order kinetic. A fixed exchange coefficient between soil and water is considered. The amount of herbicide in soil is daily computed according to the degradation and transfer processes.

5. THE DECISION MODEL

The purpose of the management model is to simulate farmers technical decisions concerning maize crops on the catchment that explain herbicide pollution. This includes soil tillage, sowing and herbicide applications during springtime. This model takes as input the spatial distribution of crops on the catchment. This allocation of crops to parcels is produced by the spatial model.

5.1. Weed control programs in maize crops

Weeding programs in maize crops consist in planned but adaptive combinations of weed control operations like post-emergence mechanical cultivation, post-emergence thermal weeding, herbicide spraying or cover crop seeding, used under different primary tillage methods. As the objective of the SACADEAU project is to analyse the transfer of herbicides over a catchment area, we were especially interested here in modeling herbicide spraying strategies.

A survey identified three main weed control strategies based on herbicide spraying on the Fremeur catchment area:

- a pre-emergence strategy with a single application after sowing;
- a post-emergence strategy with two applications at 3 leaves and 5-7 leaves stages ;
- an intermediate strategy, with two applications after sowing and at 5 leaves stage.

Chemicals are chosen in relation with weeding strategies of the farmers. The recommended rates are generally applied. Application dates on a parcel are directly related to the type (pre-emergence, post-emergence) of the weeding strategy and on the maize plants growth . This one depends mainly on sowing date and weather conditions after sowing. Two main periods for sowing were observed on the Fremeur catchment (beginning and end of April), generally depending on the parcel locations (upslope and downslope). Finally, tillage practices, which influence the infiltration/surface runoff process, had also to be considered.

5.2. Simulating farmers' decisions

Given management strategies for each maize parcels of the catchment, the management model simulates, on a daily basis, tillage, sowing and weeding applications as a function of weather conditions. At the parcel level, this model relies on a classical temporal windows approach previously used in decision models like Otelo (Aubry et al., 1994), Déciblé (Aubry et al., 2004), Moderato (Bergez et al., 2002) or Control-Dièse (Martin-Clouaire et al., 2005). Operations like sowing or weeding applications are only possible within some temporal intervals. The beginning and ending of these intervals are defined by crop evolution. When the temporal window of an operation is opened, the simulation model checks day after day whether the operation can be performed or not (Figure 2).

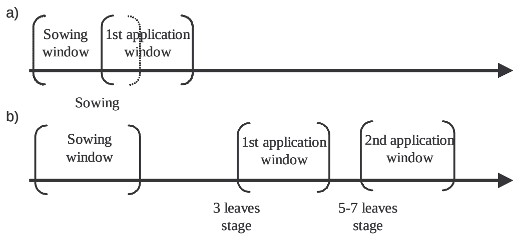

Figure 2: temporal scheme of decisions for pre-emergence and post-emergence strategies.

In order to execute an operation, different constraints need to be satisfied. At the stand level, constraints are related to weather and soil conditions. Farmers do not work on sowing or weeding applications when the weather is rainy. We modelled this condition by a threshold on the daily rainfall amount. Farm machines cannot work on a parcel when the soil bearing capacity is not high enough. We modelled this restriction as a number of days farmers have to wait before working on a parcel after an high rainfall event.

Decisions at parcel level are also constrained by machines and time availability. These constraints need to be considered in order to represent correctly the spatial and temporal distribution of operations on the catchment (Leenhardt et al., 2002). We thus introduced two upper organizational levels on the catchment. On farm level, farmers manage the set of their parcels. A farmer cannot work on all his parcels at the same time, and sowing and weeding operations have durations that depend on the speed and number of machines and the surface of the parcels. On a given day, the cumulated time of work cannot go over the maximum working hours of farmers. At the farm group level, sharing farm machinery takes place between farmers. Sowing and tillage machines are available at this level. They are allocated to farmers depending on the number and priority of parcels they need to work on. Farmers have to wait for these machines before starting the sowing or tillage operations.

When all these conditions are fulfilled on a parcel, the current operation can be performed. The operation window is then closed and simulation goes on, waiting for the next window opening on this parcel, until the end of the simulation period.

6. ACQUIRING KNOWLEDGE BY MACHINE LEARNING

The complexity of inputs and outputs of the simulation model makes interpretation of results difficult. We thus developed an automatic tool for learning from simulation data qualitative rules expressed in a high level language.

6.1. A High-Level scenario Language

A scenario is a set of simulations under certain constraints. It can be seen as a qualitative question. An example of scenario is « What happens if a post-emergence weeding strategy is applied instead of a pre-emergence strategy on all the plots close to the river when the climate is rainy with many high rainfall events from May to July? ». The objective is to give a qualitative answer like « Herbicide concentrations are high when applications just precede (within 2 days) heavy rainfalls (qty > 10mm) ».

An initial step was to gather a set of scenarios suggested by experts so that relevant problems concerning streamwater quality were understood by members of the project. We then defined a scenario simulation methodology (Figure 3). The three main processes of this methodology are translation between qualitative descriptions and quantitative constraints, generation of a set of simulations and analysis of simulation results.

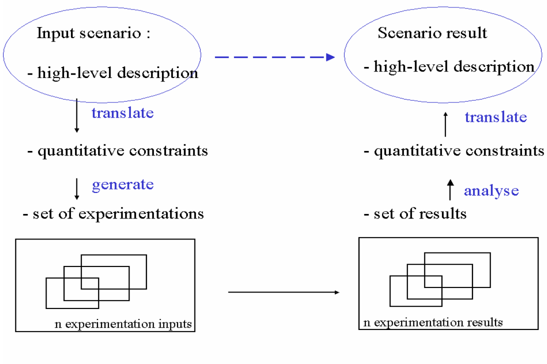

Figure 3: simulating scenarios

Translation between qualitative descriptions and quantitative constraints is a fundamental process if we want to construct comprehensive results for decision-makers. This can be done *a priori* by experts or by automatic process. For example, when data is described by continuous attributes, the translation consists in finding relevant thresholds provided by experts or machine learning techniques (clustering and automatic discretization). This can be also done *a posteriori*. For example, a climate corresponding to a rainy spring (regarding water contamination) is, for experts, a springtime with frequent and high rainfall events. After some experiments, we defined a new parameter which is a potential rainfall amount that may produce surface runoff, cumulated during the study period and using soil infiltrability. The second process is the definition of simulations. It can be a random choice of a subset of simulations for which quantitative constraints are verified. The last process is the analysis of simulations results. To get understandable results, simulation results have to be summarized. This step can be achieved by automatic learning techniques like rule learning.

6.2. Learning from Simulation Results

A lot of simulations are needed if we want to get relations concerning all the inputs and outputs of the model: farmers' strategies (pesticides and quantities used, spatial and temporal localisations of applications), catchment area topology, climate characteristics, pollution peaks. For these reasons, an automatic tool for summarizing the highly structured results was adapted and we opted for an Inductive Logic Programming (ILP) approach. Let us explain the goal of ILP for our streamwater quality concern.

Assume that the simulation outputs are of two types: the pollution classes + (no pollution) and – (pollution). The aim of ILP is to find logical discriminative relations between variables, that is relations – or rules - that are verified by many simulations of class – and none of the simulations of class +, and vice versa. Here is an example of such rule: "If atrazine is spraying on a parcel close to the river at most two days before a rainfall then simulation is of class –". This rule can be written with predicates *weeding*, *close* and *rainfall* as below:

$$\textit{weeding}(\text{day1,parcel,atrazine}) \wedge \textit{close}(\text{parcel,river}) \wedge \textit{rainfall}(\text{day2}),\ 0< \text{day2–day1} < 2 \Rightarrow \text{class -}$$

The singularity of an ILP learner is to find first order logic relations. Actually, a same variable (like variable parcel in the example) can be found in many predicates (*weeding* and close). To find such discriminative rules, the number of rules to assess is huge and has to be restricted. We chose ICL software (Van Laer, 2002) to construct this kind of relations. ICL uses a grammar to restrict the number of rules to assess.

7. RESULTS

In a first step, this approach has been tested with a simplified model that only simulates water and herbicide transfer per parcel, due to surface runoff, and agregates runoff on virtual catchments in a very simple way. The parameters were defined in relationships to regional data. Input data were: farmer's program and chemicals, defined per parcel; common soil characteristics and climate to all the parcels of the catchment; extension of buffer zone and slope distribution that buffers the outflow per parcel at catchment scale. Mean daily discharge acted as a factor of dilution at catchment scale. We simulated all the scenarios (i.e. inputs were not constrained) and propositional relations were learnt (i.e. simpler relations than first order rules were considered). Six types of parameters were defined; five of them were defined by experts and the sixth by automatic learning. Parameters defined by experts were:

- Soil characteristics of the catchment. The kinetic of the soil surface degradation, and so infiltrability, depends on organic matter content for soils with similar texture. Two soil types have been considered: 20 $g.kg^{-1}$ or 50 $g.kg^{-1}$ organic matter content.
- General topography of the catchment. Two values “concave” or “convex” may be chosen. A concave catchment area has more low slopes close to the river than a convex one.
- Farmer’s strategy, which takes the value “pre-emergence” or “post-emergence”. A pre-emergence strategy consists in a first application of pesticides just after the seeding date of maize whereas a post-emergence consists in a first application one month after the seeding date.
- The chemical used for maize crop, which can take one of the two values “atrazine” or “new”. Atrazine is now forbidden in France. New molecules which have shorter half time of degradation are now used.
- The extension of buffer zones close to the river: these act as sinks for water and herbicides. We considered two values: no buffer zones or extended in 90 percent of the border of the stream.

The climate typology was defined by automatic learning. To describe a climate (defined as temporal series of rainfall between April and August), we considered two attributes: the number of days which rainfall amount is higher than a threshold of 10 mm and the potential rainfall amount on the study period that may produced surface runoff, as defined previously. Thus, five types of climate series were distinguished by ways of clustering methods. A climate of type 0 is a very dried climate and a climate of type 5 is a very wet climate.

Setting values for these parameters, we generated 896 instances. For each instance, a pollution class (among five classes) was associated. The class 0 represents no pollution and class 4 represents high pollution. These classes of pollution were defined with the sum of herbicide concentrations (between april and august) and the number of high values during the same period. Propositional rules were learnt with ICL. 70 rules were learned to summarize the 896 instances. An analysis of the relative importance of the attributes is given below (we call an *example* a list of attribute-value pairs and a pollution class):

All (100%) the rules of class 0 explain an example by a few rainy climate and a soil with an 50 $g.kg^{-1}$ organic matter content. 74% of the rules of class 1 and 59% of rules of class 2 explain an example by the presence of buffer zones in 90% of the parcels close to the river. 74% of the rules of the class 1 explain an example by the use of new chemicals. 67% of the rules of class 3 explain an example by an 20 $g.kg^{-1}$ organic matter content. 88% of the rules of class 4 explain an example by an 20 $g.kg^{-1}$ organic matter content and no buffer zone.

From these results, we concluded that the soil characteristics (organic matter content here) had an important role in water contamination. Let us give two rules, of different classes, with the same attribute-value list except for the soil characteristics:

1. if climate=3 and organic_matter=20 $g.kg^{-1}$ and buffer zone=90% and chemicals=new then class=2
2. if climate=3 and organic_matter=50 $g.kg^{-1}$ and buffer zone=90% and chemicals=new then class=1

The difference of one class corresponds to a factor of 10 on the sum of herbicide concentrations. A soil with a lower organic matter content resulted in higher levels of pesticide contamination of streamwater. Considering the others attributes, we concluded that the use of new chemicals, no rainy climate and a concave basin were favourable factors.

Conclusions given above on the relative importance of the attributes were largely expected. The impact of farmers’ strategies was less expected. The pre-emergence strategy appeared to be a more favourable strategy than the post-emergence one. A possible reason was that late herbicides applications met degraded soil surface conditions that improved surface runoff.

8. CONCLUSIONS

We developed a machine learning approach for acquiring knowledge about streamwater pollution by maize crop herbicides from a simulation model that represents transfer and farmers' decision processes. First results have been obtained with a simplified model. We were able to check the coherence and the feasibility of our approach, and to build a first view of the role of some attributes in stream-water quality. When the SACADEAU model is fully operational, it should be possible to develop more specific rules that incorporate a greater level of details about spatial and temporal variations.

9. ACKNOWLEDGMENTS

The SACADEAU project is in its third year of development. Contributing members are M. O. Cordier, V. Masson, A. Salleb (IRISA/Univ. Rennes 1, Rennes), C. Gascuel-Odoux, F. Tortrat, P. Aurousseau, R. Trepos (INRA-ENSAR/UMR SAS, Rennes), F. Garcia (INRA/BIA, Castanet Tolosan), B. Chanomordic (INRA/LASB, Montpellier), M. Falchier, D. Heddadj and L. Lebouille (Chambres d'agriculture de Bretagne). SACADEAU is funded by Conseil Général du Morbihan and INRA.

10. REFERENCES

Aubry, C., F. Papy, and A. Capillon (1998), Modeling of decision-making processes for annual crop management, *Agricultural Systems*, 56, 45-65.

Bergez, J.-E., Debaeke, P., Deumier, J.-M., Lacroix, B., Leenhardt, D., Leroy P., and Wallach, D (2001). MODERATO: an object-oriented decision tool for designing maize irrigation schedules, *Ecological Modelling*, 137-1, 43-60.

Chatelin, M.-H., Aubry, C., Poussin, J.-C., Meynard, J.-M., Massé, J., Verjux, N., Gate, P. and Le Bris, X. (2005). DéciBlé, a software package for wheat crop management simulation, *Agricultural Systems*, 83-1, 77-99.

Beven J.K. et M.J. Kirkby (1979). A physically based variable contributive area model of basin hydrology. *Hydrol. Sc. Bull.*, 24, 43-69.

Bruneau P., Gascuel-Odoux, C., Robin, P., Merot, P. et Beven, K.J. (1995). Sensitivity analysis to time and space resolution on an hydrological modelling based on Digital Elevation Model. *Hydrol. processes*, 9, 69-81.

Cerdan O., Souchère V., Lecomte V., Couturier A., Le Bissonnais Y. (2001). Incorporating soil surface crusting processes in an expert-based runoff model : Sealing and Transfert by Runoff and Erosion related to Agricultural Management. *Catena*, Vol. 46, 189-205.

Crave A. et Gascuel-Odoux C. (1997). The influence of the topography on time and space distribution of soil surface water content. *Hydrol. process.*, 11 (2), 203-210.

Leenhardt, D. and Lemaire, P. (2002), Estimating the spatial and temporal distribution of sowing dates for regional water management, *Agricultural Water Management*, 55-1, 37-52.

Martin-Clouaire, R. and Rellier, J.-P. (2005). Representing and interpreting flexible production management plans. *Int. Conf. on Conceptual Modeling and Simulation*, Marseille, France.

Molénat J. et Gascuel-Odoux C (2002). Modelling flow and nitrate transport in groundwater for the prediction of water travel times and of consequences of land use evolution on water quality. *Hydrol. process.*, 16, 479-492.

Molénat J., Gascuel-Odoux C., Davy P., Durand P. (2005). How to model shallow water-table depth variations: the case of the Kervidy-Naizin catchment, France. *Hydrol. process.*, 19, 901-920.

Moussa R., Voltz M., Andrieux P. (2002). Effects of the spatial organization of agricultural management on the hydrological behaviour of a farmed catchment during flood events. *Hydrol. Process.*, 16, 393-412.

Souchère V., King D., Daroussin J., Papy F., Capillon A. (1998). Effects of tillage on runoff directions: consequences on runoff contributing area within agricultural catchments. *Journal of Hydrology*, 206, 256-267.

Tortrat, F., Aurousseau P., Squividant H., Gascuel-Odoux C., Cordier M.O. (2004). Modèle Numérique d'Altitude (MNA) et spatialisation des transferts de surface: utilisation de structures d'arbres reliant les exutoires de parcelles et leurs surfaces contributives. *Bulletin SFPT*, 172, 128-136.

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, 161p.

Van Laer W. (2002) From Propositional to First Order Logic in Machine Learning and Data Mining. *Thesis*. Department of Computer Science, Leuven University, Belgium.