The feasibility of co-existence between conventional and genetically-modified crops: Using machine learning to analyse the output of simulation models

Aneta Ivanovska[1], Celine Vens[2], Nathalie Colbach[3], Marko Debeljak[1], and Sašo Džeroski[1]

[1] Jožef Stefan Institute, Department of Knowledge Technologies
Jamova 39, 1000 Ljubljana, Slovenia
{aneta.ivanovska, saso.dzeroski, marko.debeljak}@ijs.si
[2] K.U.Leuven, Department of Computer Science
Celestijnenlaan 200A, B-3001 Leuven, Belgium
celine.vens@cs.kuleuven.be
[3] UMR 1210, Biologie et Gestion des Adventices, INRA
21000 Dijon, France
colbach@dijon.inra.fr

Abstract. Simulation models are a commonly used tool for the study of the co-existence between conventional and genetically-modified (GM) crops. Among other things, they allow us to investigate the effects of using different crop varieties, cropping systems and farming practices on the levels of adventitious presence of GM material in conventional crops. We propose to use machine learning methods to analyze the output of simulation models to learn co-existence rules that directly link the above mentioned causes and effects. The outputs of the GENESYS model, designed to study the co-existence of conventional and GM oilseed rape crops, were analysed by using the machine learning methods of regression tree induction and relational decision tree induction. Co-existence and adventitious presence of GM material were studied in several contexts, including gene flow between pairs of fields, the interactions of this process with farming practices (cropping systems), and gene flow in the context of an entire field plan. Accurate models were learned, which also make use of the relational aspects of a field plan, using information on the neighboring fields of a field, and the farming practices applied in it. The use of relational decision tree induction to analyze the results of simulation models is a novel approach and hold the promise of learning more general co-existence rules by allowing us to vary the target field within a chosen field plan, as well as consider completely different field plans at the same time.

1 Introduction

Crop varieties developed by genetic engineering were first introduced for commercial production in 1996. Today, these crops are planted on more than 167 million acres worldwide. Genetically-modified (GM) crops are usually engineered to tolerate herbicides and/or resist pests. Crops carrying genes coding for herbicide tolerance were developed so that farmers could spray their fields with non-selective herbicides to eliminate weeds irrespective of species and stage without damaging the crop. Likewise, pest-resistant crops have been engineered to contain a gene for a protein from the soil bacterium, *Bacillus thurigiensis*, which is toxic to certain pests. This protein, referred to as Bt, is produced by the plant, thereby making it resistant to insect pests like the European Corn Borer (*Ostrinia nubilalis*) or Cotton Boll Worm (*Helicoverpa zea*). Other pest-resistant GM crops on the market today have been engineered to contain genes that confer resistance to specific plant viruses. So the main purpose of growing genetically-modified crops in a developed European agriculture is not to achieve higher yields, but to reduce producers' inputs and operating costs.

However, genetically-modified crops were not primarily developed with environmental benefit in mind and the introduction of transgenic crops and foods into the existing food production system has generated a number of questions about possible negative consequences. These concern the co-existence issue, i.e., the economic damage caused by GM contamination of conventional crops; the unwanted ecological influences of GM crops on habitats in natural and agricultural environments; and the consequences of exposure of humans to transgenic proteins.

The possible unwanted influence of consuming GM crops on the human health and the influence of growing GM crops on the habitats in natural and agricultural environments are topics of ongoing research. The main concern in this paper is the co-existence issue, i.e., the possibility of GM plants mixing with conventional or organic crops. GM crops can contaminate other crops simply by pollen being transported from one field to another. In addition, for species such as oilseed rape, seeds lost before or during the harvest survive in the soil and give rise to volunteers in subsequent crops. If these volunteers emerge in later non-GM oilseed rape crops, they lead to the adventitious presence of GM seeds in non-GM harvests.

Corn (maize) and oilseed rape (OSR) are the most important transgenic crops in Europe. EU regulations allow 0.9% of adventitious presence of GM material in conventional harvests and the co-existence is concerned with achieving the prescribed level of adventitious presence in regions with both conventional and transgenic cultivars. Therefore, there is a need to find appropriate measures at the farm and regional levels to minimize gene flow from GM crops.

To study the co-existence issue for the above two crops, computer simulation models have been developed (e.g., GENESYS - for oilseed rape, MAPOD - for corn) [7, 8, 19]. Given a specific situation, e.g., a specific field plan and a set of chosen farming practices, the simulation models give predictions for the levels of adventitious presence in the fields under study. By analyzing and aggregating the results of many such simulations, one can gain insight about the conditions under which co-existence is possible. For example, a JRC (Joint Research Center of the European Commission) study [19] gives some recommendations regarding co-existence of conventional and GM corn, produced by analysing the results of MAPOD simulations.

In this study we propose the use of machine learning methods to analyze the results/outputs of the simulation model GENESYS to gain insight into co-existence issues. Machine learning methods derive general knowledge from specific examples. By applying machine learning methods, we would generalize over the specific outputs of individual simulations and derive more general rules concerning the co-existence of conventional and GM crops.

We use machine learning to analyze the outputs of two sets of GENESYS simulations. The first studies the effects of relative size and position of fields on gene flow (via pollen or seed) between pairs of fields. The second examines the adventitous presence of GM seeds in the central field of a high-risk field pattern. To the outputs of each set of simulations, we apply suitable machine learning techniques: we use regression trees for the first and relational decision trees for the second.

2 The GENESYS simulation model

The computer model GENESYS was used to assess probable effects of changing farming practices on contamination rates. GENESYS [7, 8] was developed by INRA (French National Institute for Agronomy Research) to rank cropping systems according to their probability of gene flow from herbicide-tolerant winter oilseed rape to rape volunteers and neighbor crops, both in time via seeds and in space via pollen and seeds. The model works for seed as well as crop production. GENESYS integrates various input variables (Figure 1):

- The field plan of the region, comprising cultivated fields as well as uncultivated field- and road-margins (hence "borders"). Borders consist of strips of spontaneous vegetation where rape volunteers can appear, produce pollen and seeds that are dispersed to fields and other borders;
- The crop rotation of each field;
- The cultivation techniques applied to each crop (summer tillage, primary tillage and tillage for seed bed preparation, sowing date and density, herbicide applications, cutting dates and seed loss at rape harvest) as well as the management of the borders (herbicides and/or cutting), and
- The type of the simulated gene (dominant A or recessive a), as well as the genotype of the rapeseed varieties.

The model is based on the life-cycle of oilseed rape, and includes both cropped and volunteer plants, starting with the seed bank at harvest and continuing with seedling emergence. Some of these seedlings become adults, flower and produce new seeds, part of which replenish the seed bank at the end of the season. The model calculates for each stage of the annual rapeseed life-cycle and for each field or border the number of individuals per m^2 (number of seeds in the seed bank, of seedlings etc.) and the proportions of these individuals with and without transgenes (e.g. contamination with GM seeds).

GENESYS has already been evaluated using independent data collected on farmers' fields and on the GMO field trials set up and managed by INRA and CETIOM (Centre Technique Interprofessionnel des Oléagineux Métropolitains, France) and other technical institutes [5]. The first comparisons of simulation and

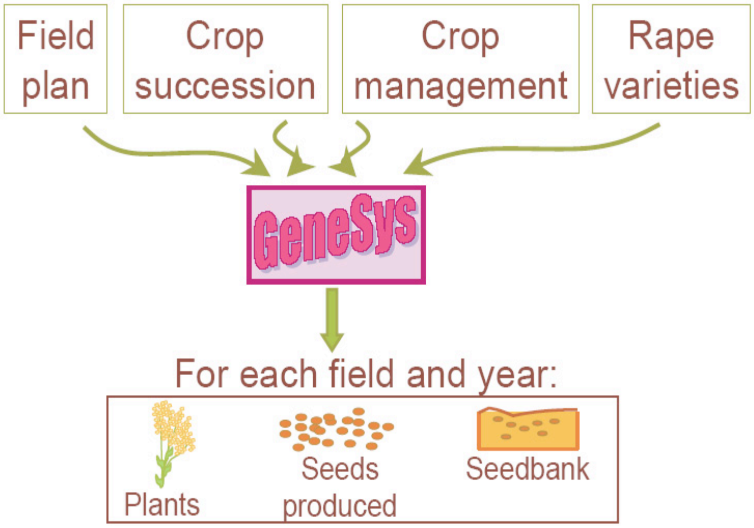

Fig. 1. GENESYS: input and output [6]

trial results show that the rates of contamination of harvested seeds are underestimated but that the orders of magnitude are reliable and that the various situations are ranked correctly. GENESYS may therefore be used to compare the effects of different cropping practices or of various varietal characteristics for decreasing the probability of contamination in the field.

3 Machine learning methods

This section describes the machine learning methods used to analyse the GENESYS simulation outputs. We first describe regression trees, which were used to learn co-existence rules for paired fields. Then we describe relational decision trees, which were used to learn co-existence rules and predict the rate of adventitious presence of GM seeds in the central field of a large-risk field plan.

3.1 Regression trees

In order to explain regression trees, we first describe decision trees [3]. Regression trees are namely a special type of decision trees.

Decision trees predict the value of a dependent variable (called target) from the values of a set of independent variables (called attributes), by partitioning the space of attributes into axis-parallel rectangles and fitting a model for each of these partitions. A decision tree (see for example Fig.3 or Fig.4) has a test in each inner node that tests the value of a certain attribute and compares it with a constant. Leaf nodes give a prediction that applies to all instances (examples) that reach the leaf. To predict the target of an unknown instance, it is routed down the tree according to the values of the attributes tested in successive nodes, and when a leaf is reached the instance is given the prediction, assigned to the leaf. If the dependent variable is nominal, the task is called classification, the predictions in the leaves are called classes, and the decision trees are called classification trees. If the dependent variable is numeric, then in each leaf there is a model for predicting it: the model can be a linear equation (model trees) or a constant (regression trees).

In order to build a decision tree, one makes use of a dataset of examples, for which the target is known. This dataset is called the training set. Tree construction proceeds recursively, starting with the entire training set. At each step a node is created and the most discriminating attribute is placed in the node. A number of new branches are created according to the values of the selected attribute. For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created based on that threshold. Technically speaking, the most discriminating attribute test is the one that most reduces the entropy/variance (for classification and regression trees respectively) of the values of the target. The training set is split into subsets by sorting down each example following the appropriate branch. For each subset, the tree construction algorithm is called recursively. Tree construction stops when the entropy/variance of the target values of all examples in a node is small enough (or if some other stopping criterion is satisfied). Such nodes are called leaves and are labeled with a class or a model (constant or linear equation) for predicting the target value.

An important mechanism used to prevent trees from over-fitting data is tree pruning. Pruning can be employed during tree construction (pre-pruning) or after the tree has been constructed (post-pruning). Typically, a minimum number of examples in branches can be prescribed for pre-pruning and a confidence level in the error estimates in the leaves for post-pruning.

A number of systems exist for inducing regression trees, such as CART [3] and M5 [20]. M5 is one of the most well-known programs for regression and model tree induction. We used the system M5' [22], a re-implementation of M5 within the software package WEKA [23].

A decision tree can be easily transformed into a set of rules. One rule is generated for each leaf. The rules are of form:

IF *conditions* **THEN** *prediction*

The antecedent of the rule includes a condition for every node on the path from the root to that leaf, and the consequent of the rule is the constant or the linear model assigned by the leaf. This procedure produces rules that are unambiguous in that the order in which they are executed is irrelevant.

3.2 Relational decision trees

Most machine learning algorithms assume that the training set is stored in a single table where each example is represented by a fixed number of attributes. These are called attribute-value or propositional techniques (as the patterns found can be expressed in propositional logic). Propositional machine learning techniques (such as the classification or regression decision trees discussed in the previous section) are popular, mainly because they are efficient, easy to use and are widely accessible.

In practice, however, the single table assumption turns out to be a limiting factor for many machine learning tasks that involve data residing in multiple related tables. An example of such a problem is the analysis of co-existence of GM and non-GM crops in a region with many fields, where there is a need to examine the relations among the fields. Typically, the data consists of several pieces of information; one could imagine having a table storing general information on each field (e.g. area), a table storing the cultivation techniques for each field and each year, and a table storing relations (e.g. distance) among pairs of fields. Data scattered over multiple relations (or tables) can be transformed into a propositional table (attribute-value representation) by means of propositionalization, so that conventional machine learning techniques can be applied to the transformed data [15]. This allows a wide choice of robust and well known algorithms. A disadvantage is that propositionalization almost inevitably leads to a loss of information due to aggregation or to the generation of a (possibly huge) amount of redundant data [12]. Also, if different examples can have a different number of fields (e.g., by varying the field plan), the propositionalization approach is not feasible. Alternatively, the relational approach takes into account the structure of the original data by providing functionalities to navigate relational structure in its original format and generate potentially new forms of evidence not readily available in a flattened single table representation.

Since decision tree induction is one of the major approaches to machine learning, upgrading this approach to a relational setting has been of great importance. Like in the propositional case, a table or relation is given, which contains at least two columns where the IDs of the examples and the values of the target variable are stored. An example of such a relation is *contamination(sim1, positive)*, which means that simulation 1 (example ID) is labeled as contaminated (target). (Recall from the introduction that a field is considered as contaminated if it contains more that 0.9% GM material.) In addition, a set of background knowledge relations, stored in other tables, may be given, as illustrated above.

Relational decision trees have much the same structure as propositional decision trees. Internal nodes contain tests, while leaves contain predictions for the target value. If the target variable is discrete/continuous, we talk about relational classification/regression trees. For regression, linear equations may be allowed in the leaves instead of constant class-value predictions: in this case we talk about relational model trees.

The major difference between propositional and relational decision trees is in the tests that can appear in the internal nodes. In the propositional case, the tests compare the value of an attribute to a constant. In the relational case the tests are conjunctions of relations, instantiated with variables (starting with upper case) and constants, and are mapped against the examples. For each example, a test results in 'yes' or 'no'. The conjuncts in the tests refer to background relations, while the leaves predict a value for the target in the target relation.

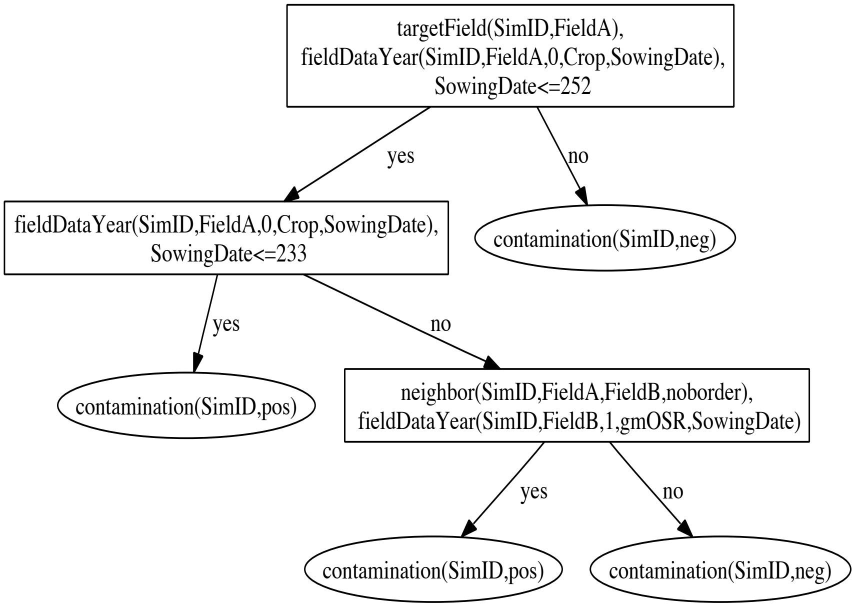
targetField(SimID,FieldA),
fieldDataYear(SimID,FieldA,0,Crop,SowingDate),
SowingDate<=252
yes
no
fieldDataYear(SimID,FieldA,0,Crop,SowingDate),
SowingDate<=233
contamination(SimID,neg)
yes
no
contamination(SimID,pos)
neighbor(SimID,FieldA,FieldB,noborder),
fieldDataYear(SimID,FieldB,1,gmOSR,SowingDate)
yes
no
contamination(SimID,pos)
contamination(SimID,neg)

Fig. 2. An example of relational classification tree predicting whether a field in a large-risk field plan is contaminated by a GM crop (Section 5).

An example of a relational classification tree for predicting the contamination of the central field of a large-risk field plan is given in Figure 2. The top node of the tree calls *FieldA* the target field we are interested in (*targetField(Sim,FieldA)*) and checks whether the sowing date of FieldA in the present year (year 0) is before the 252^{th} day of the year, i.e., 9 September (*fieldDataYear(Sim,FieldA,0,Crop,Sowing-Date), SowingDate<252*). If not, then the field is predicted not to be contaminated. If yes, there is another test that checks if the sowing date of FieldA in the present year is before the 233^{th} day of the year (21 August). If it is the case, then the field is predicted to be contaminated. If not, then the contamination depends on whether the target field has a neighboring field (called FieldB) with which it is adjacent (*neighbor(Sim,FieldA,FieldB,adjacent)*), and which had GM oilseed rape in the previous year (*fieldDataYear(Sim,FieldB,1,gm-OSR,SowingDate)*). Remark that this kind of test can not be found by a propositional system. A propositional decision tree can only refer to a particular field, e.g., it can check whether field 20 had GM oilseed rape in the previous year, but it can not check this for any neighbor field without enumerating them all.

For easier inspection and comprehensibility relational decision trees can be transformed/reformulated into relational decision lists, i.e., ordered lists of relational rules. When applying a decision list to an example, we always take the first rule that applies and return the answer produced. A decision list is produced by traversing the relational decision tree in a depth-first fashion, going down left branches first. At each leaf, a rule is output that contains the prediction of the leaf and all the conditions along the left (yes) branches leading to that leaf.

The two major algorithms for inducing relational decision trees are upgrades of the two most famous algorithms for inducing propositional decision trees. SCART [17, 18] is an upgrade of CART [3], while Tilde [1, 13] is an upgrade of C4.5 [20]. Both SCART and Tilde have their propositional counterparts as special cases. The actual algorithms thus closely follow CART and C4.5.

In our relational data analysis, we used the system Tilde for building relational classification trees. The algorithm is included in the ACE-ilProlog data mining system [4].

3.3 Evaluating predictive performance

For classification problems it is natural to measure a classifier's performance in terms of accuracy. The classifier (in our case a predictive model in the form of a propositional or relational classification tree) predicts the class of each example: if the prediction is correct, that is counted as success; if not, it is an error. The accuracy is the proportion of successful predictions (classifications) made over the whole set of instances, and it measures the overall performance of the classifier [23].

The performance of machine learning methods for regression (such as propositional and relational regression trees) can be measured by the *correlation coefficient*, which measures the statistical correlation between the predicted and the real values of the target variable. The correlation coefficient ranges from 1 for perfectly correlated results, through 0 when there is no correlation, to -1 when the results are correlated perfectly, but negatively. Other performance measures for regression include root mean squared error (RMSE), relative RMSE (RRMSE) and mean absolute error (MAE).

As mentioned before, the data on which we build a predictive model is called the *training set/data*. Normally, we are interested in the future performance of the model on new data. The accuracy on the training set is not a good indicator for future performance, since the classifier has been learned from the very same training data and any estimate of performance based on that data will be very optimistic. Thus, to evaluate the performance of a classifier, we have to assess its accuracy on a dataset that played no part in the formation of the model. This independent dataset is called *test set*. It is assumed that both training data and the test data are representative samples of the underlying problem.

If large sets of data are available, a large sample is taken for training, and another, independent large sample of different data for testing. However, in many real problems the data is limited, and in this case a certain amount of the dataset is set aside for testing, and the remainder is used for training (this is called a *holdout* procedure).

In general, we cannot tell whether a sample chosen from the dataset is representative or not, so to avoid any bias caused by the particular sample chosen for holdout, the whole process, training and testing, is repeated several times with different random samples. This technique is called *cross-validation.* In cross-validation, we decide on a fixed number of *folds*, or partitions of the data. Suppose we use k folds, then the data is split into k approximately equal partitions and each in turn is used for testing and the remainder is used for training. This procedure is repeated k times so that, in the end, every partition has been used exactly once for testing. This is called k-fold cross-validation. At the end, the accuracies on the different iterations are averaged to yield an overall accuracy.

In practice, 10-fold cross-validation is used. Extensive tests on numerous datasets, with different learning techniques, have shown that 10 is about the right number of folds to get the best estimate of the accuracy, and there is also some theoretical evidence that backs this up [23]. Although there is still a debate what is the best scheme for evaluation, 10-fold cross-validation has become a standard method for evaluation of machine learning methods.

4 Learning co-existence rules for pairs of fields

This part of the analysis presents the application of propositional machine learning techniques on outputs from GENESYS simulations.

To learn co-existence rules for pairs of fields, three different output variables were analysed:

- the proportion of pollen dispersed from a donating to a receiving field,
- the same for seeds,
- the proportion of GM seeds in non-GM oilseed rape harvests (hence harvest contamination).

4.1 Predicting the proportion of pollen/seed dispersal

In each simulation, the field plan was limited to two individual field. The simulations were obtained by taking all possible combinations of the following properties: (1) the distance between the plots (0, 10, 50, 100, 500, 1000, 1500, 2000 or 3000 m); (2) their areas (9, 100, 961 or 10000 m^2); (3) their shapes (square, linear with 1-m-width or intermediate with length equal to three times the width); and (4) the orientation of the two plots (parallel or perpendicular). In total, there were $9 \cdot 4^2 \cdot 3^2 \cdot 2 = $2592 couples of plots tested. For each of these couples, the proportions of pollen and seeds from the first plot to itself, from the first plot to the second plot, from the second plot to itself and from the second plot to the first plot were simulated and analysed, resulting into $4 \cdot 2592 = $10368 situations (examples).

To summarize, the field descriptors used as attributes in the analysis were the following:

- DispersalType (type of dispersal, i.e., identical vs. distinct donating and receiving fields),
- Distance (between fields),
- Orientation (of the fields related to each other),

- Area_D (area of donor field),
- Area_R (area of receiver field),
- RatioLW_D (shape: ratio length to width of donor),
- RatioLW_R (idem for receiver),
- Circumference_D (circumference of donor field),
- Circumference_R (circumference of receiver field).

For each of the output variables, a model tree was fitted. In all the cases, the correlation coefficient r^2 of the regression trees, obtained with 10-fold cross-validation, was extremely large (0.99). The structures of the pollen and seed dispersal trees were similar (Fig. 3).

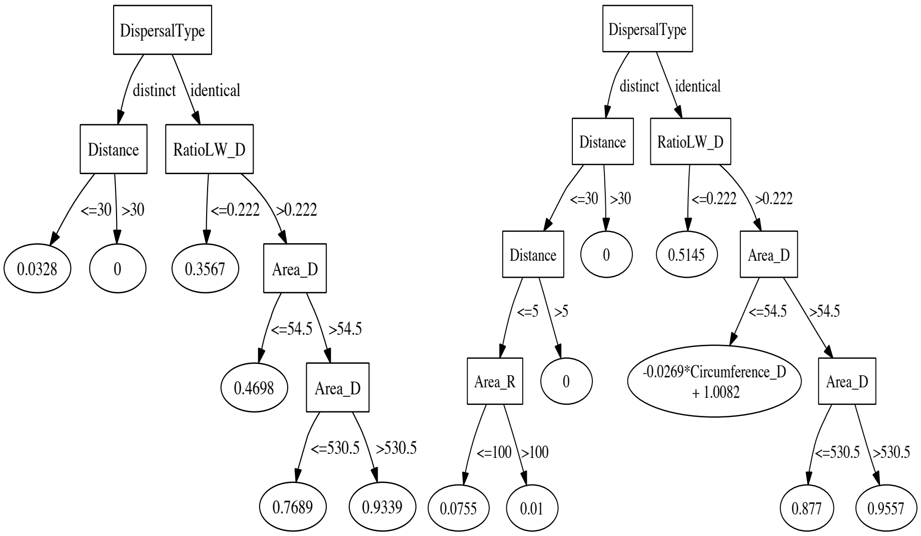

Fig. 3. Model trees for predicting pollen (left) and seed (right) dispersal (proportions dispersed from a donating field to a mean m^2 of a receiving field). Explicative variables are distance between fields (in m), length/width ratio (RatioLW_D) and circumference (Circumference_D, in m) of donating field, areas of donating (Area_D, in m^2) and receiving fields (Area_R, in m^2), and dispersal type (identical vs. distinct donating and receiving fields).

The main factor explaining the proportion of immigrating pollen was the "type" of dispersal, i.e. dispersal was larger for pollen movement from a plot to itself than from a plot to a distinct neighbor plot (Fig. 3, left). In the case of distinct plots, the only other factor was the distance between fields: below 30 m, mean pollen dispersal to a mean m^2 of the receiving field was 0.03 of the production of the donating field; above 30 m, it was nil. In case of seed dispersal to neighbor fields (Fig. 3, right), the distance threshold was 5 m, and the area of the receiving plot also had an effect: the larger this area, the smaller the dispersal because the incoming seeds were distributed over a larger reception area.

In case of self-dispersal, i.e. dispersal from a plot to itself, the shape of the field was most important: self-dispersal was lower for rectangular (low width/length ratio) vs. square plots. In the case of "squarer" plots, both pollen and seed dispersal increased with field area.

4.2 Predicting the adventitious presence of GM seeds

For predicting the third output variable (harvest contamination) the same 2592 plot couples were used to simulate harvest contamination, but additional input variables were necessary, comprising cropping systems, initial seed bank and the characteristics of the oilseed rape varieties. Each simulation covers a period of 25 years, which is far longer than the time during which the initial seed bank influences harvest contamination [9]. As only the last year was used for analysis, we were able to ignore the effect of the initial seed bank present at the onset of the simulation and started all simulations with an empty seed bank. Six contrasted cropping systems were identified by Colbach et al. (2004). They comprised two high-risk systems (with frequent GM rape), two intermediate systems and two low-risk systems (with GM rape only every 10 or 25 years). The varieties used in these systems were high-risk genotypes (low self-pollination of non-GM plants, large pollen emission of GM plants, etc.) or low-risk genotypes (high non-GM self-pollination, low GM pollen emission etc.). The remaining cultivation techniques were also chosen according to these contrasted cropping systems.

For each of the 6 cropping systems and 2592 plot couples, 7 repetitions were simulated, resulting into 108864 simulations (examples). The 7 repetitions resulted from starting each time with a different crop from the 7-year rotation (e.g. rape/winter wheat/spring barley/set-aside/rape/winter wheat/spring barley) simulated in the plot couples.

So in this analysis we used the following attributes:

- CroppingSystem (6 options)
- FirstCrop (7 options)
- Distance (between fields)
- Orientation (of the fields related to each other)
- Area_D (area of donor field)
- Area_R (area of receiver field)
- Circumference_D (circumference of donor field)
- Circumference_R (circumference of receiver field)

The structure of the regression tree for harvest contamination (Fig. 4) was very different from the structures of the trees for predicting pollen and seed dispersal. The main factor was the effect of cropping system. In case of the maximum-risk systems as well as intermediate and low-risk systems, field characteristics had no influence at all. Only in the case of the high-risk systems was there any effect of field characteristics, which were similar to those observed for pollen and seed dispersal: harvest contamination increased with the area of the gene-donating field and was more important in case of rectangular vs. square donating fields.

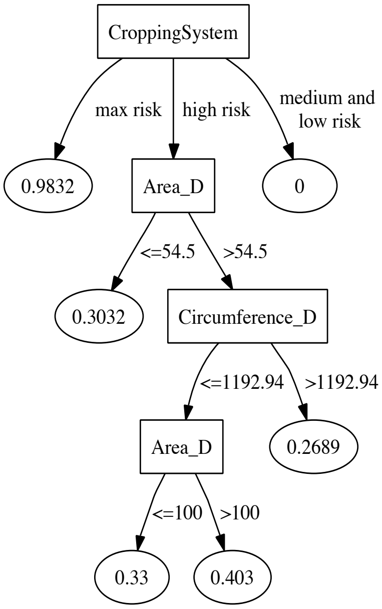

Fig. 4. A regression tree for predicting harvest contamination (rate of adventitious presence of GM seeds in non-GM oilseed rape harvests). Explicative variables are area of donating field (Area_D, in m^2), cropping system (labeled as max risk, high risk, medium and low risk), and circumference of donating field (Circumference_D, in m).

The identified effects of field characteristics were consistent with previous sensitivity analyses [10] and the knowledge on dispersal mechanisms: dispersal decreases with distance from the pollen or seed source, large areas emit more pollen or seeds and "dilute" the incoming material, rectangular plots emit more material because most of their surface is close to a neighbor field etc. In contrast to the previous study, the present work improves the knowledge on interactions, e.g. that the shape of fields is most important for small fields. The most interesting result was the interaction between cropping systems and field characteristics, showing that the latter were only important in certain situations such as the high-risk system in the present study. This system comprised frequent GM and non-GM oilseed rape crops both in space and in time and, most importantly, non-GM varieties with low self-pollination rates (50%) [9]. In case of omnipresence (maximal-risk cropping system) or low frequency of GM pollen and seeds (intermediate and low-risk systems), field plan characteristics present a negligible effect.

The results of this section confirmed the overall importance of the cropping systems, overriding most of the field plan effects. To obtain satisfactory simulations with GENESYS, it is thus most important to concentrate on gathering input data on cropping system while errors on field coordinates should have less impact.

5 Learning co-existence rules for a large-risk field plan

The aim of this analysis was to estimate how the properties of the farming region and the cropping system influence the rate of contamination of non-GM crops with GM seeds. In this part of the analysis, the focus was not on predicting gene flow and contamination between pairs of fields, but on predicting the rate of adventitious presence of GM seeds in the central field of a large-risk field pattern (Figure 5).

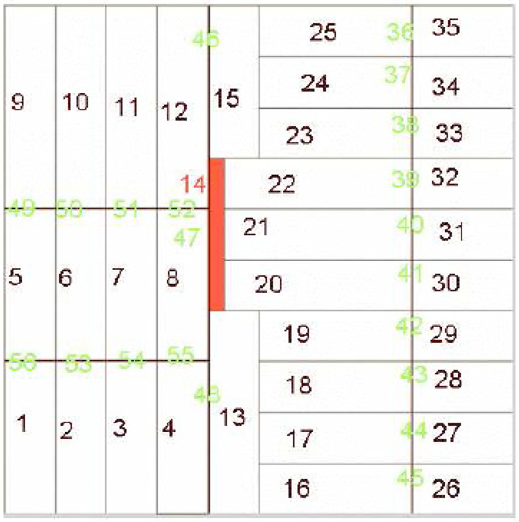

Fig. 5. Large-risk field plan. Out-crossing rate for the central field (dark-shadowed field with number 14) was predicted. Neighbor fields are numbered from 1 to 13 and 15 to 35. (Borders are numbered from 36 to 56 and are small grass strips between cultivated fields, but in our analysis only the large-risk field plan without borders was used.) [10]

The large-risk field plan consists of a small and rectangular central field (field number 14) surrounded by large neighbor fields, a combination which maximises pollen and seed input into the central field. The dataset used in this analysis was based on previous sensitivity analyses of GENESYS to field patterns [7, 8, 10]. Each simulation starts with an empty soil seedbank and covers a period of 25 years. Each year, the crops and the management techniques for crops were chosen randomly, as well as the genetic variables describing the oilseed rape varieties. The only exception was the crop grown during the 25^{th} year in the central field which was always non-GM oilseed rape. Our target variable was the rate of harvest contamination (adventitious presence of GM seeds) in this crop. 100000 simulations of crop rotation on the large-risk field plan without borders were performed. Of the 25 simulated years of each simulation, full details were kept only for the last 4 years.

The dataset produced by GENESYS was analyzed using relational decision trees. Our assumption was that the contamination of a field with GM seeds depends a lot on the cropping techniques and crops grown on the surrounding fields (e.g., the level of contamination of a field may be influenced by the crop grown at or the level of contamination of its neighboring fields). So it seems worthwhile to exploit neighborhood relations in the predictive model and create a relational representation of the problem. Also, the probability of contamination might increase if the field plan contains a lot of contaminated fields. Therefore it would be useful to investigate properties at the regional level, which can be obtained by aggregating over the individual fields. For this study we used Tilde [1], a system that builds relational decision trees.

According to the previous analysis of factors for presence and abundance of GM oilseed rape [14], a field is most likely to be contaminated if GM oilseed rape had been grown in the same field previously. Having this in mind, we filtered the dataset originally consisting of 100000 examples, excluding the examples in which there was GM oilseed rape grown on the target field in the last four years. The reason to do this was to avoid generating very obvious rules (for example: if there was GM oilseed rape on the target field in the last four years, the probability that it will now be contaminated is almost 100%) and try to see what is the role of the neighboring fields. At the end the dataset consisted of 64877 examples.

We used the following relational representation of the data. The target relation was *contamination(SimID,RateAdvPres)*, where *RateAdvPres* is the target variable, denoting the rate of adventitious presence of GM varieties of the non-GM central target field and *SimID* is the number (from 1 to 100000) of the simulation.

The background relations were related to the cultivation techniques, the year that oilseed rape was last planted at a given field, and the geometry of the field plan. A first relation is *targetField(SimID,FieldID)*, denoting that *FieldID* is the target field of the field plan. In this analysis, *FieldID* always refers to field 14 (see Figure 5), although the applied method allows to vary the target field per example. In the relation *fieldDataYear(SimID,FieldID,Year,CultivationTechniques)*, *CultivationTechniques* is a list of variables describing the cropping techniques. Here we use only crop and sowing date and ignore the other cropping techniques, like tillage, sowing density, efficiency for herbicides on non-GM/GM volunteers, $1^{st}/2^{nd}$ cutting, harvest loss and grazing. *Year* takes values from [0, 1, 2, 3], 0 denoting the present year and 3 - three years ago. In the relation *lastOSR(SimID,FieldID,LastGM,LastNonGM)*, *LastGM* is the number of years ago [1..25] in which GM oilseed rape was last grown on *FieldID*, and *LastNonGM* is the number of years ago in which non-GM oilseed rape was last grown on *FieldID*.

The relation *neighbor(SimID,Field1ID,Field2ID,NeighType)* holds if the minimum distance between Field1 and Field2 is zero. If they have a common edge of non-zero length, *NeighType* is *adjacent*, and if they have only one point in common (touching with only one corner), then *NeighType* is *corner*. Additional information on the area of fields, their mutual distances (average and minimal), and length of the common edges was available, but was not used in our analyses.

Experiments and results. For the experiments we discretized the target attribute, in order to obtain a classification problem. If the rate of harvest contamination exceeds 0.9%, an EU labeling threshold, the target field is considered contaminated, otherwise not.

Given the size of the dataset, we used a sampling strategy to build the tree: at each node only 10000 examples are used to evaluate the tests and select the best test. Afterwards, the whole dataset is split according to this best test. The minimum number of examples a leaf has to cover was set to 600, and a random proportion of 20% of the data was set aside as a validation set for pruning.

We tried the following experimental settings:

- *Propositional*: besides the target relation *contamination(SimID,RateAdvPres)*, only (propositional) data for the target field is included (not using any relations among the fields), i.e., the following predicates are used:
  - *fieldDataYear(SimID,FieldID,Year,Crop,SowingDate)*, for the target field
  - *lastOSR(SimID,FieldID,LastGM,LastNonGM)*, for the target field
- *Neighbor*: the same relations were used as in the *Propositional* setting, but now other fields are introduced via the *neighbor* relation, starting at the target field:
  - *neighbor(SimID,Field1ID,Field2ID,NeighType)*

  Note that the information on the neighboring fields from the relations *fieldDataYear* and *lastOSR* can also be used.

For each of these settings, we report the tree size (number of nodes) and the predictive performance in Table 1. The accuracy was measured by three-fold (and not 10-fold) cross-validation due to high computational complexity resulting from the large size of the dataset.

	Propositional	Neighbor
tree size	15	13
accuracy	78.35%	79.66%

Table 1. Tilde's experimental results.

In addition, we give examples of rules obtained in each of the experimental settings tried. The following rule is an example from the *Propositional* experiments:

```
contamination(S,neg):-targetfield(S,T), fieldDataYear(S,T,0,Crop,SowingDate), SowingDate<252,
                                                    yearsSinceOSR(S,T,Gm,NonGm), Gm>5, !.
```

The above rule states that the target field will be predicted as not contaminated, if the sowing date in the present year is before the 252nd day of the year (9 September) and the last GM oilseed rape grown on it was more than 5 years ago.

The relational model contains 4 nodes referring to neighboring fields. The next rule is an example from the relational model (*Neighbor* experiments) which uses information about a neighboring field:

```
contamination(S,pos):-targetField(S,T), fieldDataYear(S,T,0,Crop,SowingDate), SowingDate<252,
                    neighbor(S,T,FieldA,adjacent), fieldDataYear(S,FieldA,1,gm-OSR,SowingDate),!.
```

This rule can be interpreted as follows: if the sowing date of the target field in the present year is before the 252^{nd} day of the year (9 September) and it has a neighboring field (FieldA) with which it is adjacent, and the neighboring field had GM-OSR last year, then the target field is predicted to be contaminated.

The results from the analysis in every experimental setting showed that the most important attribute for determining the contamination of the target field is the sowing date as was also shown in [16]. The later the sowing date the lower the contamination, because the GM volunteers can appear and be destroyed prior to the sowing of the non-GM oilseed rape, thus decreasing the possibility of its contamination with GM material. Another important factor that influences the contamination of one field is the crop grown on its neighboring fields. If GM crops are grown in the nearest neighborhood of the target field, then it is very probable that it will be contaminated. Also, as said previously, if the target field had GM crops grown on it in the past years, then it is almost certain that it will be contaminated.

From the results and the comparison of accuracies of the relational to the propositional experiments we have noticed that the former provided only a small improvement in accuracy (1%). However, this study is only a first step in using the relational data mining methods for analysis of outputs of complex simulation models. Exploring the possibility of varying the field plans and target fields within them might use the advantages of relational methods in their full and result in higher improvement in accuracy.

6 Conclusions and further work

In this paper, we have studied the use of machine learning for analyzing the output of complex simulation models in the context of understanding the co-existence of GM and non-GM crops and the conditions for its feasibility. The outputs of the GENESYS model, designed to study the co-existence on conventional and GM oilseed rape crops, were analysed by using the machine learning method of decision tree induction. Co-existence and adventitious presence of GM material were studied in several contexts, including gene flow between pairs of fields, the interactions of this process with farming practices (cropping systems), and gene flow in the context of an entire field plan. The results of our study confirmed that machine learning is a powerful tool for learning co-existence rules for GM and non-GM crops from the output of complex simulation models.

We first used machine learning to learn co-existence rules for pairs of fields, predicting the pollen and seed dispersal, as well as the proportion of GM seeds in non-GM oilseed rape harvests (harvest contamination). For each of the three target attributes, very accurate model and regression trees were built (with cross-validated correlation coefficients of 0.99). The trees were also simple enough to be inspected and understood. The results showed that the identified effects of field characteristics were consistent with previous sensitivity analyses [10] and the knowledge on dispersal mechanisms: dispersal decreases with distance from the pollen or seed source, large areas emit more pollen or seeds and "dilute" the incoming material, rectangular plots emit more material because most of their surface is close to a neighbor field etc. In contrast to the previous study, our analysis improves the knowledge on interactions between individual influencing factors, e.g., that the shape of fields is most important for small fields. The most interesting result was the interaction between cropping systems and field characteristics, showing that the latter were only important in certain situations such as the high-risk system in the present study. This analysis confirmed the overall importance of the cropping systems, overriding most of the field plan effects. To obtain satisfactory simulations with GENESYS, it is thus most important to concentrate on gathering input data on cropping system while errors on field coordinates should have less impact.

We then used machine learning to analyze the influence of the farming region and various cropping systems on the contamination of non-GM crops with GM material. For the purpose of this analysis, we used a large-risk field plan and learned to predict the contamination of the central field based on the cropping systems and farming practices of the central field and its neighbors. Given the relational nature of this problem (namely, the relations to neighboring fields in the field plan are expected to play an important role), we used the relational decision tree learning system TILDE. The actual target variable that we predicted was whether the level of adventitious presence exceeds the 0.9% threshold set by EU regulations. We also constructed classical (propositional) decision trees which only used the properties of the central field. The cross-validated classification accuracies reached around 80%, with the relational approach achieveing a higher (albeit only by 1%) accuracy. The learned model also clearly made use of the relational aspects, referring to the properties of and farming practices applied to the neighboring fields of the target (central) field.

While data analysis and machine learning methods had previously been used to analyze the output of simulation models for studying the co-existence of GM and non-GM crops, the use of relational learning methods is a novelty and a unique contribution of our study. The relational learning methods allow us to use the relational aspects, both spatial and temporal, of the information concerning the field plan and farming practices applied to the field in it. In fact, these methods would allow us to vary the target field within a chosen field plan, as well as consider completely different field plans at the same time, and thus obtain more generally valid co-existence rules. This is a unique advantage as compared to the data analysis method applied so far to the problem at hand.

The most natural direction for further work would be to use a larger amount of simulation data that would exploit the advantages of the relational learning methods. This would mean running GENESYS simulations with different field plans, as well as with different target fields within each field plan. In this way, we would exploit the relational capability of the learning methods better and obtain more accurate and more general co-existence rules. Another direction for further work would be to use the same general approach of using machine learning, and the more specific approach of using relational learning, to analyse the simulation results of other models designed to study the co-existence of GM and non-GM crops. Finally, the general methodology we propose would be applicable to the analysis of results of simulation models in other areas of ecology.

Acknowledgements

The authors acknowledge the support of the specific targeted research project SIGMEA (Sustainable Introduction of Genetically-Modified organisms into European Agriculture) funded by the European Commission under contract 502981. The authors wish to thank Antoine Messéan(INRA, Grignon, France) and Hendrik Blockeel (Department of Computer Science, K.U.Leuven, Belgium) for their help and useful discussions and Damjan Demšar (Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia) for obtaining the regression trees in WEKA for the field-to-field experiments.

References

1. Blockeel, H., De Raedt, L.: Top-down induction of first order logical decision trees. Artificial Intelligence **101**(1-2) (1998) 285–297
2. Bratko, I.: Prolog Programming for Artificial Intelligence, 3rd edition. Addison-Wesley, Harlow, England (2001)
3. Breiman, L., Freidman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Ed: Wadsworth & Brooks, Monterey, California (1984)
4. Blockeel, H., Dehaspe, L., Ramon, J., Struyf, J., Van Assche, A., Vens, C., Fierens, D.: The ACE data mining system: User's manual. http://www.cs.kuleuven.be/ dtai/ACE (2006)
5. Champolivier, J.: Etude de l'impact de colzas résistans aux herbicides dans les systèmes de culture. 1. année d'expérimentation. Synthèse des essais inter-instituts-Campagne 1995-96, CETIOM, 22p.
6. Colbach, N., Meynard, J.M., Clermont-Dauphin, C., Messéan, A.: GeneSys: A model of the effects of cropping system on gene flow from transgenic rapeseed. Gene Flow and Agriculture - Relevance for transgenic crops, Keele University (UK) (1999), 89–96
7. Colbach, N., Clermont-Dauphin, C., Meynard, J.M.: GENESYS: A model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. I. Temporal evolution of a population of rapeseed volunteers in a field. Agriculture, Ecosystems and Environment **83** (2001) 235–253
8. Colbach, N., Clermont-Dauphin, C., Meynard, J.M.: GENESYS: a model of the influence of cropping system on gene escape from herbicide tolerant rapeseed crops to rape volunteers. II. Genetic exchanges among volunteer and cropped populations in a small region. Agriculture, Ecosystems and Environment **83** (2001) 255–270
9. Colbach, N., Molinari, N., Clermont-Dauphin, C.: Sensitivity analyses for a model simulating demography and genotype evolutions with time. Application to GENESYS modelling gene flow between rapeseed varieties and volunteers. Ecological Modelling **179** (2004) 91–113
10. Colbach, N., Molinari, N., Meynard, J. M., Messéan, A.: Integrating spatial aspects into sensitivity analyses for models simulating demography and genotype evolutions with time. Application to GENESYS modelling gene flow between rapeseed varieties and volunteers. Agronomy for Sustainable Development **25** (2005) 355–368

11. Colbach, N., Fargue, A., Sausse, C., Angevin, F.: Evaluation and use of a spatio-temporel model of cropping system effects on gene flow. Example of the GeneSys model applied to three co-existing herbicide tolerance transgenes. European Journal of Agronomy **22** (2005), 417–440
12. De Raedt, L.: Attribute-value learning versus inductive logic programming: the missing links (extended abstract). In D. Page (Ed.), Proceedings of the Eighth International Conference on Inductive Logic Programming, Volume 1446 of Lecture Notes in Artificial Intelligence, pp. 18. Springer-Verlag.
13. De Raedt, L., Blockeel, H., Dehaspe, L., Van Laer, W.: Three Companions for Data Mining in First Order Logic. In [15] (2001) 105–139
14. Debeljak, M., Squire, G., Demšar, D., Young, M., Džeroski, S.: Data mining methods reveal soil-related and community-dependent factors in the presence and abundance of weedy oilseed rape before GM crop trials. Ecological Modelling (2006) (in press)
15. Džeroski, S., Lavrač, N.: Relational Data Mining. Springer, Berlin (2001)
16. Ivanovska, A., Panov, P., Colbach, N., Debeljak, M., Džeroski, S., Messean, A.: Using simulation models and data mining to study co-existence of GM/non-GM crops at regional level. In Proceedings of 20^{th} International Conference on Informatics for Environmental Protection (EnviroInfo) (2006) 489–500, Graz, Austria
17. Kramer, S.: Structural regression trees. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (1996) 812–819, MIT Press, Cambridge, MA
18. Kramer, S., Widmer, G.: Inducing Classification and Regression Trees in First Order Logic. In [15] (2001) 140–159
19. Messéan, A., Angevin, F., Gomez-Barbero, M., Menrad, K., Rodriguez-Cerezo, E.: New case studies on the coexistence of GM and non-GM crops in European agriculture. Technical Report Series of the Joint Research Center of the European Commission, EUR 22102 EN (2006)
20. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., Redwood City, CA. (1993)
21. Van Assche, A., Vens, C., Blockeel, H., Džeroski, S.: First order random forests: Learning relational classifiers with complex aggregates. Machine Learning (2006) Accepted.
22. Wang, Y., Witten, I.H.: Induction of model trees for predicting continuous classes. In Proc. Poster Papers Europ. Conf. Machine Learning. Prague: University of Economics, Faculty of Informatics and Statistics. (1997)
23. Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (1999)