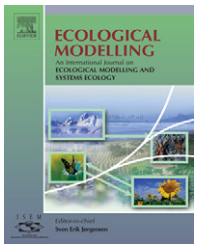
ECOLOGICAL MODELLING 212 (2008) 147–154

available at www.sciencedirect.com

ScienceDirect

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



Characterizing the presence of oilseed rape feral populations on field margins using machine learning

Sandrine Pivard[a,*], *Damjan Demšar*[b], *Jane Lecomte*[a], *Marko Debeljak*[b], *Sašo Džeroski*[b]

[a] *Univ Paris-Sud, CNRS, AgroParisTech, Laboratoire Ecologie, Systématique et Evolution, UMR 8079, Orsay F-91405, France*

[b] *Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia*

ARTICLE INFO

Keywords:
Oilseed rape
Feral population
Risk assessment
Data mining
Attribute ranking
Classification tree

ABSTRACT

Many cultivated species, such as oilseed rape, sunflower, wheat or sorghum can escape from crops, and colonize field margins as feral populations. The general processes leading to the escape and persistence of cultivated species on field margins are still poorly investigated. An exhaustive 4-year survey was conducted in the centre of France at a landscape level to study the origin of feral oilseed rape populations. We present here results obtained with machine learning methods, which are increasingly popular techniques for analysing large ecological datasets. As expected, the dynamics of feral populations relies on large seed immigration from fields and transport. However, the seed bank was shown to be the keystone of their persistence rather than local recruitment.

1. Introduction

Most agricultural landscapes are a mosaic of fields, semi-natural habitats, human infrastructures (e.g. roads), and occasional natural habitats (Marshall and Moonen, 2002). Within intensively-managed farming systems, fields margins represent most of the semi-natural habitats and are thus a key feature of agricultural landscapes, present in some form at the edges of all agricultural fields (Burel et al., 1998). Major changes in agriculture, such as production intensification and land re-allotment programmes, have triggered recent changes in field margin features, leading in dramatic declines in biodiversity (Kleijn et al., 1997; Freemark et al., 2002). Therefore, they become remnant refugia for farmland biodiversity in virtual wildlife deserts (Kleijn and Verbeek, 2000).

The high frequency of disturbances, combined with the high fertility of the habitat, enhances the growth of annuals and ruderal perennial species (Wilson and Tilman, 1991). Thus, many cultivated species, such as oilseed rape, sunflower, wheat or sorghum can escape from crops, colonize field margins as feral populations and even manage to persist there (Hancock et al., 1996; Stewart et al., 2003; Clements et al., 2004). While research effort is in progress to understand the ecology of field margins at a range of spatial scales (Marshall and Moonen, 2002), the general processes leading to the escape and persistence of cultivated species on field margins are still poorly investigated (Garnier and Lecomte, 2006).

To investigate theses questions, we chose oilseed rape (*Brassica napus* L.) as a model. Indeed, oilseed rape feral populations are common features of field margins in agricultural landscapes (Charters et al., 1999; Pessel et al., 2001). Some studies suggest that oilseed rape behaves as an early successional ruderal, incapable of regenerating in undisturbed habitats (Crawley et al., 1993, 2001). Feral populations are thus

* *Corresponding author at*: Univ Paris-Sud, Laboratoire Ecologie, Systématique et Evolution, UMR 8079, Orsay F-91405, France. Tel.: +33 1 69 15 76 57; fax: +33 1 69 15 46 97.
E-mail address: sandrine.pivard@polytechnique.org (S. Pivard).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.10.012

often considered as resulting mainly from seed immigration: spillage from farm machinery or from neighbouring fields cultivated the current or previous year (Crawley and Brown, 1995; Lutman, 2003). However, there is also evidence that feral populations can persist many years via local recruitment (local production of new individuals) or via a seed bank (Charters et al., 1999; Pessel et al., 2001).

In order to study the origin of feral oilseed rape populations, a survey was conducted from 2000 to 2003 in the centre of France at a landscape level in a farmland area. First analyses showed that the feral populations did not rely only on seed immigration from neighbouring crops and seed transport, but also managed to persist, through local recruitment or seed bank or even frequent seed spillage over particularly favourable habitats (Pivard et al., in press). Information on the presence of feral plants over time did not reveal in itself the processes involved in population persistence (Charters et al., 1999). To explore more deeply these processes, we needed to use machine learning methods. Notably, attribute ranking methods (*Information gain* and *ReliefF*) were used on the collected data set to estimate the relative importance of 23 explanatory variables. The latter concerned different types of oilseed rape presence in the past (cultivated fields, feral populations), including information about the period of observation and the possible production of seeds, and permanent features characterising the place (road type, vicinity to a junction, vicinity to a village). Classification trees were then built to predict the feral population presence in 2003 from these variables.

Ecosystems characteristically exhibit highly complex non-linear relationships between their associated variables. Machine learning methods offer some advantages over traditional statistical analysis techniques when modelling such relationships because they introduce less (if any) prior assumptions about the relationships between the variables. Machine learning methods are thus increasingly popular techniques for analysing ecological datasets (Lek and Guegan, 1999; Debeljak et al., 2001; Recknagel, 2001; Dzeroski and Todorovski, 2003). For example, where a classical statistical approach would assume independence of the attributes, some machine learning methods can explicitly explore the dependencies among attributes. Another example is the use of linear regression to model the relationships in an ecosystem (with the associated distributional assumptions), whereas a typical machine learning approach would be to use model trees, which would approximate a non-linear dependency through a tree-structured set of piece-wise linear dependencies. Advantages, disadvantages and complementarities of machine learning and statistical methods are reviewed by Dzeroski (2001).

2. Materials and methods

2.1. Study area and survey

The study area was a 41 km^2 production area of winter oilseed rape centred on the grain silo of Selommes (Loir-et-Cher, centre of France). Twice a year, from 2000 to 2003, we monitored the GPS coordinates of all cultivated oilseed rape fields and feral populations present on road verges along a 110 km-long road network. Census was made from a van driving at maximum 15 km/h in early April during oilseed rape flowering and in early July, before harvesting. As it was impossible to do an individual plant survey of feral plants, we considered in the same set of GPS coordinates all the plants separated by less than 10 m, so called "populations".

The study area took place in an intensively managed reallotted farmland. As margins between fields were trifling, we observed field margins only along roads. Conversely, 97% of the road verges were field margins. We thus confounded here field margins and road verges. Each year, 10–14% of the road verges were bordered by oilseed rape crops, and approximately the same percentage were occupied by feral

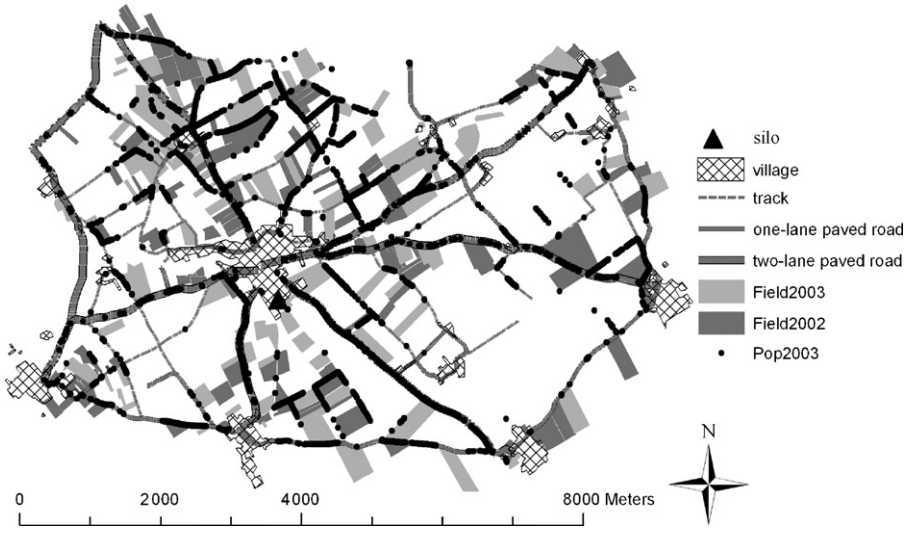


Fig. 1 – GIS map of the feral populations monitored in 2003 (black points) and rape fields monitored in 2002 (dark-grey polygons) and 2003 (light-grey polygons). The road network is composed of tracks (grey lines), one-lane (thin continuous lines) and two-lanes (large continuous lines) paved roads. The two-lane roads are mainly directed toward the silo (black triangle).

populations (Pivard et al., in press). The road network was composed of tracks (42% of the roads), one-lane (29%), and two-lane (29%) paved roads.

The original data were collected in a database created with PostgreSQL DBMS, coupled with ArcView GIS software, used for data visualisation (Fig. 1).

2.2. Data set

To create the data set, we used the roads as the reference system. We divided them in 3 m-long oriented segments, distinguishing the two roadsides. We thus obtained 74,002 segments, designated as the spatial unit. GPS coordinates of fields and feral populations collected between 2000 and 2003 were projected onto these segments, together with their attributes. The response variable Pop03, i.e. the presence of a feral population in 2003, took for each segment the value 1 if a feral population was projected on it or 0 otherwise. We considered 23 independent variables, including historical presence of oilseed rape (cultivated fields for each of 2000–2003, feral populations 2000–2002) and elements of landscape heterogeneity, whose purpose was to investigate extra sources of variability, such as seed transport intensity or weed management:

- The road type (*Rdtype*) took three values: track, one-lane or two-lane paved road. Seed spillage frequency and intensity was assumed to differ on these different types of road.
- The vicinity to a junction (*Junction*) was a binary variable taking the value 1 if the segment was less than 10 m from a junction, which was true for 6% of the segments. Crawley and Brown (1995) demonstrated that seed spillage was more important there.
- The presence of a village (*Village*) was a binary variable set to 1 if the segment was in a village. This was true for 2.5% of the segments.

To take into account the natural seed shedding from field or feral populations, a neighbourhood for each segment was defined, composed of the segment itself and its two adjacent segments on the same roadside. The 20 remaining variables described the history of the segment and of its neighbourhood:

- *Field03*, *Field02*, *Field01 and Field00* were binary variables standing for the presence of an adjacent rape field, i.e. in the neighbourhood of the segment in 2003, 2002, 2001 and 2000, respectively. Fields could be a source of seeds during sowing or harvesting or via natural seeds shedding. They could be responsible for the presence of a feral population in 2003 directly or via a seed bank. *OField03*, *OField02*, *OField01*, *OField00* were similar variables, except that they described the presence of a rape field in the neighbourhood of the opposite segment.
- *PopApr02*, *PopApr01*, *PopApr00* were binary variables standing for the presence of adjacent feral populations in April, at flowering period. *OPopApr02*, *OPopApr01*, *OPopApr00* were similar variables, except that they represented the presence of a feral population in the neighbourhood of the opposite segment.
- *PopJul02*, *PopJul01*, *PopJul00* stood for the observation of adjacent feral populations, in July, at crop harvest. They took tree values: 0 if there is no population, 2 if the population produced seeds, 1 if not. *OPopJul02*, *OPopJul01*, *OPopJul00* were similar variables, in the neighbourhood of the opposite segment. The presence of past feral populations could be responsible for the presence of a feral population in 2003 via direct local recruitment or constitution of a seed bank.

2.3. Analysis and methods

We considered here the machine learning task of predicting the probability of presence of a feral population in 2003 (called from now a class attribute, with two classes: 1, present or 0, absent) from the 23 explanatory variables (called attributes). We first applied two machine learning methods for attribute (feature) ranking. We then applied classification tree induction to construct a model capable of predicting new feral populations.

To evaluate the predictive performance of the methods, we used n-fold cross-validation (with $n = 10$), which is known to be one of the most acceptable methods to evaluate the performance of machine learning algorithms (Stone, 1977): the whole dataset is split randomly (if possible so that the class and attribute distribution is preserved) into n-folds. Then the algorithm is run n times, each time using one of the folds as the testing set and the rest $n-1$ folds as the training dataset. The n accuracies average gives an estimated accuracy of the tree learned from the entire dataset on new (unseen) data.

Attribute ranking (or feature ranking) using the measures of *Information gain* (Hunt et al., 1966) and *ReliefF* (Kononenko, 1994) as implemented in the software package WEKA (Witten and Frank, 1999) was used to rank the 23 attributes. For attribute ranking, the average merit of the attributes on the 10 folds was reported and the attributes were ordered by average rank.

- *Information gain* estimates the difference between the prior entropy of the class variable in the entire dataset and the residual entropy after the dataset is split by using one attribute. Since the prior entropy is the same, regardless of the attribute tested, the attributes are thus ranked according to the residual entropy (the smaller the better). Residual entropy can be calculated as:$\mathrm{Ires}(A) = -\sum_v p_v \sum_c p(c|v) \log p(c|v)$, where v are the values of the attribute A, c is the value of the class attribute, and p_v is the probability that attribute A has value v. *Information gain*, the difference between the prior and residual entropy of the class, is quickly and easily calculated and is used in a number of machine learning methods, especially in methods for decision tree induction. It assumes that all attributes are independent.
- The *ReliefF* algorithms, on the other hand, are general and successful measures for evaluating attribute quality and are especially good in detecting conditional dependencies (Robnik-Sikonja and Kononenko, 2003). The key idea is to evaluate the partitioning power of attributes according to how well their values distinguish between similar observations. An attribute is given a high score if its values separate observations with different prediction values and do not

separate observations with similar prediction values. For that purpose it takes an observation (here a segment) and it searches for several sets of nearest neighbours (segments with similar values of attributes): one set of observations with the same class as the selected class (for instance Pop03 = 1), and one set for every other possible class (in our case, only Pop03 = 0). Mathematically, the attribute quality evaluation is an approximation of the difference of these probabilities: p(different value of A| nearest observations from different class) − p(different value of A| nearest observations from the same class). In other words, the first term rewards the attribute A if it separates similar observations with different prediction values and the second term punishes it if it separates similar observations with similar prediction values (Robnik-Sikonja and Kononenko, 2003).

These two probabilities contain the additional condition that the observations are close in the problem space and form an estimate of how well the values of the attribute distinguish between the observations that are near to each other. The assigned quality evaluations are in the range [−1,1], however, values below zero are only assigned to completely irrelevant (random) attributes. The quality evaluation of attribute A assigned by *ReliefF* can also be interpreted as an approximation of the contribution of that attribute to the explanation of the predictions, which is better if data are abundant (Robnik-Sikonja and Kononenko, 2003).

Classification trees (Breiman et al., 1984), often called decision trees (Quinlan, 1986), predict the value of a discrete dependent variable with a finite set of values from the values of a set of attributes, which may be either continuous or discrete. The common way to induce decision trees is the top-down induction of decision trees (TDIDT, Quinlan, 1986). Tree construction proceeds recursively starting with the complete data set. At each step, the most informative attribute is selected as the root of the (sub)tree and the current training set is split into subsets according to the values of the selected attribute. Splits are selected based on information statistics (such as *Information gain*) that measure how well the split decreases the impurity (heterogeneity or variance) of the class variable values within the resulting subsets.

Tree construction normally stops when all examples in a node are of the same class. But it is possible to prune the tree to prevent the model from being over-fit to the sample data, and to reduce tree complexity. Pruning can be employed during tree construction (prepruning) or after the tree has been constructed (post-pruning). Typically, a minimum number of examples in leaves can be prescribed for prepruning and a confidence level in accuracy estimates for leaves for post-pruning. Post-pruning entails combining pairs of terminal nodes into single nodes to determine how the misclassification error rate changes as a function of tree size. The classification tree as the final result of the method can be used either for prediction of class values for new examples, or as a source of new knowledge accessible with the interpretation of the tree.

We performed classification tree induction to predict the presence/absence of feral populations in 2003 according to the 23 attributes. We used the method J48, a reimplementation of C45 (Quinlan, 1993), from WEKA. To measure impurity decrease due to a test on a given attribute, J48 uses the gain ratio, which is *Information gain* (Quinlan, 1986), normalized (divided) by the entropy of the attribute in question. We produced models until a tree was found with an acceptable compromise of size versus test error rate (the tree with the smallest error rate had 581 leaves), where error rate was estimated using 10-fold cross-validation. Once the tree has been developed (or "grown"), it can be interpreted as a set of decision rules that define the range of conditions (attributes values) best used to predict the class attribute, here the presence/absence of feral population in 2003.

Methodological remarks: note that we used feature ranking and classification trees independently, i.e., we did not use the results of feature ranking to pre-select or filter the attributes before passing the dataset to classification tree construction. This, however, is often done in practice and is known under the name of attribute or feature selection. Feature selection can greatly improve the performance of classification algorithms, especially when the dataset contains many features irrelevant to the class as the classification algorithm is sensitive to irrelevant features. In our study, all features appeared to be relevant to some extent and best trees were achieved introducing all features (see Section 3), thus feature selection was not performed. We used feature ranking additionally to the classification trees, for understanding the relative importance and contribution of the attributes.

3. Results

3.1. *Attribute rankings*

The two machine learning attribute ranking methods we used gave different attribute rankings (Fig. 2). *Information gain* ranked the presence of an adjacent rape field (on the same side of the road) in the previous year (2002) as the most important attribute. This was followed by the presence of adjacent feral populations at flowering period in the three previous years and adjacent feral populations in July 2001. The other feral populations and fields attributes were then closely overlapped. *ReliefF*, on the other hand, judged the road type to be the most important attribute. Then it showed a well-defined pattern: the method ranked first all the attributes concerning rape fields presence, followed by all the attributes describing the presence of feral populations at flowering period and it ended with all the attributes describing the presence of feral populations at harvest time. All the assigned quality evaluations were in the range [0,1], which meant that none of the attributes were completely irrelevant, although the last ones were quite unimportant.

In both rankings, the presence of an adjacent rape field in 2002 was a highly relevant attribute. Moreover, the attributes concerning oilseed rape presence on the same side of the road were always ranked before the attributes representing oilseed rape presence on the opposite roadside, the difference being larger in *ReliefF*.

3.2. *Attribute contributions according to ReliefF*

When the data are abundant, the normalised quality evaluations assigned by ReliefF are approximations of the

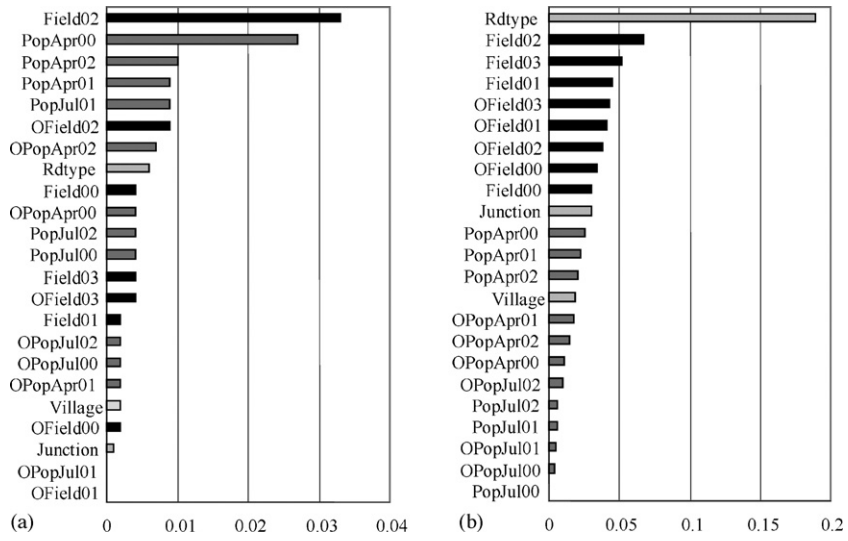


Fig. 2 – Results of the attribute evaluations obtained by 10-fold cross-validation: (a) *Information gain* attribute evaluation; (b) *ReliefF* attribute evaluation. The average merit (over the 10 folds) of an attribute is depicted by the bar length. Attributes are listed according to the average rank. As the algorithms are different, the values between figures cannot be compared, but the rankings can. To favour an easier comparison of the rankings, the field attributes are depicted with black bars, the population attributes with medium-grey bars and the permanent ones (*Rdtype*, *Village*, *Junction*) with light-grey bars.

contribution of each attribute to the explanation of the predictions. On the basis of this interpretation, the road type was responsible for 25% of the variance of the feral populations presence in 2003. Each field contributed between 4 and 9% (the highest for the presence of an adjacent field in 2002) while the adjacent feral populations (flowering and harvest periods confounded) explained 4% each.

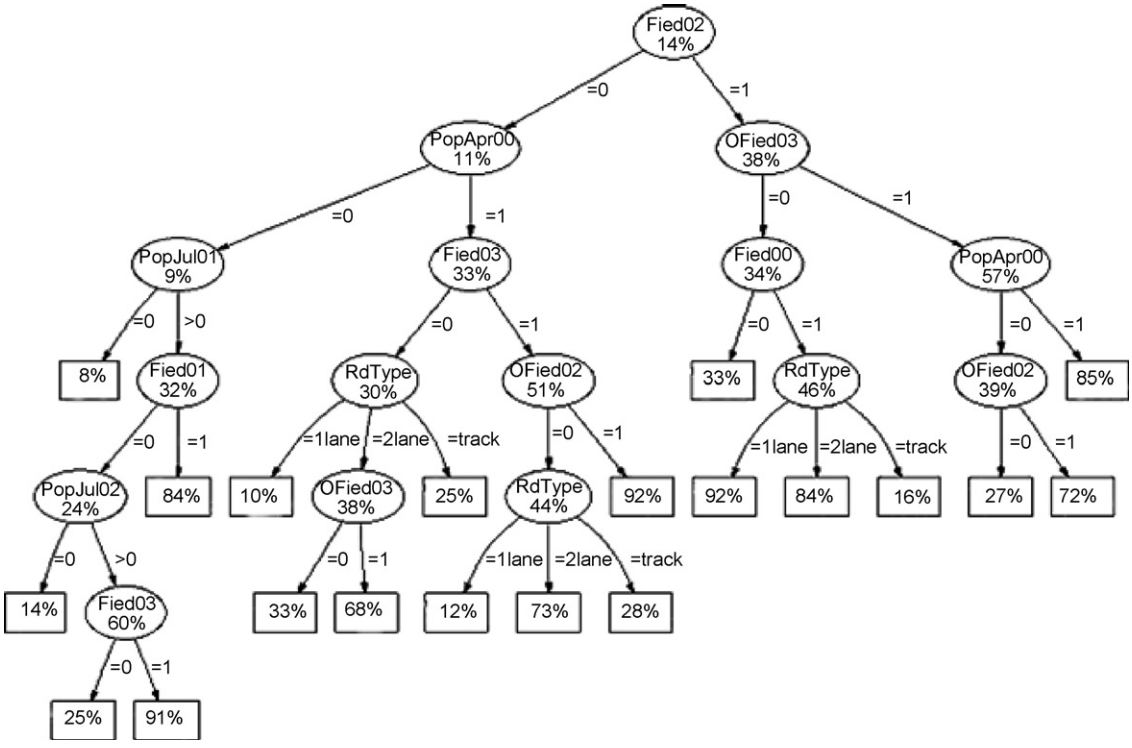


Fig. 3 – The classification tree constructed by J48, modelling the presence of oilseed rape feral population. The percentages give the predicted probability of presence of a feral population in 2003 according to the situation. For instance, this probability on the whole area is 14% (at the root); in absence of an adjacent field in 2002 (*Field02* = "0"), which is the best attribute to explain the presence of a feral population in 2003, it is only 11% (on the left); in presence of an adjacent field in 2002 (*Field02* = "1"), it increases to 38% (on the right); and so on.

3.3. Classification trees for predicting the probability of presence of a feral population

We tried constructing trees with only the best-ranked attributes (feature selection), however these trees were evaluated as worse than the presented tree, constructed using all attributes (Fig. 3). The accuracy of this pruned tree (estimated by 10-fold cross-validation) was 87.2%. The best accuracy produced by modelling was 91.0%, still using all attributes. It was significantly better but the tree, with 581 leaves, was irrelevant for presentation or even interpretation (but could be used for prediction).

The classification tree for predicting the probability of presence of a feral population in 2003 had the presence of an adjacent rape field in 2002 in the root. In node 2, thus in absence of an adjacent field in 2002, the presence of feral populations in April 2000 best split the data. On the other side of the tree, in node 3 (thus in presence of an adjacent field in 2002), the presence of an opposite field in 2003 best separated the data. In the following nodes, we found once again the presence of an adjacent feral population in April 2000, the presence of adjacent feral populations in July 2001 and 2002, the road type (three times) and the presence of adjacent fields in 2000, 2001 and 2003 and opposite fields in 2002 and 2003. We noted two things. First, in the nodes concerning feral populations at harvest, it did not matter whether the population produced seeds or not. The splits occurred only on the presence/absence of the populations. Secondly, among the attributes structuring the tree, we found the top five attributes of the two ranking methods, or attributes well ranked in both methods (except the presence of an adjacent feral population in July 2002).

4. Discussion

We estimated the importance of the 23 attributes with *Information gain* and *ReliefF*, then built classification trees for predicting the probability of presence of a feral population in 2003. The results revealed that some attributes were highlighted more often than others: the road type, the presence of adjacent fields in any year, of fields on the opposite roadside in 2002 and 2003, of adjacent feral populations, mostly at flowering period and sometimes at harvest. In contrast, some other attributes were clearly not important, notably the presence on the other side of the road of feral populations, whenever the period is, and the presence of a nearby village or junction.

4.1. Contrasting ranking methods

Looking more closely, the different ranking methods gave us contrasted results: in both rankings, the presence of an adjacent rape field in 2002 was a highly relevant variable, while the road type was judged really more important by *ReliefF*. Moreover, the presence of adjacent fields tend to take in *ReliefF* the ranks occupied by the presence of adjacent feral populations at the flowering period in *Information gain* ranking (and vice versa). Nevertheless, we kept both, judging the two methods interesting and the differences highly informative. The main difference between impurity-based methods, as *Information gain*, and *ReliefF* is the independency assumption. The first ones use correlations between the attribute and the class disregarding the context of other attributes. As for *ReliefF*, the algorithm takes into account the context of other attributes, i.e. the conditional dependencies between the attributes given the predicted value, which can be detected in the context of locality. From the global point of view, these dependencies are hidden due to the effect of averaging over all training observations, and that makes the impurity functions myopic (Robnik-Sikonja and Kononenko, 2003). The power of *ReliefF* is its ability to exploit information locally, taking the context into account, but still to provide a global view.

4.2. Dependency between attributes

Thus, the lack of independency between some attributes, in particular the ones concerning the presence of feral populations in the past, could have been a reason why they were downgraded in the *ReliefF* classification. Biologically, it would not be surprising. Indeed, if we admit that feral populations of 2003 are dependent on the presence in the past of some fields and feral populations, as all methods demonstrated, then we can reasonably admit that the feral populations of 2002 and 2001 depend on the presence of past populations too.

We thus expected all the attributes concerning the feral populations presence to be more and less dependent: dependencies among years as some populations might have produced seeds that produced plants the following years, others might have grown from the same old seed bank, or some could even result from frequent seed spillage over particularly favourable habitats (Charters et al., 1999; Pivard et al., in press) and dependencies within years, between the attributes concerning the presence a given year of feral populations at flowering and at harvest period. Conversely, the field attributes are of course independent of the presence of feral populations in the past and almost completely independent between them.

4.3. Validity of the evaluation of attribute contribution through ReliefF

The *ReliefF* attribute quality evaluation can be interpreted as an approximation of the contribution of each attribute to the explanation of the predictions. If several attributes are involved in the same way of the explanation, they share the credit in their quality estimate, and are thus each less important than they appear through impurity functions. Ideally, if important attributes are not equally important then *ReliefF* should share the credit among them in proportion of the explained concept (Robnik-Sikonja and Kononenko, 2003). In practice (with limited number of examples), less important attributes get less than this because the differences in predicted value caused by less important attribute are overshadowed by larger changes caused by more important attributes. In concrete terms, among important attributes, *ReliefF* could have underestimated the contribution of some feral populations in the past while favouring independent field attributes.

4.4. *Classification tree*

The presented classification tree reaches a compromise between the two ranking methods. We note that the tree is structured by the top five attributes of each method, plus several well-ranked attributes on both methods. That is logical: if the splits are built on impurity decreases, the tree does take into account dependencies, when it does splits subtrees—splits there depend on the split and the attribute value in the top node.

4.5. *Biological interpretations*

Many known results about seed input from fields were confirmed:

- The presence of an adjacent rape field in 2002 was among the whole set of attributes the most relevant cause of feral population presence in 2003 (second in *ReliefF*). Notably, it tripled the mean probability to find a feral population in 2003. Some other field attributes appeared recurrently to be important, especially in the *ReliefF* ranking. The probability of feral population presence in 2003 almost doubled with the presence of adjacent or opposite fields the same year too, meaning that seed input during sowing was not insignificant. Moreover, the probability of presence of a feral population in 2003 increased with the presence of an adjacent field in 2000 and 2001, indicating that a seed bank was probably created in 2000 or 2001, which allowed plant emergence in 2003. Many observed feral populations do result from seed spillage during sowing, harvesting or natural seed losses from edge plants, delayed sometimes by seed storage in the seed bank (Charters et al., 1999; Lutman, 2003; Pivard et al., in press).
- In the ranking methods, the attributes concerning adjacent field presence were almost always ranked before field presence on the opposite roadside. Although sowing or harvesting machinery obviously projected seed across the road, the effect was bigger if the field was really close by. Natural shedding could explain those differences. Price et al. (1996) showed indeed that natural shedding represents between 40 and 60% of the overall losses in winter oilseed rape.

Depicted as the most relevant attribute by the *ReliefF* method (responsible for one-fourth of the variance), the road type appeared as a key attribute structuring the classification tree, in interaction with the presence of oilseed rape. Paved roads tended to display higher probabilities to find a feral population, especially the two-lane roads very often directed toward the silo. Considering more intense seed transport there is an attractive assumption but this is not the only environmental factor alleged to depend on the road type. Among others, weed management is known to differ on each road type. On tracks, field margins are weeded by farmers whereas paved road sides are either under the responsibility of the village or the district, depending on the width. We could thus have found more feral populations along two-lane paved roads because weed management was milder there. In any event, seed losses during transport do exist. They were estimated to be at the origin of about 15% of the observed feral populations for the area (Pivard et al., in press) but could reach 60% (possibly confounded with old seed bank) in Tayside in Scotland (Charters et al., 1999).

Studying thoroughly the processes (local recruitment or seed bank) involved in the persistence of feral populations was our main challenge. Indeed, information on the presence of feral plants over time does not reveal in itself the processes involved in population persistence (Charters et al., 1999; Pivard et al., in press). Allowing us to handle a large number of data and attributes, machine learning methods enabled us to study finely the effect of past feral populations, especially considering the period of observation and the possible production of seeds.

We found here quite clear results, although we had to keep in mind the effect of feral populations presence in the past could have been overestimated by the *Information gain* ranking and underestimated by *ReliefF* (because it was outranked by larger changes due to the presence of fields or the road type).

The feral populations present on the opposite roadside never appear important, whether at flowering period or at harvest. Contrary to the fields, no machinery helps the seeds to cross the road. Information gain incidentally showed the presence of opposite feral populations in 2000 and 2002 not too badly ranked (but not *ReliefF*). They could be correlated to the presence of adjacent feral populations the same year (both founded by symmetric losses from a truck), or more simply result from a field, the same or previous years.

The presence of adjacent feral populations in the past was obviously more important, and especially at flowering period. Notably, both evaluation methods ranked the presence of feral populations at flowering period far before the July attributes (when seeds are mature). This point was crucial because it showed that persistence relied more weakly than expected on local recruitment. The classification tree allowed us to go further. Two nodes concerned the presence of feral populations in July 2001 and 2002, which were thus important attributes to predict the presence of feral populations in 2003. But the tree split on the presence/absence of the populations and did not differentiate whether it produced seeds or not, which confirmed a weak contribution of local recruitment to the presence of feral populations (a few percentages according to *ReliefF*). In the absence of an adjacent field the previous year, the presence of adjacent feral populations in April 2000 was the best attribute to predict the feral populations presence in 2003. Thus, not only did the whole tree provided evidence for persisting seed banks, but that particular node also suggested that the seed banks were often older than 4 years.

5. Conclusion

Machine learning methods have an inherent ability to discover patterns in the data that are not easy to detect using conventional statistical models (Dzeroski, 2001). *ReliefF* in particular is a powerful method to handle complex relationships between attributes and proved here to supply reliable results compared to statistical methods (Pivard et al., in press). Using several machine learning methods, two feature ranking methods and a decision tree construction helped to highlight information

that neither of the three methods alone would have emphasized about the dynamics of feral populations and in particular their persistence ability. The results of the analyses converged to show that local recruitment had a low contribution compared to that of seed banks in the persistence of feral populations, and confirmed the seed bank to be the keystone of the dynamics of oilseed rape feral populations, with large seed immigration from fields. From this experience, we strongly recommend the use of several complementary methods as an optimal strategy to put into perspective results.

Acknowledgements

We are very grateful to D. Mckey, J. Shykoff and the anonymous referees for their helpful comments and corrections on the manuscript. This study was financially supported by the European project "Sustainable Introduction of GMOs into European Agriculture" (SIGMEA, contract no. 502981)

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group.

Burel, F., Baudry, J., Butet, A., Clergeau, P., Delettre, Y., Le Coeur, D., Dubs, F., Morvan, N., Paillat, G., Petit, S., Thenail, C., Brunel, E., Lefeuvre, J.-C., 1998. Comparative biodiversity along a gradient of agricultural landscapes. Acta Oecol. 19, 47–60.

Charters, Y.M., Robertson, A., Squire, G.R., 1999. Investigation of Feral Oilseed Rape Populations. Department Environment Transport Regions.

Clements, D.R., DiTommaso, A., Jordan, N., Booth, B.D., Cardina, J., Doohan, D., Mohler, C.L., Murphy, S.D., Swanton, C.J., 2004. Adaptability of plants invading North American cropland. Agric. Ecosyst. Environ. 104, 379–398.

Crawley, M.J., Brown, S.L., 1995. Seed limitation and the dynamics of feral oilseed rape on the M25 motorway. Proc. R. Soc. Lond. B 259, 49–54.

Crawley, M.J., Brown, S.L., Hails, R.S., Kohn, D.D., Rees, M., 2001. The performance of transgenic crops in natural habitats: a 10-year perspective. Nature 409, 682–683.

Crawley, M.J., Hails, R.S., Rees, M., Kohn, D., Buxton, J., 1993. Ecology of transgenic oilseed rape in natural habitats. Nature 363, 620–623.

Debeljak, M., Dzeroski, S., Jerina, K., Kobler, A., Adamic, M., 2001. Habitat suitability modelling for red deer (Cervus elaphus L.) in South-central Slovenia with classification trees. Ecol. Model. 138, 321–330.

Dzeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. Ecol. Model. 146, 263–273.

Dzeroski, S., Todorovski, L., 2003. Learning population dynamics models from data and domain knowledge. Ecol. Model. 170, 129–140.

Freemark, K.E., Boutin, C., Keddy, C.J., 2002. Importance of farmland habitats for conservation of plant species. Conserv. Biol. 16, 399–412.

Garnier, A., Lecomte, J., 2006. Using a spatial and stage-structured invasion model to assess the spread of feral populations of transgenic oilseed rape. Ecol. Model. 194, 141–149.

Hancock, J.F., Grumet, R., Hokanson, S.C., 1996. The opportunity for escape of engineered genes from transgenic crops. HortScience 31, 1080–1085.

Hunt, E., Martin, J., Stone, P., 1966. Experiments in Induction. Academic Press, New York.

Kleijn, D., Ineke, G., Snoeijing, J., 1997. Field boundary vegetation and the effects of agrochemical drift: botanical change caused by low levels of herbicide and fertilizer. J. Appl. Ecol. 34, 1413–1425.

Kleijn, D., Verbeek, M., 2000. Factors affecting the species composition of arable field boundary vegetation. J. Appl. Ecol. 37, 256–266.

Kononenko, I., 1994. Estimating attributes: analysis and extension of RELIEF. In: Proceedings of the Seventh European Conference on Machine Learning, Springer-Verlag, Berlin, Catania, Italy, pp. 171–182.

Lek, S., Guegan, J.F., 1999. Application of artificial neural networks in ecological modelling. Ecol. Model., 120.

Lutman, P.J.W., 2003. Co-existence of conventional, organic and GM crops—role of temporal and spatial behaviour of seeds. In: Boelt, B. (Ed.), Proceedings of the First European Conference on the Co-existence of GM Crops with Conventional and Organic Crops (GMCC-03). Danish Institute of Agricultural Sciences, Denmark, pp. 32–42.

Marshall, E.J.P., Moonen, A.C., 2002. Field margins in northern Europe: their functions and interactions with agriculture. Agric. Ecosyst. Environ. 89, 5–21.

Pessel, F.D., Lecomte, J., Emeriau, V., Krouti, M., Messean, A., Gouyon, P.H., 2001. Persistence of oilseed rape (*Brassica napus* L.) outside of cultivated fields. Theor. Appl. Genet. 102, 841–846.

Pivard, S., Adamczyk, K., Lecomte, J., Lavigne, C., Bouvier, A., Deville, A., Gouyon, P.H., Huet, S. Where do the feral populations come from? A large-scale study of their possible origin in a farmland area. J. Appl. Ecol., in press.

Price, J.S., Hobson, R.N., Neale, M.E., Bruce, D.M., 1996. Seed losses in commercial harvesting of oilseed rape. J. Agric. Eng. Res. 65, 183–191.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn. 1, 81–106.

Quinlan, J.R., 1993. Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

Recknagel, F., 2001. Application of machine learning to ecological modelling. Ecol. Model. 146, 1–3 (Special issue).

Robnik-Sikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learn. J. 53, 23–69.

Stewart, C., Halfhill, M., Warwick, S., 2003. Transgene introgression from genetically modified crops to their wild relatives. Nat. Rev. Genet. 4, 806–817.

Stone, M., 1977. Asymptotics for and against cross-validation. Biometrika 64, 29–35.

Wilson, S.D., Tilman, D., 1991. Interactive effects of fertilization and disturbance on community structure and resource availability in an old-field plant community. Oecologia 88, 61–71.

Witten, I.H., Frank, E., 1999. Data Mining: Practical Machine Learning Tools with Java Implementation. Morgan Kaufmann, San Francisco.