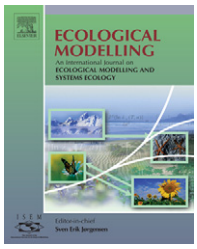
available at www.sciencedirect.com

ScienceDirect

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



Relations between the oilseed rape volunteer seedbank, and soil factors, weed functional groups and geographical location in the UK

Marko Debeljak[a,*], *Geoff R. Squire*[b], *Damjan Demšar*[a], *Mark W. Young*[b], *Sašo Džeroski*[a]

[a] *Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*
[b] *Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK*

ARTICLE INFO

Article history:
Published on line 26 November 2007

Keywords:
Oilseed rape (*Brasica napus*)
Volunteer weed
Soil seedbank
Plant functional groups
Data mining
Farm scale evaluations of GMHT crops

ABSTRACT

Data mining techniques were applied to model the presence and abundance of volunteer oilseed rape (OSR) (*Brasica napus* L.) in the seedbank at 257 arable fields used for baseline sampling in the UK's Farm Scale Evaluations of genetically modified herbicide tolerant (GMHT) crops. Constructed models were supported by statistical tests. Volunteer OSR was most likely present if a previous OSR crop had been grown in the same field, but it was also present at sites where it had not been grown in the previous 8 years (24% of all fields). In 136 fields where it was found, it showed a slow decline in abundance since the last crop. However, data mining indicated previously unfound correlations between oilseed rape abundance, total seedbank and several other factors, notably percent of nitrogen and percent of carbon in the soil, all of which were smallest in the centre of arable production in southern England and greatest in the surrounding south-west, west and north. In a separate analysis, its abundance was also associated with particular plant life history groups, which include broadleaf weeds such as *Capsella* and *Matricaria* species, having a similar phenology to oilseed rape, between rapidly developing annuals and the biennials and perennials. The findings are a reference point in the evolution of oilseed rape as a weed and potential GM impurity. Data mining approaches provide models that may be used to assess the status of volunteer OSR in other countries or at a later time in the UK.

© 2007 Elsevier B.V. All rights reserved.

1. Introduction

Crops and weeds of the same or similar species have lived side by side since the beginning of agriculture. In modern crops, 'wild' traits have mostly been reduced or removed for the convenience of agronomy, but one which has been difficult to remove is the shattering of fruiting structures before harvest and the resultant deposition of seed to the soil. If such seed becomes incorporated, and persists through primary or induced dormancy, it may emerge as a volunteer weed in later crops or as a feral plant around agriculture. This ability of crops to leave volunteer weeds is now widespread in maize, rice, other temperate and tropical cereals, beet, soybean, oilseed rape, and many minor crops, and increases the agricultural weed burden worldwide (Gressel, 2005). Here, the distribution and abundance of volunteer or weedy oilseed rape are examined as an example of this widespread phenomenon of crops becoming weeds.

* *Corresponding author*. Tel.: +386 1 4773124; fax: +386 1 4773315.
E-mail address: marko.debeljak@ijs.si (M. Debeljak).
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.10.019

The status of rapeseed or oilseed rape (*Brassica napus* L. and the oilseed form of the related *Brassica rapa* L.) as a volunteer has changed greatly since its first recorded presence in Mediaeval times (Thirsk, 1997). Its area of production expanded and contracted, but its use in earlier centuries appeared not to leave established permanent weedy populations, since it was not recorded in arable seedbank surveys in Britain in the 20th century before the 1970s (e.g. Milton, 1943; Champness and Morris, 1948; Roberts and Stokes, 1966). The greatest change occurred between 1970 and the mid-1990s when new varieties with oils fit for consumption by people and livestock increased in sown area from 1% to 10–15% of the arable land surface of Britain, where it functions largely as a break crop every three to five years in cereal fields (Squire et al., 2003). This rise in area, similar in many parts of Europe, brought with it the highly visible, crop-derived oilseed rape along roadsides over much of the UK, except in upland or mountainous areas (Preston et al., 2002), and the wide occurrence of oilseed rape as a volunteer weed in arable crops (Gruber et al., 2004a,b; Lutman et al., 2005; Pekrun et al., 1998).

Specific forms of field management can be applied to contain its contribution to the weed burden (Pekrun et al., 1998), but keeping populations low enough to prevent their being an impurity, at say 0.9% for GM in non-GM (EC, 2000) would be more challenging. For this reason, the problem of volunteer oilseed rape as an impurity is one of the major factors to be considered in drawing up measures for the coexistence of GM cropping with other types of agriculture (Messean et al., 2006).

Solutions to the problems of oilseed rape and other volunteer weeds are often hampered by a lack of broad-scale, quantitative, reference data against which trends, future change and agronomic remedies can be assessed. Information would be particularly valuable on the population size and distribution of the volunteers during the early phase of their expansion, but up to the late 1990s, most experiments on the seedbank dynamics of oilseed rape had been on experimental stations. Too little was known of its occurrence and abundance in the wider range of soils, cropping sequences and field management in commercial agriculture. However, an opportunity to establish a baseline for volunteer oilseed rape was provided by the first widespread survey of the arable seedbank in Britain since the 1970s, made as part of baseline measurements on the 257 fields used for the Farm Scale Evaluations of GM herbicide-tolerant crops (Firbank et al., 2003; Squire et al., 2003; Bohan et al., 2005). The seedbank is recognised as a slowly shifting indicator of recent field-history and of the problems likely to be caused by volunteer OSR.

This paper uses data-mining techniques (Debeljak et al., 2001; Džeroski, 2001; Wang and Witten, 1997), supported by statistical tests, to define the main influences among the many potential interactions between volunteer oilseed rape, previous crops, other weeds and their environment in these 257 fields. The initial question is whether volunteer oilseed rape had become widespread or is still confined to fields that recently grew oilseed rape as a crop. The data are then examined for relations between its abundance in the seedbank and contextual factors that include location in the country, soil type, crop rotation, and the other seedbank species.

2. Data and methods

2.1. *The dataset*

Seedbanks and site characteristics were measured during 2000, 2001 and 2002 as baseline samples at 257 sites in the Farm Scale Evaluations of GM herbicide-tolerant crops (Firbank et al., 2003; Squire et al., 2003). Sites were located from the north-east of Scotland to the south coast of England. All measurements here were made before the fields were divided into experimental treatments. Each site was described by four groups of attributes: the location of the site in Britain; the soil conditions, including texture, pH, carbon and nitrogen; the crops, including previous oilseed rape; and the seedbank itself. The crops grown up to 8 years before sampling (C1-C8) were categorised by main crop type (cereal, grass, lay and set aside, oilseed, vegetable including legumes), miscellaneous if more than one crop was sown, e.g. C2 = vegetable – two years ago the crop was vegetable, etc.), spring or winter sowing, the main target of weed control (grass weed in broadleaf crop, broadleaf weed in grass crop, and not determinable or none). Groups of sites in Scotland, north England and south-west England were geographically distinct. The remaining large southern grouping was further split between south-middle and south-east.

Seedbanks were estimated by the 'emergence' method, which was similar to that used in previous studies of arable and horticultural seedbanks (cf. Milton, 1943; Champness and Morris, 1948; Roberts and Stokes, 1966). Soil samples, each of 1 kg, were taken from a depth of 0–15 cm at 16 systematic locations in each field. Soil was taken between late March and early May for the spring-sown beet, maize and spring oilseed rape and in late July and August for winter oilseed rape and sent to the Scottish Crop Research Institute (SCRI) for assessment of the seedbank. The total soil processed was 16 kg per site and >3800 kg in total, 10 times more soil than in any previous survey of arable or horticultural seedbanks in the Britain. Soil was pressed through sieves, the stones removed, the fine soil placed in trays to occupy typically 1 L of space to a depth of 3 cm and kept moist in glasshouses for several months at SCRI. Seedlings that emerged from each tray were counted and identified to species. Seedbanks were described in terms of number of species in the seedbank, total seeds per square metre, absence or abundance of *B. napus* and life history groups of the seedbank species.

Around 200 plant taxa were recorded from all 257 sites. Sites had between 5 and 36 species and the moderately frequent or infrequent species were often confined to a few sites. Therefore to enable comparison across sites, species were reduced to functional classifications based on life history traits widely used in crop studies (Squire, 1990), namely annuality (annual, non-annual), speed of progress to reproduction for the respective annuality class (fast, slow), compactness (compact, spreading), determinacy (determinate, indeterminate) and position in the canopy (in/below canopy). Traits were taken from SCRI's arable species database, augmented by standard flora when necessary (Clapham et al., 1962; see also Hawes et al., 2005). For example, one of the most populous groupings was 'annual, fast, compact, indeterminate,' and included species such as *Poa annua*; while 'annual, slow, com-

pact, indeterminate' included many of the broadleaf weeds that reproduce more or less in synchrony with the crop and mature in and around July. Descriptions of the life history categories for the common seedbank species are provided at (http://sigmea.scri.sari.ac.uk/general_release/).

Soil from each of the 16 seedbank samples from a site was retained for measurement of pH and total carbon and nitrogen content by standard laboratory procedures. The values used in the analysis for a site are the means of determinations on each of the 16 samples. The carbon and nitrogen percentages were measured by mass spectrophotometry and two replicates per sample averaged.

2.2. *Data mining and statistical analysis*

Based on many positive results of applying machine learning methods in general and data mining in particular in ecological modelling (Stankovski et al., 1998; Debeljak et al., 2001; Džeroski, 2001; Jerina et al., 2003), the machine learning technique of model tree induction was applied in modelling the effects of environmental attributes on the seedbanks.

Model trees are a representation for piece-wise constant or piece-wise linear functions. Like classical regression equations, they predict the value of a dependent variable (called class) from the values of a set of independent variables (called attributes). Data represented in the form of a table can be used to construct a regression tree. In the table, each row (sample) has the form (x_1, x_2, ..., x_N, y), where x_i are values of the number of attributes (e.g., texture, organic matter content, pH, N, etc.) and y is the value of the class (e.g. total number of seeds per volume of soil).

Unlike classical regression approaches, which find one single equation for a given set of data, model trees partition the space of examples into axis-parallel rectangles and fit a model

Table 1 – Attributes used in modeling effects of soil and crop dependent factors on the oilseed rape soil seedbank

Attribute	Description	Units
Location	Location from where the sample of seedbank was taken	1: Scotland, 2: north UK, 3: south-east UK, 4: south-middle UK, 5: south-west UK
Sample year	Year when the seedbank sample was taken	Year
Number of species	Total number of plant species recorded in the seedbank sample	#
Total seeds per square metre	Total seedbank per square metre to soil depth of 0.15 m	#
pH	pH of the soil	#
Nitrogen	Nitrogen content of the soil	%
Carbon	Carbon content of the soil	%
Texture	Soil texture	f, fine; m, medium; c, coarse
Crop-1 to 8	Crops in the 8 years before the soil samples were taken (1 is the 1st year before)	96 different combinations of crops
Last OSR crop	Years since the last OSR crop was sown	Years
Any OSR crop	Presence of OSR crop in last 8 years before the soil samples were taken	Yes, no
Number of winter cereal crops in 8 years	Number of winter cereals crops in last 8 years before soil samples were taken	#
Percentile of winter cereals in last 8 years	Percentile of winter cereals in last 8 years before soil samples were taken	%
Number of winter cereals since last rape	Number of winter cereals crops since last OSR crop was sown	#
Percentile of winter cereals since rape	Percentile of winter cereals since the last OSR was sown	%
C1-8 crop grown	Crops grown in the 8 years before the soil seedbank samples were taken (1 is the 1st year before)	96 different combinations of crops
C1-2 Season	Season when the crop was sown in the 8 years before the soil seedbank samples were taken (1 is the 1st year before)	Winter, spring
C1-8 Type	Types of the crops grown in the 8 years before the soil seedbank samples were taken (1 is the 1st year before)	Cereal, Ley (grass ley or set aside), Misc. (miscellaneous), Oilseed, Vegetable
C1-8 Target	Herbicide applied on specific target groups of plants	Broadleaf, grass, not determinable
Functional groups	Number of seeds for particular functional group of plants in the soil seedbank (48 groups)	#/site
B. napus per square metre	Density of *Brasica napus* per m^2	#/m^2
B. napus seeds promile	Promile of *Brasica napus* in seedbank	Promile
B. napus per m^2 rank	Ranks of relative density of *Brasica napus* per m^2	9 = absent, 1 = 0–25%, 2 = 25–50%, 3 = 50–75%, 4 = 75–100%
B. napus seeds present	*Brasica napus* seeds present	Yes, no

to each of these partitions. A model tree has a test in each inner node that tests the value of a certain independent variable (attribute), and in each leaf a model for predicting the value of a dependent variable (class): the model can be a linear equation or just a constant.

A number of systems exist for inducing regression trees, such as CART (Breiman et al., 1984) and M5 (Quinlan, 1992). The system M5′ (Wang and Witten, 1997), a reimplementation of M5 within the software package WEKA (Witten and Frank, 1999) is here used, the parameters of M5′ being set to their default values. The quality of the model (i.e. predictive performance) was evaluated with the Pearson correlation coefficient and several other error measures (i.e. mean average error, root mean square error, and relative average error and root relative square error) using 10-fold cross validation. The dataset is split into a ten parts. All parts except one are combined and a model tree is induced on these data and tested on the remaining parts. This procedure is repeated for all the parts and in the end the average correlation coefficient is computed which can be used as a measure of the model's predictive power.

Statistical analysis, using STATISTICA data analysis software, was used to support relations predicted by data mining. Basic descriptive statistics, frequency analysis and various tests for differences between arithmetic means were applied using both parametric (*t*-test) and nonparametric tests (Mann–Whitney *U* test, Wals–Wolfowitz runs tests, and Kolmogornov–Smirnov test). Spearman rank order correlations, Gamma correlations and Kendall Tau correlations were used in estimating correlations between site properties. To address the complex interdependent effects of environmental attributes on seedbanks, ANOVA and parametric multiple regression were applied.

The dependent variable (OSR in seedbank) was treated in the experiments in a three different ways: (1) as a numeric variable (absolute and relative concentration or abundance), (2) as an ordered rank variable with four quartiles of the relative concentration, and as a discrete variable denoting presence/absence (Table 1).

3. Results and interpretation

The approach here makes no presumptions about the association between oilseed rape and any variables of crop management, site, or other seedbank species. Indeed, very little is known about any such associations. Correlations indicated by data mining are then examined by statistical methods.

3.1. *Site, soil and cropping*

The models explored two main types of variable that might influence presence/absence: those concerned with the type of crop, season and weed management target, and those with site characteristics. When only crop factors were used, 61% (estimation by cross-validation) of sites were classified correctly (Fig. 1). The main explanatory variable was whether or not oilseed rape had been grown in the field over the past 8 years. No important, independent effects of season of sowing or weed management target were detected. When only

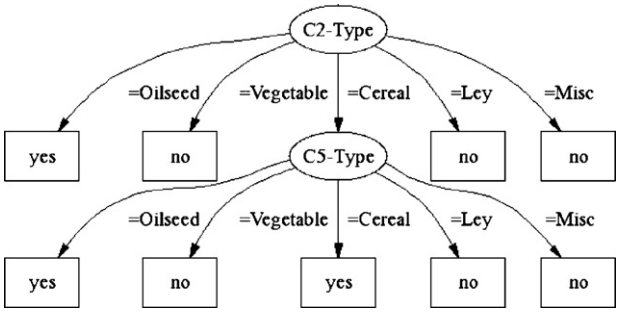


Fig. 1 – Classification of presence of oilseed rape by crop type (C2-Type: crop type 2 years before the sampling date; C5-Type: crop type 5 years before the sampling date; types are Oilseed, Miscellaneous (Misc.), Cereal, Vegetable, grass ley or set aside (Ley) (correctly classified instances: 60.7 %).

site factors were used, 71% (estimation by cross-validation) of sites were classified correctly, based mainly on location in the country and soil texture. Sites were more likely to contain oilseed rape at locations 1 and 2 (Scotland and north England) and more likely at other locations in specific soil textures (Fig. 2)

The importance of a recent oilseed rape crop was not unexpected, but that geographical location and soil factors could influence the presence of oilseed rape in the seedbank had not been found previously in the UK. A more discriminating analysis was made by using the quantitative, continuous variables, of % of carbon and % of nitrogen, pH, total seedbank and number of species. As a first comparison, the soil factors, species number and total seedbank density were compared at sites with and without oilseed rape by three statistical tests (*t*-test, Mann–Witney *U* test, and Wald–Wolfowitz runs test).

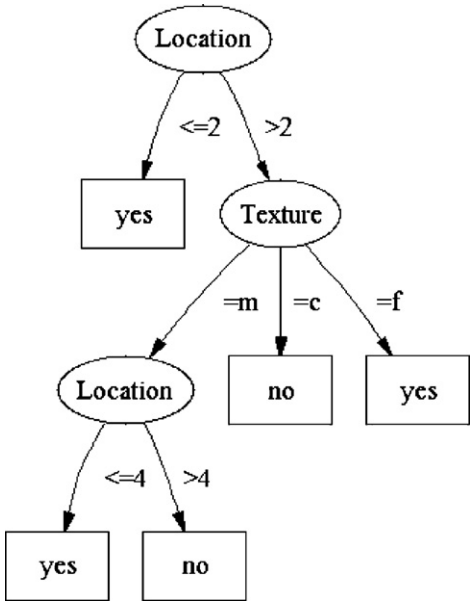


Fig. 2 – Classification of presence of oilseed rape by location and soil texture (location: 1, north-west Scotland; 2, north England; 3, south-east England; 4, south middle England; 5, south-west England; texture: f, fine; m, medium; c, coarse) (correctly classified instances: 70.8%).

Table 2 – Comparison of sites with (N = 136 or 53.9%) and without (N = 121) oilseed rape in the seedbank

Variable	Sites with oilseed rape				Sites without oilseed rape				P
	Mean	Minimum	Maximum	S.D.	Mean	Minimum	Maximum	S.D.	
n species	19.8	5.0	36.0	5.85	18.3	3.0	37.0	6.82	u*, w
Seeds (m^{-2})	2913	422	15966	2425	2986	169	24281	3518	n.s.
pH	6.46	5.30	7.52	0.55	6.53	4.90	7.53	0.56	n.s.
% of nitrogen	0.302	0.100	2.11	0.276	0.266	0.080	1.36	0.208	u*
% of carbon	3.67	0.91	26.6	3.62	3.11	0.96	15.5	2.38	n.s.

* Indicates significant difference ($P < 0.05$) detected by Mann–Witney *U* test (u) or Wald–Wolfowitz runs test (w), n.s.: not significant by any of the tests.

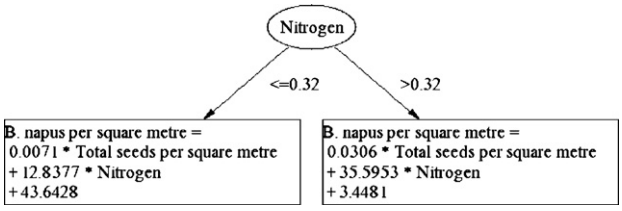


Fig. 3 – Prediction models for density of oilseed rape seeds in the seedbank per square metre, which are dependent on the percent of nitrogen in the soil (correlation coefficient: 0.3567, relative absolute error: 97.6%).

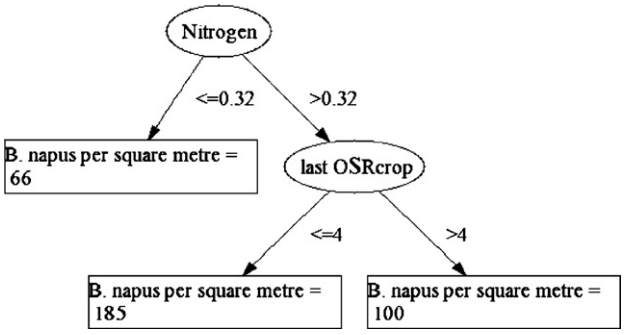


Fig. 4 – Prediction models for density of oilseed rape seeds in the seedbank per square metre, which depend on the percent of nitrogen in the soil and the time since the last oilseed rape (last OSRcrop) was grown in the field (unit: year) (correlation coefficient: 0.2128, relative absolute error 97.43%).

Sites with and without did not differ in pH or total seeds m^{-2}, but number of species was significantly higher with OSR in two of the three tests (and almost significantly different, $P = 0.053$, using the *t*-test), while soil nitrogen was higher with OSR in one test (Table 2).

For abundance, the model with highest correlation ($r = 0.31$) showed high OSR seedbanks were found where % of nitrogen was >0.32% and seedbank density was also high (Fig. 3). When crop and site factors were combined, the model for abundance giving the highest correlation ($r = 0.21$) indicated high seedbanks when nitrogen was high (>0.32%) and time since the previous oilseed rape crop <4 years (Fig. 4).

Statistical tests were therefore carried out assess the relations between OSR density and site factors. In the first analysis, the rank order of oilseed rape was compared against site factors using three rank-order methods (Spear-

Table 3 – Relations between geographical location, the OSR seedbank and site factors showing means and superscript text below the means, sites whose values are significantly different at <0.05% (P value was <0.001 in most instances)

Location	1: North-west Scotland	2: North England	3: South-east England	4: South-middle England	5: South-west England
OSR seedbank (m^{-2})	156 3, 4	73	60 1, 5	63 1, 5	144 3, 4
Total seedbank (m^{-2})	5370 2, 3, 4, 5	2847 1	2643 1, 5	1738 1, 5	3783 1, 3, 4
Species number	21.7 2	17 1, 5	19.4 5	19 5	22.8 2, 3, 4
pH	5.6 2, 3, 4, 5	6.04 1, 3, 4, 5	6.73 1, 2, 5	6.55 1, 2	6.5 1, 2, 3
% of nitrogen	0.35	0.26	0.27 5	0.26 5	0.44 3, 4
% of carbon	3.93	4.4	3.16 5	2.8 5	5.73 3, 4

Only sites with OSR in seedbank are included (N = 136).

man, Gamma, Kendall Tau). All showed highly significant positive correlations between oilseed rape and species number (*P* values, respectively, 0.0049, 0.00124 and 0.00124), % of nitrogen (0.0033, 0.00102, 0.00103) and % of carbon (0.021, 0.012, 0.012).

The second analysis separated the 136 fields into geographical locations then compared the means of the variables between locations (Table 3). Major differences were revealed which together showed low values of the OSR seedbank and low values of total seedbank, species number, % of nitrogen, and % of carbon in the south-east and south-middle of the UK, and higher values elsewhere. % of nitrogen and % of carbon are themselves highly correlated both as means among the locations and across individual sites. This explains why when one of them (% of nitrogen) gave explanatory power in the models, the other provided no greater correlation. The values of pH differed between locations but not systematically in relation to the centre and periphery. The analysis has therefore revealed important effects of location on oilseed rape density that appear related to the total seedbank and to the % of carbon and % of nitrogen of the soil, the greatest effects being the more than two-fold suppression of all factors in the south-east and south-middle regions relative to the north and south-west.

The indication of time since the previous oilseed rape crop as a factor was further examined. The regression of oilseed rape abundance on time for sites with OSR in soil seedbank was not significant; neither was ANOVA on the mean abundance for each of the years 1–8. However, *t*-tests showed the mean abundance of OSR in years 1–4 (109.03 m^{-2}) was significantly greater ($P = 0.00317$) than the mean abundance in years 5–8 (38.26 m^{-2}). The abundance of the seedbank as a whole was similar between these respective periods being 2997 m^{-2} in years 1–4 and 2991 m^{-2} in years 5–8 (not significant).

In order to examine the relationship between locations (1, north-west Scotland; 2, north England; 3, south-east England; 4, south middle England; 5, south-west England), and soil texture (f, fine; m, medium; c, coarse) additional statistical tests (Pearson Chi-square test) were conducted. The results show no significant relationships between locations included in the research ($N = 257$) and soil texture ($P = 0.26197$) and no significant relations between locations where OSR was sown any time in last 8 years ($N = 123$) and soil texture ($P = 0.5786$). However the statistical test was significant ($P = 0.01739$) for locations where the OSR was sown any time in last 8 years and where OSR was present in soil seed bank ($N = 103$). The results of the last experiment fit very well with the classification model presented in Fig. 2.

3.2. *Association with plant functional groups*

Independently of predictions based on site characteristics, the models were used to predict the relation between plant functional groups on the three attributes of the OSR seedbank - presence or absence among 257 sites, rank in four quartiles (1 lowest, 4 highest) and absolute density (m^{-2}). To simplify the procedure, while still using all data, five sets of functional groupings were chosen for analysis based on each of the hierarchies: annuality, speed, compactness, determinacy and position in canopy. The combination of functional groups

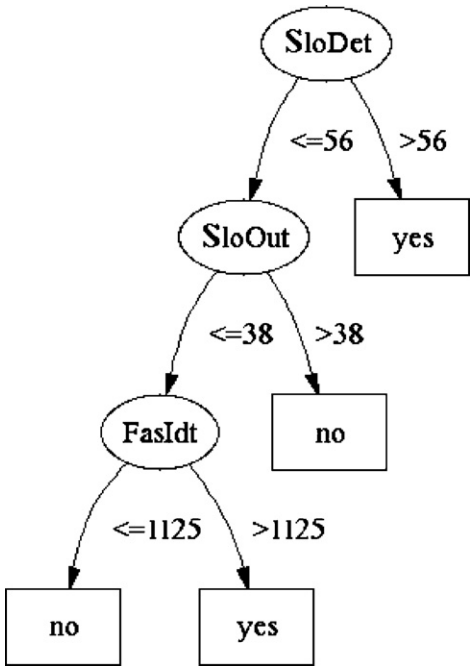


Fig. 5 – Classification of presence of oilseed rape by the abundance of plants (m^{-2}) of particular functional groups (SloDet – slow, determinate development; SloOut – slow development living below the crop canopy; FasIdt – fast indeterminate development) (correctly classified instances: 63.8%).

that gave the highest correlation omitted annuality (probably since it was so prevalent) and used the attributes of speed, compactness, determinacy and position in canopy. The model for presence/absence gave a correct categorization for 64% of the 257 sites, and indicated the primary distinguishing group was 'slow determinate': sites were likely to have oilseed rape if they had >56 m^{-2} seeds of this group in the seedbank (Fig. 5). The average for this group over all sites was 612 m^{-2} out of a total mean seedbank of 2940 m^{-2} for all species. The low cut-off of 56 m^{-2} and the fact that the group was not highest in abundance – constituting about 1/5th of the total seedbank – implies a strong dependence between this group and oilseed

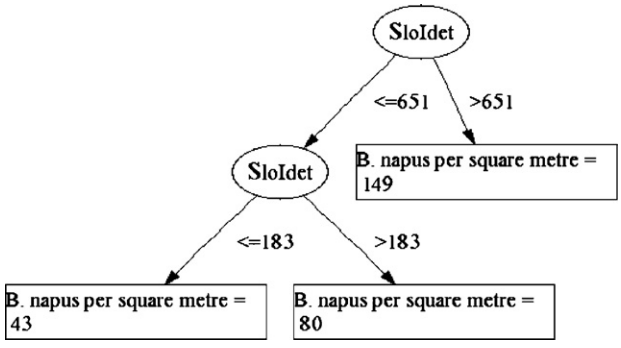


Fig. 6 – Classification of the abundance of OSR by functional groups (SloIdet – slow indeterminate development) (correlation coefficient: 0.4378; relative absolute error: 86.6 %).

Table 4 – Comparison of the abundance (seeds m^{-2}) of plant functional groups at sites with (136 sites) and without (121) volunteer oilseed rape in the seedbank

Functional groups	Abundance in seedbank (m^{-2})		P
	Sites with OSR	Sites without OSR	
Annual (all)	2535	2492	n.s.
Non-annual (all)	370	491	n.s.
Annual determinate	318	147	0.032
Compact determinate	336	166	0.0041
Slow determinate	283	126	0.0061
Below-crop determinate	199	79	0.00054
Non-annual slow	1.4	7	0.0028

The broad categories of annual and non-annual are shown for reference, did not differ significantly (n.s.) and are compared with representative categories that were significantly different, most of which were variations on the determinate functional trait.

rape. If sites fell below this value, then they were less likely to be associated with oilseed rape, except if they had small populations of 'slow in-crop' and large populations of 'fast indeterminate' groups.

For sites with OSR in the seedbank, the model for rank (not shown) correctly classified 35% of sites, and pointed to an association of OSR sites with slow, compact and indeterminate plants. The model for absolute density (Fig. 6) – which gave the highest correlation coefficient of 0.44 of all correlations with density – partitioned the sites into three categories of abundance (with breaks at 149, 80 and 43 m^{-2}) determined solely by the abundance of slow indeterminate plants (e.g. *Chenopodium* sp., *Persicaria* sp.).

The analyses of presence, rank and abundance therefore together pointed to an association between oilseed rape and plants (mostly annual) that were relatively slow to reproduce and variously determinate (presence–absence) and indeterminate (abundance). None of the models indicated that the attributes non-annual (biennial *Compositae*, woody species) or spreading (*Polygonum aviculare*, *Galium aparine*, *Stellaria media*) were associated with OSR in the seedbank, while the models were mostly neutral towards the attribute of 'position in the canopy'.

Statistical tests were performed for the means of abundance for each functional group for sites with (1 3 6) and without (1 2 1) oilseed rape. The groupings that were the most abundant (annual or fast indeterminate) were generally not discriminating. The groupings that were different between with-OSR and without-OSR sites were those based on slow, compact, determinate development (Table 4). The only exception was the group 'non-annual slow' composed of shrub and tree species, which was in any case very low in abundance.

For those sites with oilseed rape in the seedbank, the abundance of functional groups was compared with the abundance of OSR. The abundance of the whole group 'annual' was positively related to the abundance of oilseed rape ($P = 0.0027$), as were the categories 'annual slow' (0.00027), 'annual determinate' (0.002), 'slow indeterminate' (<0.0001) and several other groups among the annuals including 'in-canopy' and 'belowcanopy'. An important distinction revealed by this analysis is that OSR abundance was not related to certain major and distinctive groups – the non-annual group as a whole, and the major sub-group 'fast indeterminate' whose mean abundance was around 2000 m^{-2} and dominated in these seedbanks by the annual arable weeds of the genus *Poa*.

The statistical analysis of functional groups therefore largely confirmed the results of applying machine learning to the whole dataset. For determining presence/absence, the association between volunteer OSR and plants of the 'slow determinate' (mostly broadleaf) group links it with commonest plants in that category, *Capsella bursa-pastoris*, which is also the most frequent Crucifer, and *Matricaria* species. Among sites where OSR is found, it is associated less with these slow, more or less determinate plants than with the similar but slightly more populous indeterminate group to which many phenotypes of oilseed rape belong.

The association between presence of OSR and presence of the functional group dominated by *Capsella*, *Matricaria and related species* deserves further examination. Direct interactions between the volunteers and this functional group are probably not responsible, since both are typically present as plants at low density compared to the density of the crop. The more likely explanation is that the group is able to emerge and re-seed only or mostly when oilseed rape is present as the 'break' crop in the cereal rotation: the oilseed rape crop both supplies the volunteers and offers competitive conditions and forms of weed control that this group is able to exploit. This hypothesis is strengthened by the finding of low abundance in this functional group at sites where the break crops exclude oilseed rape, but include the much more intensely managed potato and sugar beet in which few weeds typically emerge. Where oilseed rape was absent and one of these crop species occur in the rotation (50 sites), the abundance of the slow determinate functional group was 101 m^{-2}; where oilseed rape was absent and both potato and sugar beet were in the rotation (15 sites), the abundance of the group dropped to 37 m^{-2} (different from 101, $P < 0.05$, by t-test). Soil characteristics also differed among these treatments: for example the corresponding mean nitrogen was 0.25% where either beet or potato occurred and 0.13% where they both occurred (difference significant, $P < 0.001$).

4. Conclusions

From this large bank of information on site attributes, crop rotations, soil types and weed functional groups, data mining

detected certain crucial associations which on further analysis confirmed the role of important and previously unknown factors in the distribution and abundance of volunteer oilseed rape. The presence of volunteers was still strongly dependent on a crop having been grown at a site, but once grown, the resulting volunteers were not excluded specifically from any part of the country or from sites having particular abiotic characters such as high pH or low % of nitrogen. Volunteers had, moreover, become present at 24% sites where there had been no OSR crop in the last 8 years, presumably as a result of a previous crop (beyond the 8 years recorded) or import to the site in farm machinery. Their abundance, moreover, varied systematically with factors that are generally associated with the intensity of farming, notably total seedbank abundance, species number in some instances, and most consistently % of nitrogen and % of carbon in the soil. All these were linked to an extent with geographical region, being smallest in the arable south-central and south-east and largest in the north and south-west. The salient finding is the size of the regional variation, since no previous surveys in the UK had included enough sites to be able to detect any such differences.

Where volunteer oilseed rape had become established, it had characteristics similar to those of particular functional groups of the seedbank. An association was repeatedly indicated with slow, i.e. later flowering, determinate or indeterminate types, which constitute around 20% of the seedbank. Conversely, volunteer OSR was not positively or negatively associated with the 'earliest' weeds of the fast indeterminate category, the most abundant in terms of seed numbers in the seedbank, and mostly grasses, or with the much more slowly developing non-annual forbs, shrubs and trees, the latter of which exist as seedlings in arable fields. While knowledge of the factors driving these functional groups in general is still highly uncertain, the association with the category of annual, slow, determinate/indeterminate arguably fits many observed forms of oilseed rape volunteer which seed about the same time as the cereal or broadleaf crop rather than well before it. The severe suppression of these groups in certain parts of the country and in certain rotations raises the possibility that the presence of oilseed rape is not so dependent on their having been an oilseed rape crop previously at the site, as implied earlier. A contrary hypothesis is that volunteer seed gets widely distributed through being sown and in being carried in farm machinery, but along with similar annuals is highly suppressed at certain types of site, for instance, having beet or potato in a mainly winter cereal rotation, by the very intense management inherent in such rotations. Further work is needed to resolve these hypotheses.

Nevertheless, the findings have clear implications for managing OSR as an impurity. Its mean density was such that only 1% of the seedbank emerging in a later crop of oilseed rape might compromise an impurity-threshold of around 0.9%. Its populations do not appear to have decayed very rapidly with time since the last crop, but generally its populations are low in regions of high intensity production in the south and east and high in the north and south-west, where it could therefore be a greater impurity. Thus far, farmers have been managing volunteer OSR as a general broadleaf weed. If they were to manage it as an impurity, then much more rigorous or specific control would be needed, especially in the peripheral arable regions that have the most abundant and diverse seedbanks and consequently the richest arable food webs.

Acknowledgements

Part of this work was funded through the SIGMEA project (EU FP6). Geoff Squire and Mark Young also receive funding from the Scottish Government. The Farm Scale Evaluations was funded by Defra, London. We thank Gill Banks at SCRI for supervising the soil analysis.

REFERENCES

Bohan, D.A., Boffey, C.W.H., Brooks, D.R., Clark, S.J., Dewar, A.M., Firbank, L.G., Haughton, A.J., Hawes, C., Heard, M.S., May, M.J., Osborne, J.L., Perry, J.N., Rothery, P., Roy, D.B., Scott, R.J., Squire, G.R., Woiwod, I.P., Champion, G.T., 2005. Effects on weed and invertebrate abundance and diversity of herbicide management in genetically modified herbicide-tolerant winter-sown oilseed rape. Proc. Roy. Soc., Ser. B 272, 463–474.

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

Champness, S.S., Morris, K., 1948. The population of buried viable seeds in relation to contrasting pasture and soil types. J. Ecol. 36, 149–173.

Clapham, A.R., Tutin, T.G., Warburg, E.F., 1962. Flora of the British Isles, second ed. Cambridge University Press (1st edition 1952).

Debeljak, M., Džeroski, S., Jerina, K., Kobler, A., Adamic, M., 2001. Habitat suitability modeling for red deer (*Cervus elaphus* L.) in south-western Slovenia with classification trees. Ecol. Model. 138, 321–330.

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

EC, 2000. Commission Regulation (EC) No. 49/2000 of 10 January 2000 amending Council Regulation (EC) No. 1139/98 concerning the compulsory indication on the labelling of certain foodstuffs produced from genetically modified organisms of particulars other than those provided for in Directive 79/112/EEC. Off. J. Eur. Commun. L6, 13–14.

Firbank, L.G., Heard, M.S., Woiwod, I.P., Hawes, C., Haughton, A.J., Champion, G.T., Scott, R.J., Hill, M.O., Dewar, A.M., Squire, G.R., May, M.J., Brooks, D.R., Bohan, D.A., Daniels, R.E., Osborne, J.L., Roy, D.B., Black, H.I.J., Rothery, P., Perry, J.N., 2003. An introduction to the farm scale evaluations of genetically modified herbicide-tolerant crops. J. Appl. Ecol. 40, 2–16.

Gressel, J. (Ed.), 2005. Crop Ferality and Volunteerism. CRC Press, Boca Raton, FL, USA, p. 422.

Gruber, S., Pekrun, C., Claupein, W., 2004a. Population dynamics of volunteer oilseed rape (*Brassica napus* L.) affected by tillage. Eur. J. Agron. 20, 351–361.

Gruber, S., Pekrun, C., Claupein, W., 2004b. Seed persistence of oilseed rape (*Brassica napus*): variation in transgenic and conventionally bred cultivars. J. Agric. Sci. 142, 29–40.

Hawes, C., Begg, G., Squire, G.R., Iannetta, P.P.M., 2005. Individuals as the basic accounting unit in studies of ecosystem function: functional diversity in shepherd's purse, Capsella. Oikos 109, 521–534.

Jerina, K., Debeljak, M., Dzeroski, S., Kobler, A., Adamic, M., 2003. Modeling the brown bear population in Slovenia: a tool in the conservation management of a threatened species. Ecol. Model. 170, 453–469.

Lutman, P.J.W., Berry, K., Payne, R.W., Simpson, E., Sweet, J.B., Champion, G.T., May, M.J., Wightman, P., Walker, K., Lainsbury, M., 2005. Persistence of seeds from crops of conventional and herbicide tolerant oilseed rape (*Brassica napus*). Proc. Roy. Soc. Lond. B272, 1909–1916.

Messean, A., Angevin, F., Gómez-Barbero, M., Menrad, K., Rodríguez-Cerezo, E., 2006. New case studies on the coexistence of GM and non-GM crops in European agriculture. Technical Report EUR 22102 EN. Eur. Commun., 2006.

Milton, W.E.J., 1943. The buried viable-seed content of a midland calcareous clay soil. Empire J. Exp. Agric. 20, 155–167.

Pekrun, C., Hewitt, J.D.J., Lutman, P.J.W., 1998. Cultural control of volunteer oilseed rape (*Brassica napus*). J. Agric. Sci. 130, 155–163.

Preston, C.D., Pearman, D.A., Dines, T.D. (Eds.), 2002. New Atlas of the British and Irish flora. Oxford University Press.

Quinlan, J.R., 1992. Learning with continuous classes. In: Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence, World Scientific Singapore, pp. 343–348.

Roberts, H.A., Stokes, F.G., 1966. Studies on the weeds of vegetable crops. VI. Seed populations of soil under commercial cropping. J. Appl. Ecol. 3, 181–190.

Squire, G.R., 1990. The Physiology of Tropical Crop Production. CAB international, Wallingford, UK.

Squire, G.R., Brooks, D.R., Bohan, D.A., Champion, G.T., Daniels, R.E., Haughton, A.J., Hawes, C., Heard, M.S., Hill, M.O., May, M.J., Osborne, J.L., Perry, J.N., Roy, D.B., Woiwod, I.P., Firbank, L.G., 2003. On the rationale and interpretation of the farm-scale evaluations of genetically-modified herbicide-tolerant crops. Philos. Trans. Roy. Soc. Lond. B358, 1779–1800.

Stankovski, V., Debeljak, M., Bratko, I., Adamič, M., 1998. Modelling the population dynamics of red deer (*Cervus elaphus* L.) with regard to forest development. Ecol. Model. 108, 145–153.

Thirsk, J., 1997. Alternative Agriculture: A History. Oxford University Press.

Wang, Y., Witten, I.H., 1997. Induction of model trees for predicting continuous classes. In: Proceedings of the Poster Papers of the European Conference on Machine Learning, Faculty of Informatics and Statistics, University of Economics, Prague.

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