CHAPTER 21

Applications of data mining in ecological modelling

M. Debeljak & S. Džeroski
Department of Knowledge Technologies, Jozef Stefan Institute, Slovenia.

1 Introduction

Ecology is a broad discipline comprising many sub-disciplines which can be classified on the basis of complexity into the following groups: behaviour ecology (at the level of an organism), population ecology (dynamics of populations and the interactions of these populations with environmental factors), community ecology (interactions between species within an ecological community) and ecosystem ecology (energy and matter flows in the ecosystem and between ecosystem and its environment). Thus, ecology is considered a multidisciplinary scientific area that overarches life and earth sciences and plays a central role among environmental sciences. Ecology deals with the environment at different levels and scales of complexity to obtain knowledge about properties of living systems (from organism to the biosphere) and their interactions with the environment.

To overcome the problems with different levels of complexity and scales of systems under study, ecological modelling has been introduced into ecology. Based on existing knowledge about the structure, dynamics and functional relationships between the elements that are linked together into an ecological model, new knowledge about the system under study may be obtained. However, the quality of the models does not depend solely on the existing knowledge about the interrelationships between the elements of the model, but depends also on the quantity and quality of the data used for construction, calibration and validation of the model. These three activities are the most important steps in the modelling procedure [1].

As environmental concerns grow and ecology as a scientific discipline develops, more and more data on many different aspects and scales of ecology and environment are gathered. These data are of varying quality, different spatial and temporal scales and precisions, which trigger many problems for their use for solving the problem at hand. The amount of data being collected in databases today far exceeds the ability of classical ecological modelling approaches. To analyse them, we need a new generation of computational tools and theories, which were designed for extracting useful information (knowledge) from the rapidly growing volumes of digital data. These tools and the theory behind them are the subject of the field of knowledge discovery in databases (KDD). Nowadays, KDD approaches are increasingly more often used also in ecological modelling.
Knowledge discovery in databases was initially defined as the ‘non-trivial extraction of implicit, previously unknown, and potentially useful information from data’ [2]. A revised version of this definition states that ‘KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data’ [3]. According to this definition, data mining is the central step in the KDD process concerned with applying computational techniques (i.e. data mining algorithms implemented as computer programs) to find patterns in the data. The other steps in the KDD process are concerned with preparing data for data mining, as well as evaluating the discovered patterns (the results of data mining).

The above definitions of KDD contain very imprecise notions, such as pattern and knowledge. To make these (slightly) more precise, some explanations are necessary concerning data, patterns and knowledge, as well as validity, novelty, usefulness and understandability. For example, the discovered patterns should be valid on new data with some degree of certainty (typically prescribed by the user); ‘a pattern that is interesting (according to a user-imposed interest measure) and certain enough (again according to the user’s criteria) is called knowledge’ [2].

2 Data mining

Given that data mining is the central step in the process of KDD and is concerned with applying specific algorithms to find patterns in data, the notions of most direct relevance here are the notions of data, data mining algorithms and patterns.

2.1 Data

Data are a set of facts, e.g. cases in a database [3]. Most commonly, the input to a data mining algorithm is a single flat table comprising a number of records or examples (rows) and attributes (columns) (Table 1). Attributes can be of nominal or ordinal type where the latter can have continuous or discrete values. In general, each row represents an object and columns represent

<table>
<thead>
<tr>
<th>Sampling plot ID</th>
<th>Visual angle (°)</th>
<th>Appropriate wind (%)</th>
<th>Wind tunnel length (m)</th>
<th>Min. distance (m)</th>
<th>Out-crossing rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.830</td>
<td>5.82</td>
<td>2.09</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>18.263</td>
<td>6.02</td>
<td>2.22</td>
<td>67.1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>18.694</td>
<td>5.92</td>
<td>2.37</td>
<td>65.2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>19.121</td>
<td>6.12</td>
<td>2.57</td>
<td>63.5</td>
<td>0.28</td>
</tr>
<tr>
<td>5</td>
<td>19.539</td>
<td>6.68</td>
<td>2.79</td>
<td>61.7</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>19.943</td>
<td>6.94</td>
<td>3.07</td>
<td>60.1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>20.329</td>
<td>7.24</td>
<td>3.43</td>
<td>58.6</td>
<td>0.238</td>
</tr>
<tr>
<td>8</td>
<td>20.690</td>
<td>7.65</td>
<td>3.79</td>
<td>57.1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>21.020</td>
<td>7.55</td>
<td>4.10</td>
<td>55.8</td>
<td>0.386</td>
</tr>
<tr>
<td>10</td>
<td>21.312</td>
<td>8.37</td>
<td>4.31</td>
<td>54.5</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>21.557</td>
<td>8.88</td>
<td>4.49</td>
<td>53.4</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>21.748</td>
<td>8.57</td>
<td>4.58</td>
<td>52.4</td>
<td>0</td>
</tr>
</tbody>
</table>
properties of objects. When data from more than one table in a database need to be taken into account, it is left to the user to manipulate the relevant tables. Usually, this results in a single table, which is then used as input to a data mining algorithm.

2.2 Patterns

Patterns are of central importance in data mining and knowledge discovery. Data mining algorithms search for a pattern or a set of patterns that are valid in the given data. Discovered patterns that are valid, interesting and useful can be called knowledge.

A pattern in a data set is defined as a statement that describes relationships in a subset of the data set with some certainty, such that the statement is simpler (in some sense) than the enumeration of all facts in the data set [2, 3]. A pattern thus splits the data set, as it pertains to a part of it, and may involve a spatial aspect, which can be visualized. The most common types of patterns that are considered by data mining algorithms are equations, decision trees and predictive rules. Note that the same type of pattern may be generated by different data mining algorithms addressing different tasks (e.g. trees can be used for classification, regression or clustering).

2.2.1 Equations

The first group of patterns comprises predictive models in the form of linear or nonlinear equations. Equations can be algebraic or differential where the time derivatives or rates of change of system variables are expressed. In between these are difference equations, which predict the value of a system variable at a given discrete time point from the values of system variables at a previous time point. Difference and differential equations are used to model the behaviour of dynamic systems, which change their rate over time.

The simplest and most frequently used type of equations are linear equations, which predict the value of a target (dependent) variable as a linear combination of the input (independent) variables (e.g. eqns (1) and (2)). Linear equations involving two variables (such as eqn (1) with variables \(a\) and \(n\)) can be depicted as straight lines in a two-dimensional space, while the linear equations involving three variables (such as eqn (2) with variables \(a\), \(b\) and \(n\)) can be depicted as planes in a three-dimensional space. Linear equations, in general, represent hyper-planes in multi-dimensional spaces, while nonlinear equations are represented by curves, surfaces and hyper-surfaces in two-, three- or multi-dimensional spaces respectively.

\[
y = ax + n, \\
z = ax + by + n.
\]  

2.2.2 Decision trees

The second group of patterns is decision trees, which can also be used to predict the value of one or several target (dependent) variables from the values of independent variables. The decision trees are hierarchical structures where each internal node contains a test on an attribute and each branch corresponds to one outcome of the test. At the end of each branch is a leaf node, which gives a prediction for the value of the target variables (Fig. 1).

Depending on the type of the target attribute that we are aiming to predict, we distinguish among three types of decision trees. In case the target variable is discrete with a finite set of nominal values (e.g. yes, no; or species A, species B, species C), then such a decision tree is called a classification tree (Fig. 4a and b). A decision tree for predicting a real-valued target variable that can have an infinite set of values (e.g. 10, 98, 161, ...) is called a regression tree.
Some regression trees predict the target variable by using a linear model in each leaf; such trees are sometimes called model trees (Fig. 7).

### 2.2.3 Predictive rules

The third group of patterns is ‘if-then’ rules, which denote patterns of the form ‘IF Conjunction of conditions THEN Conclusion’. For predictive rules, the conclusion gives a prediction for the value of the target (dependent) variable. If we are dealing with a classification problem, the conclusion assigns one of the possible discrete values to the class, e.g. ‘Forest type=Conifers’. A rule applies to an example if the conjunction of conditions on the attributes is satisfied by the particular values of the attributes in the given example.

A single rule is rarely enough for modeling the entire problem, and so we are usually dealing with sets of rules. Such predictive rule sets can be either ordered or unordered. In unordered rule sets, the rules are considered independently and several of them may apply to a new example that we need to classify. A conflict resolution mechanism is needed if two rules, which predict different classes, apply to the same example. Ordered rule sets, on the other hand, form a so-called decision list (Fig. 8). Rules in this list are inspected from the top to the bottom of the list and the first rule that applies to a given example is used to predict its target value. In both types of rule sets, a default rule with an empty precondition typically exists, which is applied to examples that are not covered by any other rule.

### 2.3 Algorithms

The previous section described several types of patterns that can be discovered in data, whereas this section outlines data mining algorithms (the third main notion besides data and patterns) that are applied to data to find patterns valid in the data.

In general, an algorithm is a procedure given by a finite list of well-defined instructions for accomplishing some task that, given an initial state, will terminate in a defined end-state. Data mining algorithms are algorithms for finding patterns in data. These algorithms perform a search (typically heuristic) through a space of hypotheses (patterns) that explain (are valid in) the data at hand. Similarly, we can view data mining algorithms as searching, exhaustively or heuristically, a space of patterns to find interesting patterns that are valid in the given data. Many data mining algorithms come from the fields of machine learning and statistics.
Different data mining algorithms address different data mining tasks, i.e. have different intended use for the discovered patterns, which depend on the problem we would like to resolve. Three types of data mining algorithms – bivariate and multiple regression, top-down induction of decision trees (TDIDT) and rule induction – correspond to the three types of patterns described in the previous section. Note, however, that there are many other types of patterns and many other data mining algorithms [4].

2.3.1 Bivariate and multiple regression

Linear regression is the simplest form of regression [5]. Bivariate linear regression (eqn (3)) assumes that the dependent variable ($C$) can be expressed as a linear function of one attribute ($A$),

$$C = a + \beta \times A.$$  \hspace{1cm} (3)

Given a set of data, the coefficients $a$ and $\beta$ can be calculated using the method of least squares, which minimizes the sum of squared differences $\sum_i (c_i - a - \beta a_i)^2$ between the measured values for $C$ (i.e. $c_i$), and the values predicted by a linear function from the measured values for $A$ (i.e. $a_i$).

The coefficients are given as

$$\beta = \frac{\sum_i (a_i - \overline{a})(c_i - \overline{c})}{\sum_i (a_i - \overline{a})^2},$$  \hspace{1cm} (4)

$$a = \overline{c} - \beta \overline{a},$$  \hspace{1cm} (5)

where $\overline{a}$ is the average of $a_1, ..., a_n$ and $\overline{c}$ is the average of $c_1, ..., c_n$.

Multiple linear regression extends linear regression to allow use of more than one attribute (independent variable). The dependent variable can thus be expressed as a linear function of a multi-dimensional attribute vector (eqn (6)),

$$C = \sum_{i=1}^{n} \beta_i \times A_i.$$  \hspace{1cm} (6)

This simplified form assumes that the dependent variable and the independent variables have mean values of zero (which is achieved by transforming the variables – the mean value of a variable is subtracted from each measured value for that variable). Again, the method of least squares can be applied to find the coefficients $\beta_i$. If we write the equation $C = \sum_i \beta_i \times A_i$ in matrix form $\mathbf{C} = \mathbf{A} \mathbf{\beta}$, where $\mathbf{C} = (c_1, ..., c_n)$ is the vector of measured values for the dependent variable and $\mathbf{A}$ is the matrix of measured values for the independent variables, we can calculate the vector of coefficients $\mathbf{\beta}$ as (eqn (7)):

$$\mathbf{\beta} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C},$$  \hspace{1cm} (7)

where the operations of matrix transposition $\mathbf{A}^T$ and matrix inversion $\mathbf{A}^{-1}$ are used.

The use of nonlinear transformations of attributes, such as $A_i = A_i^i$, $i = 1, ..., n$, also allows nonlinear models to be found by the above approach. Such models are linear in the parameters. Note that the coefficients $a, \beta$ and $\beta_i$ are calculated directly from a formula and no search through the space of possible equations takes place. Equation discovery approaches [6], which do not assume a particular functional form, search through a space of possible functional forms and look both for an appropriate structure and coefficients of the equation. If functional forms that are not linear in the parameters are considered, partial least squares method can no longer be applied and nonlinear optimization methods need to be used.
Linear regression is normally used for predicting continuous target variables, but can also be used for predicting discrete target variables. For this purpose, generalized linear models can be used, of which logistic regression is a typical representative. The fitting of generalized linear models is currently the most frequently applied statistical technique [7].

2.3.2 Top-down induction of decision trees
Finding the optimal decision tree for a given data set is known to be computationally expensive. Heuristic search, typically greedy, is thus employed to build decision trees. The common way to induce decision trees is the so-called TDIDT [8, 9]. Tree construction proceeds recursively, starting with the entire set of training examples. At each step, an attribute is selected as the root of the (sub)tree.

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. The current training examples are split into subsets according to the values of the selected attribute. For the subsets of training examples in each branch, the tree construction algorithm is called recursively. Tree construction stops when the examples in a node are sufficiently pure (i.e. all are of the same class) or if some other stopping criterion is satisfied (e.g. no discriminative attribute can be found at that point). The final nodes are called leaves and are labelled with the corresponding values of the dependent variable.

Different measures can be used to select an attribute in the attribute selection step. These also depend on whether we are inducing classification, regression or model trees [8]. For classification and regression, [9] uses information gain, which is the expected reduction in entropy of the class value caused by knowing the value of the given attribute. Other attribute selection measures, however, such as the Gini index or the accuracy of the majority class, can be and have been used in classification tree induction. In regression model tree induction, the expected reduction in variance of the class value can be used.

An important mechanism used to prevent trees from over-fitting the 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 the leaves can be prescribed for pre-pruning and a confidence level in accuracy estimates can be prescribed for post-pruning.

2.3.3 Algorithms for rule induction
In the simplest case of concept learning, where we have two values of the target variable (i.e. two classes) one class is referred to as positive and the other as negative. Usually, only rules for the positive class are learned, while the default class is assumed to be negative. For a classification problem with several class values, a set of rules is constructed for each class value. When rules for class \( c_i \) are constructed, examples of this class are referred to as positive and examples from all the other classes as negative.

The most commonly used algorithm for rule induction is the covering algorithm [10], which works as follows. We first construct a rule that correctly classifies some examples. We then remove the positive examples covered by the rule from the training set and repeat the process until no more positive examples remain.

Within this outer loop, different approaches can be taken to find individual rules. One approach is to heuristically search the space of possible rules top-down, i.e. from general to specific (in terms of examples covered this means from rules covering many to rules covering fewer examples) [11]. To construct a single rule that classifies examples into class \( c_i \), we start with a rule with an empty antecedent (IF part) and the selected class \( c_i \) as a consequent (THEN part). The antecedent
of this rule is satisfied by all examples in the training set, and not only those of the selected class. We then progressively refine the antecedent by adding conditions to it, until only examples of class \( c_i \) satisfy the antecedent. To allow for handling imperfect data, we may construct a set of rules, which are imprecise, i.e. they do not classify all examples in the training set correctly.

Another way to construct rules is to transcribe decision trees into rules. Each leaf gives rise to a rule, with the conditions along the path from the root to the leaf forming the antecedent part of the rule and the prediction in the leaf forming the consequent. The resulting ordered set of rules is then typically post-processed (simplified). In this way, we can obtain not only classification rules but also regression (and model) rules.

3 Applications

Knowledge discovery in databases in general and data mining can be applied to a broad range of complex problems in many interdisciplinary fields. In ecology and environmental sciences, data mining is applied side by side with classical statistical methods. In this section, we give an overview of data mining applications in ecology, complemented with case studies. The structure of this section follows the previous two, where equations, decision trees and predictive rules, as the main three types of patterns, which we are searching in data sets, are following each other.

3.1 Equations

3.1.1 Difference equations

Black alder (\textit{Alnus glutinosa} (L.) Gaertn.) wetland forests are exposed to a decrease in the groundwater table due to hydromeliorations throughout Europe. To understand the effects on growth processes, a simulation model of annual radial increments of black alder trees has been constructed on the basis of field data from north-eastern Slovenia [12]. The model was induced with the data mining algorithm Constrained Induction of Polynomial Equations for Regression (CIPER), using data that describe site conditions and applied management measures in the last 35 years of a black alder stand life. CIPER [13, 14] is an algorithm for induction of polynomial equations from the data. The model constructed by CIPER is in the form of a difference polynomial equation (eqn (8)) and its performance is given in Fig. 2, where both measured and

![Graph showing measured and modelled radial increments of black alder trees]

Figure 2: Measured (\( r \)-incr measured) and modelled (\( r \)-incr modelled) radial increments of black alder trees [12].
modelled radial increments are plotted. The correlation coefficient between the modelled radial increments and the measured increments of sample trees is 0.877.

It has been shown that the groundwater levels (minL_4-7, minL_8-10) and the duration of solar radiation (t_sun_4_7, t_sun_8_10) have the highest effect on the annual radial increments. The radial increments were the lowest in very wet and cloudy years. However, changes in the groundwater level and the duration of solar radiation, as well as the increased oscillations of groundwater level, have been found to act as environmental stressors on the black alder stand.

Radial increment (mm) = \([-0.051\times\text{minL}_8-10] + [-0.029\times\text{maxL}_8-10]\]
+ \([-0.0175\times\text{t-sun}_4-7] + [0.035\times\text{t-sun}_8-10]\]
+ \([-1.950E-05\times\text{t-sun}_8-10^2] + [-2.010\times\text{d-wf}_4-7]\]
+ \([9.356-05\times\text{minL}_4-7\times\text{t-sun}_4-7] + [-0.0002\times\text{minL}_4-7\times\text{t-sun}_8-10]\]
+ \([6.456E-05\times\text{minL}_8-10\times\text{t-sun}_8-10] + [3.065E-05\times\text{maxL}_8-10\times\text{t-sun}_4-7]\]
+ \([0.003\times\text{t-sun}_4-7\times\text{d-wf}_4-7] + [-0.001\times\text{t-sun}_8-10\times\text{d-wf}_4-7]\) + 7.911. (8)

3.1.2 Differential equations

Besides the induction of algebraic difference equations, as presented in the previous example, data mining can be applied to construct models, which have the form of differential equations. The Lagrange machine learning system [15] joins two ideas in automated modelling, i.e. compositional modelling and model induction from data. Compositional modelling builds models by assembling model fragments, typically from a library of model fragments, provided by domain experts into an adequate model. In contrast, induction methods usually tackle the same task without incorporating domain expert knowledge in the procedure for model construction.

An application of this technique is demonstrated in the study of the eutrophication process in lake Bled, Slovenia [16]. The data set comprises long-term measurements (from 1995 to 2002) of physical, chemical and biological data in lake Bled. Given the expert knowledge in terms of a simple food web concept and rules for modelling thereof, a model for long-term dynamics of the phytoplankton (data from 1995 to 2001) and a three-equation model for the annual cycle (nutrient (eqn (9)) – phytoplankton (eqn (10)) – zooplankton (eqn (11)) were induced (Fig. 3).

\[
\frac{d(ps)}{dt} = \text{ps}_{krivica} \cdot \frac{q - krivica}{7 \times 10^6} + \text{ps}_{misca} \cdot \frac{q - misca}{7 \times 10^6}
+ \text{ps}_{radovna} \cdot \frac{q - radovna}{7 \times 10^6} - \text{ps} \cdot \frac{q - jezernica}{7 \times 10^6} - \text{ps} \cdot \frac{q - natega}{7 \times 10^6}
+ 0.0022 \cdot \text{phyto}^2 \cdot 0.072 \cdot \text{temp} - 2.7 \cdot 20.4 - 2.7 + 0.07 \cdot \text{daph} \cdot 0.0026 \cdot \text{temp} \cdot 12.3
- 0.0023 \cdot \text{phyto} \cdot 0.21 \cdot \frac{\text{ps}}{\text{ps} + 0.0042} \cdot \text{temp} \cdot \text{light} \cdot \left(1 + \frac{\text{light}}{170} \right) \tag{9}
\]
3.2 Decision trees

3.2.1 Classification trees

In the example given in this section, we show the application of classification trees for modelling habitat suitability for the brown bear population in Slovenia [17]. Classification trees were induced from data about explanatory variables of several GIS layers (i.e. land cover data, forest inventory data, settlements map, road map, digital elevation model) and data on bear sightings acquired throughout Slovenia by the Hunters association and partially by radio-tracking of some bears. The modelling process takes place in the context of a raster GIS with a 500 × 500 m spatial resolution. The target variable is binary and can take the value suitable or unsuitable for each unit of spatial resolution.
Optimal habitat was defined by the presence of female bears with cubs, while maximum habitat was defined by the presence of either male or/and female on the particular location. The ‘optimal’ decision tree (Fig. 4a) accounted for differences in surrounding forest matrix size, forest abundance in each pixel, predominant land cover type, sub-regional density of human population and predominant forest association within each forest pixel. The ‘maximal’ decision tree (Fig. 4b) was much simpler, accounting only for predominant land cover type, forest abundance and proximity to settlement. The accuracies of the models are 89% for the former and 85% for the latter.

The optimal habitat covers 12.3% of Slovenia’s territory, mostly in the southern part, bordering Croatia. The possible maximal habitat extent includes an additional 26.4% of the territory, mostly in the alpine region in the northern and western part of Slovenia, thus totalling 38.7% of the country (Fig. 5). The only region completely unsuitable for bears seems to be the Pannonian region in the eastern part of the country.

### 3.2.2 Regression trees

An application of regression trees is presented with the study of presence and abundance of volunteer oilseed rape (OSR) (*Brassica napus*) in the seedbank at 257 arable fields in the UK. The data come from the baseline sampling in the UK’s Farm Scale Evaluation of genetically modified herbicide-tolerant crops [18]. Data about the seedbanks and site characteristics were collected during 2000, 2001 and 2002 [19, 20]. Our study was looking for previously unfound correlations between OSR abundance in soil seedbank, total soil seedbank and several other factors, notably the per cent of nitrogen and per cent of carbon in the soil. All of these were the smallest in the

![Figure 4: The ‘optimal’ classification tree (a) (branches indicated with * are truncated) and the ‘maximal’ classification tree (b) [17].](image-url)
centre of arable production in southern England and the greatest in the surrounding southwest, west and north. It has been shown that the density of OSR seeds per square metre depends on the per cent of nitrogen in the soil and the time since the last OSR was sown on the field for which the prediction of density of OSR in the soil seedbank was estimated (Fig. 6).

3.2.3 Model trees
We used model trees to analyse the population density of springtails (Order Collembola) and earthworms (Class Oligochaeta) in the fields where genetically modified (GM) Bt maize was

![Figure 5: Potential naturally suitable habitat for the brown bear in Slovenia [17].](image)

![Figure 6: Regression tree for predicting the density of oil seed rape seeds in the soil seedbank ('Nitrogen': % of nitrogen in the soil, 'last OSRcrop': years since the last oil seed rape was sown on the field at hand) [18].](image)
grown [21]. Samples were collected during two consecutive years at a study site located in Foulum (Denmark). Samples were taken at the beginning and at the end of the maize crop-growing season.

Farming practices, soil parameters, the biological structure of soil communities and the type and age of the crop at the sampling dates were used as attributes to predict the total abundance of springtails and biomass of earthworms. Total abundance and biomass were predicted for functional groups of earthworms and springtails. For earthworms, the epigeic, endogeic and anecic groups, while for springtails the eu-, eu to hemi-, hemi-, hemi to epi- and epiedaphic groups were considered.

Here, we give an example model tree, which predicts the abundance (number per kg of soil) of hemi-epiedaphic springtails ($r^2 = 0.59$) (Fig. 7). The model shows increased springtail abundance at the end of the maize-growing season, while higher organic matter content and higher pH tend to increase their biomass in spring. A greater abundance of springtails was also noted in the early autumn if the crop was non-Bt maize. However, the research has not found any effects of the Bt maize crop system on the studied functional groups of soil fauna.

3.3 Predictive rules for regression

An application of predictive rules for regression is a part of the study of gene flow between cropped and volunteer OSR [22]. The adventitious presence of GM varieties of crops in non-GM seeds and crops has become an issue in the field of crop ecology and poses the problem of their co-existence with conventional and organic crops. To address this issue, the GENESYS model for simulation of gene flow between cropped and volunteer OSR was developed [23, 24]. Using a large risk field pattern (small narrow and long rectangular central field surrounded by 34 large neighbour fields), a simulation of gene flow was performed with GENESYS for the period of 25 years. For each year, the crops and the cultivation techniques were chosen randomly, as well as the genetic variables. The only exception was the crop grown in the central, the most exposed field, which was always a non-GM OSR.

The goal of the research presented here was to find rules of the adventitious presence of GM OSR seeds in the central recipient field, which have been derived from input and output data of the simulation model GENESYS. The rule-based model was constructed with CUBIST.
Applications of Data Mining in Ecological Modelling

421

CUBIST builds rule-based regression models containing one or more rules, where each rule is a conjunction of conditions associated with a linear expression. The meaning of a rule is that, if a case satisfies all the conditions, then the linear expression is appropriate for predicting the target value. A CUBIST model resembles a piecewise linear model, except that the models can overlap.

Based on the 100,000 GENESYS simulations, each presenting 1899 input attributes that describe the field plan, crops grown, cultivation techniques and genetic variables for each of the 35 fields for the last 4 years, the rule-based model presented in Fig. 8 was induced. The correlation coefficient of the model was 0.77.

The model shows that the sowing date and the density of the analysed conventional rape crop, and those of the previous crops were the overall dominant factors. Irrespective of the crop, the later the sowing date, the lower the rate of adventitious presence of GM OSR in the crop. On the other hand, the higher the sowing density, the lower the rate, except if the crop is a GM OSR. In that case, the rate of adventitious presence increases with the increase in sowing density.

4 Conclusions

This chapter introduced data mining, the central activity in the process of KDD, which is concerned with finding patterns in data. It also gave an overview of data mining applications in ecological modelling, complemented with a sample of case studies, which were conducted by the authors.

On the basis of the experience with the application of data mining in ecological studies, we may conclude that this approach can help understand the studied domain better, that it enables to obtain new knowledge about studied phenomena and that it permits making reliable predictions and classifications, which can be further used in decision support for environmental management.
Data mining also helps automate the data analysis process and gives more flexibility than the classical statistical approach. It uses richer hypothesis language (e.g. piece-wise linear model vs. linear models). It can also use background knowledge and trade-off between the quality/quantity of data and background knowledge, which is a unique characteristic of this approach to data analysis.

However, data mining also has some limitations, which need to be respected. It cannot overcome the informal law of information conservation (garbage in, garbage out). The knowledge that we are seeking to discover has to come from the combination of data and background knowledge, and if we have very little data of very low quality and no background knowledge, no form of data analysis will help.

Regardless of these pro-et-contra arguments, applying data mining has also some positive side effects such as discovering problems with the data during data analysis (identification of missing data, erroneous data and inappropriately measured variables). Another very positive side effect is the identification of new problems that can be addressed by this approach and some recommendations on what data to collect and how to collect them can be specified as well.

We may conclude that data mining is a powerful tool for data analysis and knowledge discovery and application of these techniques can help understand and solve environmental problems better.

References

Applications of Data Mining in Ecological Modelling


