CHAPTER TWELVE

Data Mining for Environmental Systems

K. Gibert[a], J. Spate[b], M. Sànchez-Marrè[c], I. Athanasiadis[d], *and* J. Comas[e]

Contents

12.1. Introduction

Environmental systems (ES) typically contain many interrelated components and processes, which may be biological, physical, geological, climatic, chemical,

[a] Dep. Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Ed. C5, Campus Nord, C. Jordi Girona 1-3, Barcelona 08034, Spain

[b] 110 Wyndham Cres, Canton CF11 9EG, UK

[c] Knowledge Engineering & Machine Learning Group (KEMLG), Computer Software Department (LSI), Technical University of Catalonia (UPC), Campus Nord-Building Omega, Office 134, Jordi Girona 1-3, 08034 Barcelona, Catalonia, Spain

[d] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Galleria 2, CH-6928 Manno, Lugano, Switzerland

[e] Laboratory of Chemical and Environmental Engineering (LEQUIA), Faculty of Sciences, University of Girona, Campus Montilivi, s/n, postal code E-17071, Girona, Catalonia, Spain

Environmental Modeling and Software, 3
1574-101X.

or social. Whenever we attempt to analyse ES and associated problems, we are immediately confronted with complexity stemming from various sources:

- *Multidisciplinarity*: a variety of technical, economical, ecological and social factors are at play. Integration of knowledge as well as the use of analysis techniques from different social and scientific disciplines is necessary for proper treatment.
- *Ill-structured and non-linear domain*: ES are poor or ill-structured domains and they can be stochastic. That is, they are difficult to clearly formulate with a mathematical theory or deterministic model due to their high complexity, involving processes which are not yet well known. Many interactions between animal, plant, human and climatic system components are highly non-linear. Solutions to problems are often not unique.
- *High dimensionality and multiscalarity*: most environmental processes take place in two or three spatial dimensions, and may also involve a time component. Within this frame, multiple factors are acting at many different spatial and temporal scales (see Section 12.2.9).
- *Heterogeneity of data*: environmental real world systems are characterised by an abundance of heterogeneous data, from numerous sources, with different formats, resolutions and qualities. Qualitative and subjective information is often very relevant.
- *Intrinsic non-stationarity*: ES are in general not static, but evolve over time. The assumption of stationarity cannot be justified since ES and problems deal with many physical, chemical and biological phenomena that change over time (Guariso and Werthner, ?).
- *Stiff systems*: some states in environmental processes change fast while others change slowly. This means that cause and effect are not always closely related in time and space and that sometimes delays in responses make management more difficult.
- *Controllability*: controllability of ES is poor, due the unavailability of actuators (Olsson, ?).
- *Uncertainty and imprecise information*: because environmental data collection is often expensive and difficult, measurement error is often large, and spatial and temporal sampling may not fully capture system behaviour. Records may also contain outliers, missing values and highly uncertain information. See Section 12.2.1.
- *Cyclic behaviour*: ES have no well-defined beginning or end, but they are characterised by loops and continuously affected by feedbacks and other environmental properties.

All these features may be emphasised when natural systems are affected or altered by human activity. Because the consequences of an environmental system changing behaviour or operating under abnormal conditions may be severe, there is a great need for data analysis, modelling of ES and development of decision support systems in order to improve the understanding of ES behaviour and the management of associated complex problems (especially under abnormal situations).

The special features of environmental processes demand a new paradigm to improve analysis and consequently management. Approaches beyond straightforward application of conventional classical techniques are needed to meet the challenge of environmental system investigation. In some cases large quantities of data are available, but as the effort required to analyse the large masses of data generated by ES is large, much of it is not examined in depth and the information content remains unexploited. In this sense, the Knowledge Discovery from Databases (KDD) process and several Data Mining (DM) techniques have demonstrated that they can provide successful tools to deal with this complexity (see Section 12.2 for some examples). DM techniques provide efficient tools to extract useful information and discover knowledge from large databases, and are equipped to identify and capture the key parameters controlling these complex ES.

In 1989, the first *Workshop on Knowledge Discovery from Data* (*KDD*) was held. Seven years later, in the proceedings of the first *International Conference on KDD*, Fayyad gave one of the most well known definitions of what is termed *Knowledge Discovery from Data*:

> "*The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*" (Fayyad et al., 1996b).

KDD quickly gained strength as an interdisciplinary research field where a combination of advanced techniques from Statistics, Artificial Intelligence, Information Systems and Visualisation and new algorithms are used to face the task of knowledge acquisition from huge data bases. The term *Knowledge Discovery from Data* appeared in 1989 referring to high level applications which include particular methods of *Data Mining*:

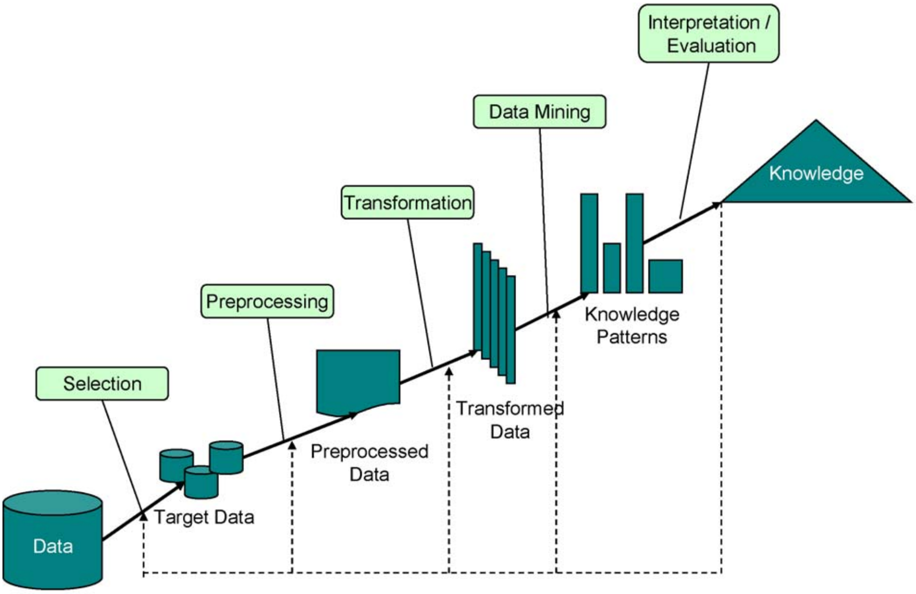
> "[. . .] *overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms.*"

Thus, KDD is the high level process of combining DM methods with different tools for extracting *knowledge* from data. The basic steps established by Fayyad are shown in Figure 12.1 and details of the different techniques involved in this process are provided in Fayyad et al. (1996a). Fayyad's proposal marked the beginning of a new paradigm in KDD research:

> "*Most previous work on KDD has focused on* [. . .] *DM step. However, the other steps are of considerable importance for the successful application of KDD in practice.*"

Fayyad's proposal included prior and posterior analysis tasks as well as the application of DM algorithms. These may in fact require great effort when dealing with real applications. Data cleaning, transformation, selection of DM techniques and optimisation of parameters (if required) are often time consuming and difficult, mainly because the approaches taken should be tailored to each specific application, and human interaction is required. Once those tasks have been accomplished, the application of DM algorithms becomes trivial and can be automated, requiring only a small proportion of the time devoted to the whole KDD process. Interpretation of results is also often time consuming and requires much human guidance.

It is convenient to remark here that in some scientific contexts, ES among them, the term *Data Mining* (DM) refers to the whole KDD process (Siebes, 1996), and not only to the application to a cleaned dataset. It is clear that either referring to



- Understanding goals
- Data Selection
- Data cleaning and preprocessing
- Data reduction and projection
- Choosing the DM task
- Selecting the DM algorithm/s and parameters
- Data mining
- Interpreting mined patterns
- Reporting and using discovered knowledge

Figure 12.1 Outline of the Knowledge Discovery from Data process.

the knowledge discovery process as *KDD* or simply as *DM*, tasks like data cleaning, variable selection, interpretation of results, and even the reporting phase are of as much importance as the data analysis stage itself. This is particularly true when dealing with environmental data which, due to the special features of ES mentioned above, may need special treatment from any modelling scheme. Some of the particular requirements/needs implied by environmental data, which make the KDD process and the application of DM techniques specifically appealing, include:

- Supporting systematic and objective exploration (data preprocessing) and visualisation of data. For this purpose, inductive techniques are an alternative for several activities of the environmental scientist, when analytical/traditional methods fail, are too slow, or simply do not exist. An example would be extracting maximum amount of useful information from on-line and heterogeneous large databases. These activities are “per se” very important although they used to be preparatory for an environmental software system development.

- Modelling and systems analysis activities: knowledge acquisition step to extract more meaningful knowledge and better models for simulation and prediction of ES. This will also contribute to build more reliable intelligent environmental decision support systems (Chapter 8; Poch et al., 2004).
- Discovering of knowledge contained in large time series.
- Facilitating the integration of different knowledge sources and expertise, and the involvement of end-user (domain expert) criteria and stakeholder points of view in algorithm design and result interpretation.
- Facilitating the sharing and rapid re-use of data and extracted technical knowledge and experiences among domain experts.
- Selecting relevant features.
- Treating outliers, missing data and uncertainty of environmental data.
- Supporting transparency in the communication of environmental data to the general population and decision-making processes.

These are some of the most important contributions where DM techniques can help environmental scientists/managers in solving real-world problems. This chapter attempts to show that DM techniques are valuable tools that could be used to good effect in the environmental and natural resource science field, aims to introduce the main concepts of DM and foster discussion of the ways in which it could be used and encouraged in ES problems.

As mentioned above, most DM techniques have not found widespread application in environmental science and management. A small number of research groups focus on applying artificial intelligence and/or DM to ES. Most notable, the BESAI (Binding Environmental Science and Artificial Intelligence, http://www.lsi.upc.edu/webia/besai/besai.html) working group has organised several international workshops as part of prestigious AI conferences (ECAI, IJCAI, AAAI), with contributions addressing DM techniques. Also, from 1998, BESAI has organised three special sessions devoted to ES and Artificial Intelligence during the iEMSs biennial conferences (2002–2006). The European Network of Excellence on KDD (KDnet, http://www.kdnet.org) has organised a workshop on KD for Environmental Management (Voss et al., 2004). In addition, four international conferences on Applications of Machine Learning to Ecological Modelling (Recknagel, 2001) have been held over the period 1997–2004, producing some of the papers discussed in this chapter.

Selected algorithms are discussed in Section 12.2, along with preprocessing methods, which will be explored in Sections 12.2.1 to 12.2.3. Also, a brief review of previous environmental DM work is given. Later in the chapter, concerns such as performance evaluation, model optimisation and validation, and dealing with disparate data sources, are addressed. Good data mining practice guidelines are outlined in Section 12.3. Available software is discussed in Section 12.4, with particular reference to the Weka (Whitten and Frank, 1991) and GESCONDA (Sànchez-Marrè et al., 2004) packages. And in Section 12.5 some conclusions and challenges for using DM techniques in environmental problem solving are highlighted.

12.2. Data Mining Techniques

Here we shall introduce a variety of DM techniques: clustering (Section 12.2.4), classification (Section 12.2.5), association rule extraction (Section 12.2.6), artificial neural networks (Section 12.2.7), and other techniques (Section 12.2.8), as well as preprocessing and other data issues. Of course, we cannot hope to detail all DM tools in a short paper. An extensive review of DM tools for environmental science is given in Spate and Jakeman (under review), and references to specific papers are given throughout the text. Key reading material introducing the reader to essential points of KDD are Han and Kamber (2001), Whitten and Frank (1991), Hastie et al. (2001), Larose (2004) and Parr Rud (2001). The techniques listed below are some of the most common and useful. For each DM technique, a brief introduction is given, followed by several applications to environmental data. Preprocessing and visualisation are also included in this section, as they are essential components of the KDD process.

12.2.1 Preprocessing: data cleaning, outlier detection, missing value treatment, transformation and creation of variables

Sometimes, a number of cells are missing from the data matrix. These cells may be marked as a ⋆, ?, NaN (Not a Number), blank space or other special characters or special numeric codes such as 99,999. The latter can induce grave mistakes in calculations if not properly treated. It is also important to distinguish between random and non-random missing values (Allison, 2002; Little and Rubin, 1987). Non-random missing values are produced by identifiable causes that will determine the proper treatment, also influenced by the goals of the task. Inputation (see Rubin, 1987) is a complex process for converting missing data into useful data using estimation techniques. It is important to avoid false assumptions when considering inputation methods, which may have a significant effect on the results extracted. All the methods have pros and cons, and the choice must be made with care. In particular, removing rows with missing cells from a dataset may cause serious problems if the missing values are not randomly distributed. It is of utmost importance to report any elimination performed.

Outliers are objects with very extreme values in one or more variables (Barnett and Lewis, 1978). Graphical techniques were once the most common method for identifying them, but increases in database sizes and dimensions have led to a variety of automated techniques. The use of standard deviations is possible when and only when considering a single variable that has a symmetric distribution, but outliers may also take the form of unusual combinations of two or more variables. The data point should be analysed as a whole to understand the nature of the outlier.

The treatment will depend on the nature of the outlier (error, member of another population, intrinsic extreme value, etc.). The influence of outliers can dramatically disturb the results or certain methods, a concern which should feature in the choice of tools used throughout the rest of the process. See Moore and Mc-

Cabe (1993) for an interesting discussion on the dangers of eliminating rows with outliers:

> "*In 1985 British scientists reported a hole in the ozone layer of the Earth's atmosphere over the South Pole.* [...] *The British report was at first disregarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software* [...] *had automatically suppressed these values as erroneous outliers! Readings dating back to 1979 were reanalysed and showed a large and growing hole in the ozone layer* [...] *suppressing an outlier without investigating it can keep valuable information out of the sight*" (Moore and McCabe, 1993).

Sometimes transformation of variables may assist analysis. For example, normality may be forced when using ANOVA or, for ease of interpretation, variables with a large number of categorical labels can be grouped according to expert knowledge. Under some circumstances, discretisation of continuous variables is appropriate (e.g. *Age* into *Child* under 18 years, *Adult* between 18 and 65 years, *Elderly* over 65 years).

Noise is often a critical issue, and especially with environmental data some bias may exist that can be removed with a filter. Transformations should always be justified and documented, and the biases that may be introduced noted (Gibert and Sonicki, 1999). Interpretability of transformed variables should be kept.

Creation of additional variables is also used in KDD. Here, expert knowledge is usually the guide. Exploratory variable creation without such assistance is almost always prohibitively time consuming, and as noted in Section 12.2.2, may obfuscate physical interpretation and exacerbate noise. Efficient techniques for data reduction, however, do exist and are well used.

12.2.2 Data reduction and projection

When the number of variables is too high to deal with in a reasonable way, which is not unusual in a data mining context, a data reduction method can be applied. This may be accomplished by eliminating some variables wholesale, or projecting the feature space of the original problem into a reduced fictitious space, with fewer dimensions. Principal Component Analysis (PCA) (Dillon and Goldstein, 1984) is one of the best known techniques used for the latter purpose. Each principal component is a linear combination of the original variables, and the aim is to work with a reduced set of these, such that the loss of information is not relevant. Interpretation of the new variables may be lost.

Regarding the former method, datasets may contain irrelevant or redundant variables (Spate et al., 2003). Automated techniques for identifying and removing unhelpful, redundant or even contradictory variables usually take one of two forms: statistical examination of the relevance of candidate variables, or searching for the best combination of attributes in terms of model performance. The former are called *filters* and the latter *wrappers* (see Hall, 1999 for details). For a survey of common attribute selection techniques, see Molina et al. (2002).

Other techniques are based on feature weighting (see for example Aha, 1998, and Núñez et al., 2003), which is a more general and flexible approach than feature selection. The aim is to assign a degree of relevance (a weight) to each attribute. Similarities (or dissimilarities) become emphasised according to the relevance of the attribute, and irrelevant attributes will not influence the results, so quality of inductive learning improves.

12.2.3 Visualisation

Visualisation is a powerful strategy for leveraging the visual orientation of sighted human beings. Sighted humans are extraordinarily good at recognising visual patterns, trends and anomalies; these skills are valuable at all stages of the KDD (Miller, 2007). For example, the presence of outliers, missing values, or errors are typical pre- and post-processing KDD tasks where visualisation techniques can be valuable.

Graphs commonly used for classical exploratory visualisation, like boxplots, histograms, time series plots or two-dimensional scatter plots, may perform poorly considering the great number of variables involved in environmental datasets, along with their complex interrelations, and spatiotemporal references. Thus, more sophisticated visualisation methods are required, as for example:

- distributional plots,
- three-, four-, and five-dimensional plots (colour and symbols may be used to represent the higher dimensions),
- dimension scaling, for example log scales,
- rotatable frames,
- animation with time and interactive graphs,
- geo-referenced visualisations and maps.

Most DM packages, such as Weka, include visualisation tools, while more advanced features are provided with widespread tools such as Matlab or a dedicated data language such as IDL or the CommonGIS tool (Andrienko and Andrienko, 2004). The reader is also directed to dedicated visualisation tools such as XGobi (Swayne et al., 1998). Visual representations are extremely effective, and may convey knowledge far better than numerical or analytical forms. They should always be considered in environmental KDD.

12.2.4 Clustering and density estimation

Clustering techniques cover an exploratory goal rather than a predictive one. They are used to divide a data set into groups, being suitable for discovering the underlying structure of the target domain, if this is unknown. Thus, they belong to the group of *unsupervised learners*. They identify distinct groups of homogeneous objects (according to some criteria) that can be considered together, which is very useful in the DM context where the number of cases to be analysed can be huge. Appropriate choice of criteria (distance, dissimilarity, logics or mixtures) for comparing objects is essential (Gibert et al., 2005a), and different measures will result in different clustering algorithms, a point which is discussed in detail in Spate (2006), Gibert et al. (2005b), Núñez et al. (2004) and Dubes and Jain (1988).

Clustering can also be viewed as a density estimation problem by assuming that the data was generated by a mixture of probability distributions, one for each cluster (e.g. Witten and Frank, **?**). A standard approach is to assume that the data within each cluster is normally distributed. Treating clustering as a density estimation problem makes it possible to objectively evaluate the model's goodness-of-fit, for example by computing the likelihood of a separate test set based on the mixture model inferred from the training data.

Applications: The use of clustering algorithms has been reported in various application fields; for dimensionality reduction in streamflow timeseries (Zoppou et al., 2002; Sanborn and Bledsoe, 2005), wastewater treatment plants (Sànchez-Marrè et al., 1997), cyclone path identification (Camargo et al., 2004), and water sample grouping based on chemical composition (Ter Braak et al., 2003).

12.2.5 Classification and regression methods

In classification and regression, the identity of the target class is known *a priori* and the goal is to find those variables that best explain the value of this target, either for descriptive purposes or prediction of the class value of a new data point. They are an example of *supervised* learning methods. A popular and accessible classification model is a decision tree, like the C4.5 method.

Classical linear regression is a technique for finding the best linear equation defining the relationship between a numerical response and the independent variables, all of which should also be numerical (Draper and Smith, 1998). It is mainly used for prediction of the target variable, but also for identifying which variables have the strongest influence on the behaviour of the response variable. In that sense it is useful for descriptive purposes. Regression is suitable under normality, homoscedasticity and independence of the regressors. Other models are to be used if those conditions do not hold (e.g. ANOVA, logistic regression, non-linear models, neural networks).

Case-Based Reasoning (CBR) is a general problem-solving, reasoning and learning paradigm (see Kolodner, 1993) within the artificial intelligence field. When CBR is used as a classification method, the assumption is that similar cases should have similar classifications: given a new case, similar cases are selected from the case library, and the new case is classified according to the classifier values of those neighbours. Quality of the case library is critical, as is the choice of an appropriate measure of similarity (see Núñez et al., 2004). It must also be noted that CBR does not produce an explicit model describing system behaviour.

Rule induction or rule extraction is the process of discovering rules that summarise common or frequent trends within a dataset (i.e. which variables and values are frequently associated). Classification rules can be induced from labelled examples. Some rule extraction routines can combine numerical and categorical data, and a time component can also be introduced into the rule format.

Applications: Classification techniques seem to be very popular. For example, in Spate (2002) and Spate et al. (2003) rainfall intensity information was extracted from daily climate data; Ekasingh et al. (2003) discusses the classification of farmers' cropping choices using decision trees; in Sweeney et al. (2007) mosquito population sites are categorised, while in Stadler et al. (2006) decision trees are applied in a European life-history trait database. Agriculture-related applications include Holmes et al. (1998) for apple bruising, Cunningham and Holmes (1999) for mushroom grading, Yeates and Thomson (1996) for bull castration and venison carcass analysis and the Michalski and Chilausky's (1980) soybean disease diagnosis work, which is a classic benchmark problem in machine learning. Considerable efforts are recorded in water-related fields, using rule-based reasoning (Zhu and Simpson, 1996; Riaño, 1998; Dzeroski et al., 1997; Comas et al., 2003, and Spate, 2005), Case-Based Reasoning (Rodríguez-Roda et al., 1999; Wong et al., 2007), regression trees (Dzeroski and Drumm, 2003) or hybrid techniques (Rodríguez-Roda et al., 2001, 2002, and Cortés et al., 2002). In the study of air quality, classification has been used for air quality data assurance issues (Athanasiadis and Mitkas, 2007) and the operational estimation of pollutant concentrations (Kaburlasos et al., **?**).

Classification has also found spatial applications. For example, fish distribution (Su et al., 2004) and soil erosion patterns (Ellis, 1996) have both been modelled with classification methods, as was soil erosion, and other soil properties in McKenzie and Ryan (1999), which also used regression trees and other techniques with a view to obtaining system information.

Comas et al. (2001) discusses the performance of several DM techniques (decision tree creation, two types of rule induction, and instance-based learning) to identify patterns from environmental data. And a comparison between statistical and classification algorithms in air quality forecasting has shown the potential of DM techniques (Athanasiadis et al., 2005).

12.2.6 Association analysis

Association analysis is the process of discovering and processing interesting relations from a dataset. The concept was originally developed for supermarket analysis, where the aim is to discover which items are frequently bought together. They have the advantage that as no initial structure is specified, the results may contain rules that are highly unexpected and which would never have been specifically searched for, because they are inherently surprising. The format summarising only frequently occurring patterns can also be useful for anomaly detection, because those data points violating rules that usually hold are easy to identify and may be examples of interesting behaviour.

Rule extraction algorithms, for both association and classification, tend to fall into two broad categories. There are those built by generalising very specific rules until they cover a certain number of instances; for example the AQ family of algorithms described in Wnek and Michalski (1991), or the RISE algorithm by Domingos (1996). And there are those that begin with a broad rule covering all or a large fraction of the data and refine that rule until a sufficient level of precision is achieved, such as the PRISM (Cendrowska, 1998) and RIPPER (Cohen, 1995) algorithms. For obvious reasons, the specific to general variety are for the most part classification rule learners. Many rule extraction algorithms are extremely fast, and can thus be applied to very large databases in their entirety. They may be used either for predictive purposes or for system investigation.

12.2.7 Artificial neural networks

We have mentioned above that DM techniques are not widely used in the area of environmental modelling and software applications. An exemption is the adoption of Artificial Neural Networks (ANNs), which have become an accepted part of the environmental modelling toolbox. The basis of the ANN methodology is to emulate the behaviour of human neural networks. Human neural networks make synaptic connections among many neurons, producing several output signals. ANNs activate some neuron output by means of weighted combination of values of some other neurons, and using an output activating function. Thus, the main role of ANNs is as an approximation function, especially suited for predicting non-linear functions.

Applications: Numerous applications have been developed, and as an indication we refer to the works of Kralisch et al. (2001) and Almasri and Kaluarachchi (2005) on nitrogen loading, Mas et al. (2004) on deforestation, Babovic (2005) on hydrology, those of Belanche et al. (2001), Gibbs et al. (2003) and Gatts et al. (2005) on water quality, or the discussion on non-linear ordination and visualisation of ecological data by Kohonen networks, and ecological time-series modelling by recurrent networks (Recknagel, 2001), along with the recent application of Dixon et al. (2007) to anaerobic wastewater treatment processes.

12.2.8 Other techniques

DM and machine learning are of course not restricted to the methods discussed here, and some less common techniques have been applied to environmental problems. In Robertson et al. (2003), Hidden Markov Models were used to model rainfall patterns over Brazil with interesting results, and Mora-López and Conejo (1998) applied Qualitative Reasoning to meteorological problems. Cloud screening for meteorological purposes was also investigated with Markov Random Fields in Cadez and Smyth (1999). Sudden death of oak trees was modelled with Support Vector Machines in Guo et al. (2005). In Comas et al. (2001), several Instance-Based and Inductive Methods were applied in wastewater treatment plants. Vellido et al. (2007) present Generative Topographic Mapping to investigate river ecology.

12.2.9 Spatial and temporal aspects of environmental data mining

Spatiotemporal DM does not simply deal with data situated in time and space, rather it has a broader goal which is to identify knowledge models (i.e. patterns) with spatial and temporal references. Roddick and Spiliopoulou (?) identified that temporal DM research deals with three “time-stamped datatypes”: conventional scalar values, events and the DM results (patterns). Similarly, in spatiotemporal DM these three datatypes can be considered to be referenced in both time and space.

Temporal relationships include “before-after” relations, while spatial relationships deal either with metric (distance) relations, or with non-metric topology relations. The classical DM methods do not consider those particularities, while KDD algorithms require to be altered for accommodating spatiotemporal data dimensions. In this respect, spatiotemporal DM is potentially useful for a variety of tasks, including:

(1) spatiotemporal pattern identification (as in pattern analysis, neighbourhood analysis),
(2) data segmentation and clustering (spatiotemporal classification),
(3) dependency analysis, correlation analysis and fault detection in data (outlier detections, surprising pattern identification),
(4) trend discovery and sequence mining (as in regression analysis and time series prediction).

However promising it might seem, spatiotemporal mining suffers from scale effects. In general, spatial resolution or time granularity affects the support of the extracted patterns. In this respect, visualising data and extracted patterns, by employing maps and GIS technology (e.g. see Andrienco, **?**) could prove valuable. Also, recently Case-Based Reasoning has been shown as a promising modelling tool for spatiotemporal databases, such as in the work by Sànchez-Marrè et al. (2005) and Martín and Plaza (2004).

12.3. Guidelines for Good Data Mining Practice

As with all modelling paradigms, good practice modelling involves far more than applying a single algorithm or technique (see Chapter 2). Each of the DM steps detailed in Section 12.2.1 must be followed with due attention. In this section, we record a few notes and considerations that may be of use to those contemplating the use of DM in ES. In this section general recommendations on the most suitable use of the presented techniques are provided, together with a set of considerations to ensure good practice in real applications.

- Devote all required time to a proper preprocessing.

  Input Data Uncertainty: Environmental data is subject to high uncertainty. Tracking and reporting of uncertainties related to measurement and other sources of noise is another area that is sometimes not treated rigorously, despite the implications. Therefore, the minimum theoretically achievable error of any model built on the data cannot be less than the error contained in the original data. Models with reported fit greater than this are overfitted and their performance measures do not reflect true predictive capacity. In general, where there is more data, there is less uncertainty, or at least that uncertainty can be better quantified.

  Cautions on data cleaning: Data cleaning is a fundamental aspect, and one which is often neglected. When working with real data, the process is often very time consuming, but is essential for obtaining good quality results, and from there useful new knowledge. The quality of the results directly depends on the quality of the data, and consequently, on the correct missing data treatment, outlier identification, etc. Data miners should become conscious of the importance of performing very careful and rigorous data cleaning, and allocate sufficient time to this activity accordingly.

Cautions on transformations: Avoidance of unnecessary transformations is recommended, especially if the transformation decreases interpretability (for example $Y = \log$ streamflow, although Y is normal). If transformations are definitely required, some bias may be introduced into the results. Thus, it is convenient to minimise arbitrariness of the transformation as much as possible (in recording *Age*, Adult may be defined from 18 to 65 or from 15 to 70), and this implies that the goals of the analysis must also be taken into account. For arithmetic transformations, imputation of missing data before the transformation is thought to be better. Note that where data are numeric and the scale changes between variables, normalisation may be necessary.

- Select DM method taking into account dataset size. Where datasets are small, choose simpler methods and be mindful of the maximum theoretical certainty that can be obtained.

Parameter selection and model fitting: While parameter-free DM algorithms do exist, most require some *a priori* set up. Parameters for DM algorithms are decided by the same methods as more common models – expert knowledge, guessing, trial and error, automated and manual experimentation. In addition, it is often helpful to learn a little about the role of the parameter within the algorithm, as appropriate values for the problem at hand can often be set or estimated this way. Some experimentation may improve the output model and reporting the process of parameter fitting in detail adds credibility to any modelling project. It is important that parameter values are not chosen based on the final test data. Otherwise optimistic performance estimates will be obtained.

- Use Principal Component Analysis for synthesising an original set of numerical variables into a small number of fictitious variables, conserving as much information as possible from the original dataset.
  - o Multiple correspondence analysis is suitable if the original variables are qualitative.

Cautions: Multivariate techniques reduce the original variables to a set of fictitious ones (or factors). Conceptual interpretation of a factor may not be clear, and if this is the case, there will be implications for interpretability of the final results. Principal Component Analysis is only recommended when all original variables are numerical. For qualitative data, multiple correspondence analyses should be used in its place (Lebart et al., 1984; Dillon and Goldstein, 1984).

- Use feature weighting techniques for calculating the importance of the variables according to the information they provide.
- Use clustering techniques for finding groups of homogeneous objects in a dataset.
  - o Density estimation techniques for finding the combination of density functions associated with homogeneous groups of objects in data.

Cautions: Most clustering methods generate a set of clusters even where no set of distinguishable groups really exists. This is why it is very important to carefully validate the correctness of the discovered clusters. Meaning and usefulness of discovered classes are one validation criteria, although this is largely subjective. A more quantitative approach is to perform multiple runs of the algorithm or different algorithms with slightly different parameters or initial values, which will give a good indication of the stability of the cluster scheme. Some software packages also contain tools to assess structural quality of the classes. As a measure of cluster 'goodness,' the ratio of average distance within the clusters with respect to average distance between clusters may be useful where a numerical distance measure exists, although it is redundant if the same criterion were used to build the clusters themselves, as is the case of using Ward's (1963) method. Cluster validation where no reference partition exists (and in real-world applications none is present, or the clustering would be unnecessary) is an open problem, but stability and meaning of classes should be observed as a minimum treatment (Gibert et al., 2005c).

- Use decision trees or classification rules for a new unclassified object in order to discriminate to which group or class it belongs or better fits.

Classification: When classifying real data, it is often useful to consider accuracy on a class-by-class basis. In this way, the modeller can keep track of where errors are occurring. These errors may be given unequal weighting, if the consequences are not equal. The most common device for this is the confusion matrix. The distribution of input data should also receive consideration, as many classification algorithms tend towards predicting the majority class. An in-depth discussion of this topic can be found in Weiss and Provost (2001). Tree (and other classifier) stability can be assessed as cluster stability (see above).

- Use statistical modelling for finding the combination of variables that better predicts a target variable:
  - o Linear regression if the target is numeric, regressors are independent, and normality and linearity holds.
  - o Non-linear regression or neural nets under the same conditions but when linearity does not hold.
  - o ANOVA for a numerical target and qualitative regressors; independence and normality are required.
  - o ANCOVA for a numeric target and both numeric and qualitative regressors. Normality, independence and linearity required.
  - o Binary logistic regression if the target is binary and regressors numeric.

Cautions: Scalar real-valued performance criteria such as the *determination coefficient* (R^2, also known as efficiency), used together with residual plots (Moore and McCabe, 1993), constitute a very useful tool for validation of the model, far more powerful than numeric indicators by themselves. Outliers, influent values, non-normalities, non-linearities and other anomalies can be investigated in this way. Note however that R^2 can be applied only to real numerical data.

- Use classical time series techniques for predicting a numeric value of a variable taking into account the past temporal values of the same variable.

- Use Case-Based Reasoning or Instance-based Learning for solving a new problem (codified as an object) by means of reusing the most similar problem (object) in the data set.
- Use Rule-induction to obtain, from a set of labelled data, a set of discriminating rules that can be used as a classification rules for new objects.
- Use Association algorithms to induce rules that can extract regular correlation patterns among several variables within the data set.
- Validate the models using proper tools.

  Caution on p values: As the amount of data increases, variance of classical estimators tends to zero, which usually implies that very small sample differences may appear statistically significant. This phenomenon requires serious attention and great care must be exercised in the interpretation of those statistical results. In fact, serious revision of classical statistical inference is necessary to enable suitable use in the context of DM.

  Uncertainty quantification and model validation: As mentioned in the note regarding input data above, proper consideration of uncertainty is essential for meaningful modelling. One must also give thought to how best to quantify and represent the performance of the final model. For some purposes, a single-valued measure such as R^2 may be sufficient provided that the model has been properly validated as unbiased, but for most applications more information is useful. It is seldom possible to represent model performance against all goals of the investigation with one number. It may also have a systematic tendency to slightly overpredict lower values to compensate for missing extreme events. All of this cannot be expressed as a single number, but a comparison of distributions will reveal the necessary information.

  Model validation is as important for automatically extracted models as it is for those constructed with more human interaction, perhaps more so. To this end we recommend the usual best practice procedures such as holding back a portion of the dataset for independent validation (if the size of database allows) and n-fold cross validation.

12.3.1 Integrated approaches

The main goal of many environmental system analyses is to support posterior decision making to improve either management or control of the system. Intelligent Environmental Decision Support Systems (IEDSSs) are among the most promising approaches in this field (see Chapter 8). IEDSS are integrated models that provide domain information by means of analytical decision models, and allow the decision maker access to databases and knowledge bases. They intend to reduce the time in which decisions can be made as well as assist repeatability and the quality of eventual decisions by offering criteria for the evaluation of alternatives or for justifying decisions (Poch et al., 2004). Often, multiple scenarios are modelled and evaluated according to environmental, social and economic criteria.

There are six primary approaches to the problem of building an integrated model (Ekasingh et al., 2005): expert systems, agent-based modelling, system dynamics, Bayesian networks, coupled complex models, and meta-modelling. Of these, the last three are most relevant to the field of DM. Opportunities exist for automation of Bayesian network and meta-model construction and parameterisation, simplification and summarisation of complex submodels, and also interpretation of results. DM techniques are important tools for the knowledge acquisition phase of integrated model building, and because integrated models are very high in complexity, results are often correspondingly difficult to interpret and the decision maker may benefit from a postprocessing DM step. Of course, data mined models may also form part of the integrated model as in Ekasingh et al. (2005).

12.4. Software – Existing and Under Development

In this section, some software tools available to perform Data Mining on real data are referenced. These software tools or packages include some of the Data Mining techniques presented in this chapter. On the one hand, there are many proprietary Data Mining packages that merit mentioning and they are briefly described in the following paragraphs.

SAS's Enterprize miner (see http://www.sas.com/technologies/analytics/datamining/miner/) SAS Enterprise Miner streamlines the entire data mining process from data access to model deployment by supporting all necessary tasks within a single, integrated solution, all while providing the flexibility for efficient workgroup collaborations. It provides tools for graphical programming, avoiding manual coding, which makes for easy to develop complex data mining processes.

It was designed for business users, and provides several tools to help with preprocessing data (descriptives, advanced statistical graphics) together with advanced predictive modelling tools and algorithms, including decision trees, neural nets, autoneural nets, memory-based reasoning, linear and logistic regression, clustering, association rules, time series. It provides a facility for direct connection with data warehouses. It also offers tools for comparing the results of different modelling techniques.

It is integrated with other tools from the wider SAS statistical framework, which at present is one of the most powerful statistical packages commercially available. SAS and Enterprise Miner are delivered as a distributed client-server system. Both are especially well suited for large organisations.

IBM has released *Intelligent Mine* (http://www-306.ibm.com/software/data/iminer/). IBM's in-database mining capabilities integrate with existing systems to provide scalable, high performing predictive analysis without moving data into proprietary data mining platforms. SQL, Web Services, or Java can be used to access DB2's data mining capabilities directly from the own user's applications or business intelligence tools from IBM's business partners. It provides a set of products for data warehouse editing, modelling, scoring or visualisation, including: market basket analysis, clustering or categorisation, or summarisation. It is available in a number or different languages including English, Danish, Spanish and Arabic. It also provides a graphical programming interface.

Clementine (http://vvv.spss.com/Clementine) was one of the first commercial tools oriented to Data Mining. Later absorbed by the firm of SPSS, which also commercialises a very popular and widely used statistical package. Clementine is designed to support the CRISP-DM, the de facto standard data mining methodology. It provides a visual interactive workflow interface supporting the data mining process and has an open architecture for integration with other systems and with all SPSS predictive analytics. It includes facilities for database access, text, survey and web data preparation, model management, automatic version control, user authentication, etc. From the point of view of data mining techniques, it provides neural networks, decision trees, rule induction, association rules, classification, data visualisation and the statistical functionalities of SPSS.

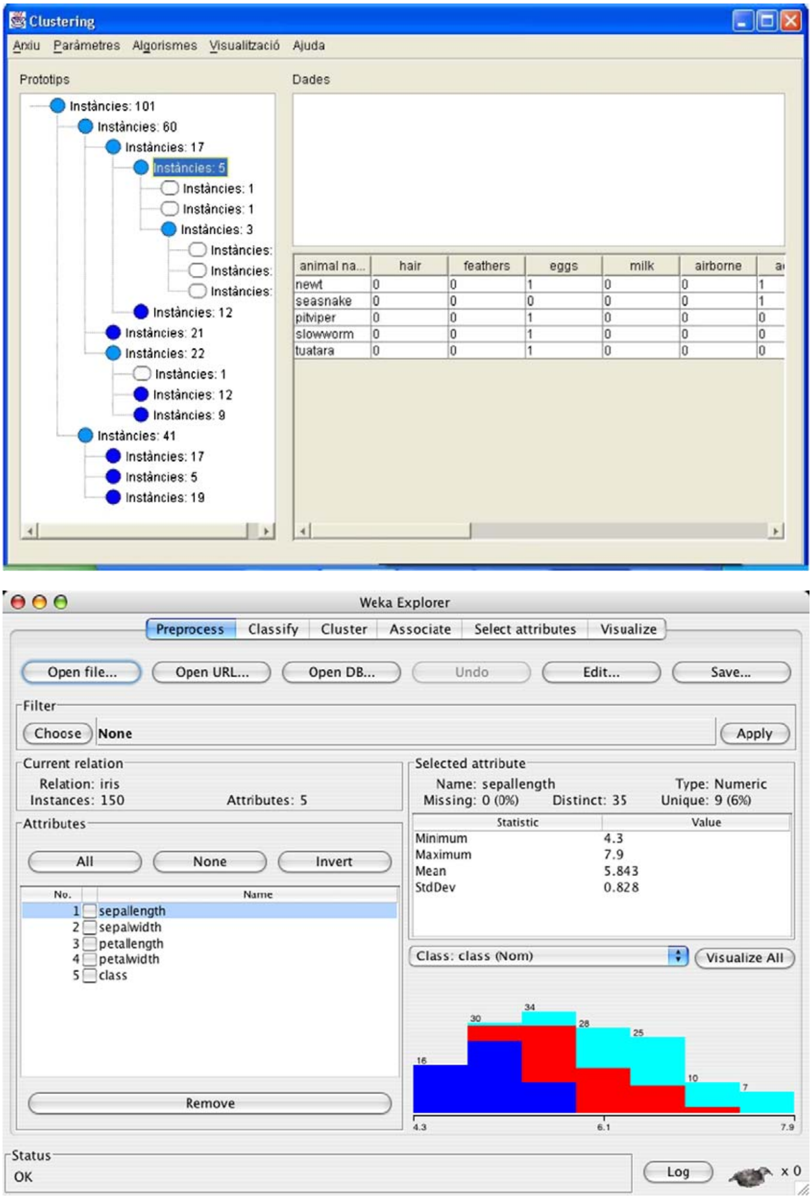
Salford System's CART (http://www.salford-systems.com/) is a decision tree tool that automatically sifts large, complex databases, searching for and isolating significant patterns and relationships. This discovered knowledge is then used to generate reliable, easy-to-grasp predictive models for applications such as profiling customers, targeting direct mailings, detecting telecommunications and credit card fraud, and managing credit risk. In addition, CART is an excellent pre-processing complement to other data analysis techniques.

The *WEKA* workbench (Witten and Frank, **?**) contains a collection of visualisation tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. Weka supports several standard DM tasks: data preprocessing, clustering, classification, regression, visualisation, and feature selection. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

Weka's main user interface is the Explorer, shown in Figure 12.2, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets rather than a single one. Weka is a general purpose package, freely available on the Internet and is well utilised in the Artificial Intelligence Community.

GESCONDA (Gibert et al., 2004; Sànchez-Marrè et al., 2004) is the name given to an Intelligent Data Analysis System developed, with partial financing of project TIN2004-01368, with the aim of facilitating KD and especially oriented to environmental databases. On the basis of previous experiences, it was designed as a four level architecture connecting the user with the environmental system or process (Figure 12.2 shows a screen capture from the clustering GUI): Data Filtering, Recommendation and Meta-Knowledge Management, Data Mining techniques for KDD, Knowledge Management.

Central characteristics of GESCONDA are the integration of statistical and AI methods into a single tool, together with mixed techniques for extracting knowledge contained in data, as well as tools for qualitative analysis of complex relationships along the time axis (Sànchez-Marrè et al., 2004). All techniques implemented in GESCONDA can share information among themselves to best co-operate for extracting knowledge. It also includes a capability for explicit management of results produced by the different methods. Portability of the software between platforms



animal na...	hair	feathers	eggs	milk	airborne	a
newt	0	0	1	0	0	1
seasnake	0	0	0	0	0	1
pitviper	0	0	1	0	0	0
slowworm	0	0	1	0	0	0
tuatara	0	0	1	0	0	0

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Undo Edit... Save...
Filter
Choose
Apply
Attributes
All None Invert
Remove
Visualize All
Log
x 0

Filter: None

Current relation
Relation: iris
Instances: 150 Attributes: 5

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute
Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

Status
OK

Figure 12.2 (Top) GESCONDA clustering interface; (bottom) The Weka Explorer user interface.

is provided by a common Java platform. The GESCONDA design document is in http://www.eu-lat.org/eenviron/Marre.pdf.

Finally, other DM libraries for general computation as well as statistical environments like Matlab and the software R have been built, providing a wide range of techniques that can be useful in complex Data Mining processes.

12.5. Conclusions and Challenges for Data Mining of Environmental Systems

In this chapter a general introduction is provided to Data Mining techniques relevant to their use for Environmental Systems. Special focus is on the contributions of DM techniques to environmental applications as well as on general guidelines for good practice in real world domains. Technical details on the specific DM techniques are not a focus of this paper. Rather the aim is to provide general guidance to a non-expert user so that they can decide which technique is appropriate for solving their problem and appreciate the cautions required to avoid erroneous analyses. For in-depth details on specific techniques, references are provided. The overview presented does not attempt to be exhaustive. Those DM techniques deemed more suitable for environmental applications are included, but there are some others that may be useful in particular applications, and several references are provided in the paper on more specific techniques.

Finally, in this section, we shall comment on the hot issues and challenging aspects in the interdisciplinary field of environmental DM sciences. Achievement of the following aims would increase utility and applicability of DM methods:

- improvement of automated preprocessing techniques,
- elaboration of protocols to facilitate sharing and reuse of data,
- development of standard procedures (benchmarks) for experimental testing and validation of DM tools,
- involvement of end-user (domain expert) criteria in algorithm design and result interpretation,
- development and implementation of mixed DM methods, combining different techniques for better KDD,
- formulation of tools for explicit representation and handling of discovered knowledge for greater understandability,
- improvement of DM techniques for on-line and heterogeneous databases,
- design and use of temporal/spatial data mining techniques,
- research towards integrated modelling that combines different partial models from different natures (statistical, qualitative, graphical) in a single global one,
- further development of guidelines and recommendations to assist with method and algorithm selection.

Another factor that is often of great importance is (conceptual) interpretability of output models. Indeed, some of the results of the Data Mining step appear under some formalisms not intuitive enough to be easily understandable by an environmental scientist (think, for example, about the results of a logistic regression, or a random forest's results). Tools that clearly and usefully summarise extracted knowledge are of great value to environmental scientists, as are those that assist in the quantification of uncertainties.

ACKNOWLEDGEMENT

Authors want to thank Pr. Eibe Frank, from the University of Waikato, New Zealand, for his valuable input regarding the description of the Weka software.

Uncited references

(Gibert et al., 2006) (Kaburlasos et al., 2007) (Spate and Jakeman, in press)

REFERENCES

Aha, D., 1998. Feature weighting for lazy learning algorithms. In: Liu, H., Motoda, H. (Eds.), Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer.

Allison, P., 2002. Missing Data. Sage, Thousand Oaks, CA, USA.

Almasri, M., Kaluarachchi, J., 2005. Modular neural networks to predict the nitrate distribution in ground water using the on-ground nitrogen loading and recharge data. Environmental Modelling and Software 20 (7), 851–871.

Andrienko, G., Andrienko, A., 2004. Research on visual analysis of spatio-temporal data at Fraunhofer AIS: An overview of history and functionality of CommonGIS. In: Proceedings of the Knowledge-Based Services for the Public Services Symposium, Workshop III: Knowledge Discovery for Environmental Management. KDnet, pp. 26–31.

Athanasiadis, I., Karatzas, K., Mitkas, P., 2005. Contemporary air quality forecasting methods: A comparative analysis between statistical methods and classification algorithms. In: Proceedings of the 5th International Conference on Urban Air Quality, Valencia, Spain.

Athanasiadis, I.N., Mitkas, P.A., 2007. Knowledge discovery for operational decision support in air quality management. Journal of Environmental Informatics 9, 100–107.

Babovic, V., 2005. Data mining in hydrology. Hydrological Processes 19, 1511–1515.

Barnett, V., Lewis, T., 1978. Outliers in Statistical Data. Wiley.

Belanche, L., Valdés, J., Comas, J., Rodríguez-Roda, I., Poch, M., 2001. Towards a model of input-output behaviour of wastewater treatment plants using soft computing techniques. Environmental Modelling and Software 5 (14), 409–419.

Cadez, I., Smyth, P., 1999. Modelling of inhomogeneous Markov random fields with applications to cloud screening. Tech. Rep. UCI-ICS 98-21. University California Irvine, USA.

Camargo, S., Robertson, A., Gaffney, S., Smyth, P., 2004. Cluster analysis of Western North Pacific tropical cyclone tracks. In: Proceedings of the 26th Conference on Hurricanes and Tropical Meteorology, Miami, pp. 250–251.

Cendrowska, J., 1998. Prism: An algorithm for inducing modular rules. International Journal of Man-Machine Studies 27 (4), 349–370.

Cohen, W., 1995. Fast effective rule induction. In: Prieditis, A., Russell, S. (Eds.), Proceedings of the Twelfth International Conference on Machine Learning. Morgan Kaufmann, pp. 115–123.

Comas, J., Dzeroski, S., Gibert, K., Rodríguez-Roda, I., Sànchez-Marrè, M., 2001. Knowledge discovery by means of inductive methods in wastewater treatment plant data. AI Communications 14 (1), 45–62.

Comas, J., Llorens, E., Martí, E., Puig, M.A., Riera, J.L., Sabater, F., Poch, M., 2003. Knowledge acquisition in the STREAMES project: The key process in the environmental decision support system development. AI Communications 16 (4), 253–265.

Cortés, U., Rodríguez-Roda, I., Sànchez-Marrè, M., Comas, J., Cortés, C., Poch M., 2002. DAI-DEPUR: An environmental decision support system for supervision of municipal waste water treatment plants. In: Proceedings of the 15th European Conference on Artificial Intelligence, ECAI 2002, Lyon, France, pp. 603–607.

Cunningham, S.J., Holmes, G., 1999. Developing innovative applications in agriculture using data mining. In: Proceedings of the Southeast Asia Regional Computer Confederation Conference.

Dillon, W., Goldstein, M., 1984. Multivariate Analysis. Wiley, New York, NY, USA.

Dixon, M., Gallop, J.R., Lambert, S.C., Healy, J.V., 2007. Experience with data mining for the anaerobic wastewater treatment process. Environmental Modelling and Software 22, 315–322.

Domingos, P., 1996. Unifying instance-based and rule-based induction. Machine Learning 24, 141–168.

Draper, N., Smith, H., 1998. Applied Regression Analysis. Wiley.

Dubes, R., Jain, A., 1988. Algorithms for Clustering Data. Prentice Hall.

Dzeroski, S., Drumm, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucospilota) on Rarotonga, Cook Islands. Ecological Modelling 170 (2–3), 219–226.

Dzeroski, S., Grbovic, J., Walley, W., Kompare, B., 1997. Using machine learning techniques in the construction of models. II. Data analysis with rule induction. Ecological Modelling 95 (1), 95–111.

Ekasingh, B., Ngamsomsuke, K., Letcher, R., Spate, J., 2003. A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. In: Singh, V., Yadava, R. (Eds.), Advances in Hydrology: Proceedings of the International Conference on Water and Environment, pp. 175–188.

Ekasingh, B., Ngamsomsuke, K., Letcher, R., Spate, J., 2005. A data mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management. Journal of Environmental Management 77 (4), 315–325.

Ellis, F., 1996. The application of machine learning techniques to erosion modelling. In: Proceedings of the Third International Conference on Integrating GIS and Environmental modelling. National Center for Geographic Information and Analysis, Santa Fe, USA.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996a. Advances in knowledge discovery and data mining. In: Data Mining to Knowledge Discovery: An Overview. American Association for Artificial Intelligence, pp. 1–34.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996b. From data mining to knowledge discovery in databases (a survey). AI Magazine 3 (17), 37–54.

Gatts, C., Ovalle, A., Silva, C., 2005. Neural pattern recognition and multivariate data: Water typology of the Paraiba do Sul River, Brazil. Environmental Modelling and Software 20 (7), 883–889.

Gibbs, M., Morgan, N., Maier, H., Dandy, G.C.H.M., Nixon, J., 2003. Use of artificial neural networks for modelling chlorine residuals in water distribution systems. In: MODSIM 2003: Proceedings of the 2003 International Congress on Modelling and Simulation, Townsville, Australia, pp. 789–794.

Gibert, K., Sonicki, Z., 1999. Clustering based on rules and medical research. Journal on Applied Stochastic Models in Business and Industry, formerly JASMDA 15 (4), 319–324.

Gibert, K., Flores, X., Rodríguez-Roda, I., Sànchez-Marrè, M., 2004. Knowledge discovery in environmental data bases using GESCONDA. In: Pahl-Wostl, C., Schmidt, S., Rizzoli, A.E., Jakeman, A.J. (Eds.), Proceedings of IEMSS 2004: International Environmental Modelling and Software Society Conference Osnabruck, Germany, pp. 51–56.

Gibert, K., Annicchiarico, R., Cortés, U., Caltagirone, C., 2005a. Knowledge Discovery on Functional Disabilities: Clustering Based on Rules Versus Other Approaches. IOS Press.

Gibert, K., Nonell, R., Velarde, J.M., Colillas, M.M., 2005b. Knowledge discovery with clustering: Impact of metrics and reporting phase by using KLASS. Neural Network World 4, 319–326.

Gibert, K., Sànchez-Marrè, M., Flores, X., 2005c. Cluster discovery in environmental databases using GESCONDA: The added value of comparisons. AI Communications 4 (18), 319–331.

Gibert, K., Sànchez-Marrè, M., Rodriguez-Roda, I., 2006. GESCONDA: An intelligent data analysis system for knowledge discovery and management in environmental databases. Environmental Modelling and Software 21, 115–120.

Guo, Q., Kelly, M., Graham, C., 2005. Support vector machines for predicting distribution of sudden oak death in California. Ecological Modelling 182 (1), 75–90.

Hall, M., 1999. Feature selection for discrete and numeric class machine learning. Tech. Rep., Department of Computer Science, University of Waikato, Working Paper 99/4. URL: http://www.cs.waikato.ac.nz/~ml/publications1999.html.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

Holmes, G., Cunningham, S., Dela Rue, B., Bollen, A., 1998. Predicting apple bruising using machine learning. In: Proceedings of the Model-IT Conference. Acta Horticulturae 476, 289–296.

Kaburlasos, V.G., Athanasiadis, I.N., Mitkas, P.A., 2007. Fuzzy Lattice Reasoning (FLR) classifier and its application for ambient ozone estimation. International Journal of Approximate Reasoning 45, 152–188.

Kolodner, J., 1993. Case-Based Reasoning. Morgan Kaufmann.

Kralisch, S., Fink, M., Flügel, W.-A., Beckstein, C., 2001. Using neural network techniques to optimize agricultural land management for minimisation of nitrogen loading. In: MODSIM 2001: Proceedings of the 2001 International Congress on Modelling and Simulation, Canberra, Australia, pp. 203–208.

Larose, D., 2004. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley.

Lebart, L., Morineau, A., Warwick, K., 1984. Multivariate Descriptive Statistical Analysis. Wiley, New York.

Little, R., Rubin, D., 1987. Statistical Analysis with Missing Data. Wiley.

Martín, F.J., Plaza, E., 2004. Ceaseless case-based reasoning. In: Proc. of 7th European Conference on Case-Based Reasoning. ECCBR, 2004. In: Lecture Notes in Artificial Intelligence, vol. 3155. Springer-Verlag, Heidelberg, Germany, pp. 287–301.

Mas, J., Puig, H., Palacio, J., Sosa-Lopez, A., 2004. Modelling deforestation using GIS and artificial neural networks. Environmental Modelling and Software 19 (5), 461–471.

McKenzie, N., Ryan, P., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma (89), 67–94.

Michalski, R., Chilausky, R., 1980. Learning by being told and learning by examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. International Journal of Policy Analysis and Information Systems 4 (2), 125–161.

Miller, H.J., 2007. Geographic data mining and knowledge discovery. In: Wilson, J.P., Fotheringham, A.S. (Eds.), Handbook of Geographic Information Science. Blackwell Publishing.

Molina, L., Belanche, L., Nebot, A., 2002. Feature selection algorithms: A survey and experimental evaluation. In: ICDM 2002: Proceedings of the IEEE International Conference on Data Mining, pp. 306–313.

Moore, D., McCabe, G., 1993. Introduction to the Practice of Statistics, 2nd ed. WH Freeman, New York.

Mora-López, L., Conejo, R., 1998. Qualitative reasoning model for the prediction of climatic data. In: ECAI 1998: Proceedings of the 13th European Conference on Artificial Intelligence, Brighton, UK, pp. 61–75.

Núñez, H., Sànchez-Marrè, M., Cortés, U., 2003. Improving similarity assessment with entropy-based local weighting. In: Proceedings of the 5th International Conference on Case-Based Reasoning. ICCBR, 2003. In: Lecture Notes in Artificial Intelligence, vol. 2689. Springer-Verlag, Heidelberg, Germany, pp. 377–391.

Núñez, H., Sànchez-Marrè, M., Cortés, U., Comas, J., Martinez, M., Rodríguez-Roda, I., Poch, M., 2004. A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. Environmental Modelling and Software 19 (9), 809–819.

Parr Rud, O., 2001. Data Mining Cookbook—Modelling Data for Marketing, Risk, and CRM. Wiley.

Poch, M., Comas, J., Rodríguez-Roda, I., Sànchez-Marrè, M., Cortés, U., 2004. Designing and building real environmental decision support systems. Environmental Modelling and Software 19, 857–873.

Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecological Modelling 146 (1–3), 303–310.

Riaño, D., 1998. Learning rules within the framework of environmental sciences. In: ECAI 1998: Proceedings of the 13th European Conference on Artificial Intelligence, Brighton, UK, pp. 151–165.

Robertson, A., Kirshner, S., Smyth, P., 2003. Hidden Markov models for modelling daily rainfall occurrence over Brazil. Tech. Rep. UCI-ICS 03-27. URL: http://www.datalab.uci.edu/papers-by-date.html.

Rodríguez-Roda, I., Poch, M., Sànchez-Marrè, M., Cortés, U., Lafuente, J., 1999. Consider a case-based system for control of complex processes. Chemical Engineering Progress 6 (95), 39–48.

Rodríguez-Roda, I., Comas, J., Poch, M., Sànchez-Marrè, M., Cortés, U., 2001. Automatic knowledge acquisition from complex processes for the development of knowledge based systems. Industrial and Engineering Chemistry Research 15 (40), 3353–3360.

Rodríguez-Roda, I., Comas, J., Colprim, J., Poch, M., Sànchez-Marrè, M., Cortés, U., Baeza, J., Lafuente, J., 2002. A hybrid supervisory system to support wastewater treatment plant operation: Implementation and validation. Water Science and Technology 45 (4–5), 289–297.

Rubin, D., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley.

Sanborn, S., Bledsoe, B., 2005. Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon. Journal of Hydrology 325 (1–4), 241–261.

Siebes, A., 1996. Data mining: What it is and how it is done. In: Proceedings of 15th Conferenza in Sistemi Evolutti per Basi di Datti, Pisa, Italy, pp. 329–344.

Sànchez-Marrè, M., Cortés, U., Béjar, J., de Gracia, J., Lafuente, J., Poch, M., 1997. Concept formation in wastewater treatment plants by means of classification techniques: A compared study. Applied Intelligence 7 (2), 147–165.

Sànchez-Marrè, M., Gibert, K., Rodríguez-Roda, I., 2004. GESCONDA: A tool for knowledge discovery and data mining in environmental databases. In: Research on Computing Science, vol. 11. Centro de Investigación en Computación, Instituto Politécnico Nacional, México, DF, México, pp. 348–364.

Sànchez-Marrè, M., Cortés, U., Martínez, M., Comas, J., Rodríguez-Roda, I., 2005. An approach for temporal case-based reasoning: Episode-based reasoning. In: Proc. of 6th International Conference on Case-Based Reasoning. ICCBR, 2005. In: Lecture Notes in Artificial Intelligence, vol. 3620. Springer-Verlag, pp. 465–476.

Stadler, M., Ahlers, D., Bekker, R.M., Finke, J., Kunzmann, D., Sonnenschein, M., 2006. Web-based tools for data analysis and quality assurance on a life-history trait database of plants of Northwest Europe. Environmental Modelling and Software 21, 1536–1543.

Spate, J., 2002. Data in hydrology: Existing uses and new approaches. Australian National University, Honours Thesis, Mathematics Department.

Spate, J., 2005. Modelling the relationship between streamflow and electrical conductivity in Hollin Creek, southeastern Australia. In: Fazel Famili, A., Kok, J., Peña, J. (Eds.), Proceedings of the 6th International Symposium on Intelligent Data Analysis, pp. 419–440.

Spate, J., 2006. Machine learning as a tool for investigating environmental systems. PhD thesis, Australian National University.

Spate, J., Jakeman, A., in press. Review of data mining techniques and their application to environmental problems. Environmental Modelling and Software.

Spate, J., Croke, B., Jakeman, A., 2003. Data mining in hydrology. In: MODSIM 2003: Proceedings of the 2003 International Congress on Modelling and Simulation, Townsville, Australia, pp. 422–427.

Su, F., Zhou, C., Lyne, V., Du, Y., Shi, W., 2004. A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution. Ecological Modelling 174 (4), 421–431.

Swayne, D., Cook, D., Buja, A., 1998. XGobi: Interactive dynamic data visualization in the x window system. Journal of Computational and Graphical Statistics 7 (1), 113–130.

Sweeney, A., Beebe, N., Cooper, R., 2007. Analysis of environmental factors influencing the range of anopheline mosquitoes in northern Australia using a genetic algorithm and data mining methods. Ecological Modelling 203 (3–4), 375–386.

Ter Braak, C., Hoijtink, H., Akkermans, W., Verdonschot, P., 2003. Bayesian model-based cluster analysis of predicting macrofaunal communities. Ecological Modelling 160 (3), 235–248.

Vellido, A., Martí, J., Comas, I., Rodríguez-Roda, I., Sabater, F., 2007. Exploring the ecological status of human altered streams through generative topographic mapping. Environmental Modelling and Software 22 (7), 1053–1065.

Voss, H., Wachowicz, M., Dzeroski, S., Lanza, A. (Eds.) 2004. Knowledge Discovery for Environmental Management. Knowledge-Based Services for the Public Sector Conference. Notes on the KDnet Workshop. Bonn, Germany.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. Journal of American Statistical Association 58, 236–244.

Weiss, G., Provost, F., 2001. The effect of class distribution on classier learning: An empirical study. Tech. Rep., Department of Computer Science, Rutgers University, Technical Report ML-TR-44. URL: http://www.research.rutgers.edu/~gweiss/papers/ml-tr-44.pdf.

Whitten, I., Frank, E., 1991. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

Wong, I.W., Bloom, R., McNicol, D.K., Fong, P., Russell, R., Chen, X., 2007. Species at risk: Data and knowledge management within the WILDSPACETM decision support system. Environmental Modelling and Software 22, 423–430.

Wnek, J., Michalski, R., 1991. Hypothesis-driven constructive induction in aQ17: A method and experiments. In: Proceedings of the IJCAI-91 Workshop on Evaluating and Changing Representation in Machine Learning, pp. 13–22.

Yeates, S., Thomson, K., 1996. Applications of machine learning on two agricultural datasets. In: Proceedings of the New Zealand Conference of Postgraduate Students in Engineering and Technology. Christchurch, New Zealand, pp. 495–496.

Zhu, X., Simpson, A., 1996. Expert system for water treatment plant operation. Journal of Environmental Engineering 122, 822–829.

Zoppou, C., Neilsen, O., Zhang, L., 2002. Regionalization of daily stream flow in Australia using wavelets and *k*-means. Tech. Rep., Australian National University. URL: http://wwwmaths.anu.edu.au/research.reports/mrr/mrr02.003/abs.html. accessed 15/10/2002.

Book:	DIEA2
Chapter:	12

Author Query Form

Dear Author,

During the preparation of your manuscript for typesetting, some questions have arisen. These are listed below. Please check your typeset proof carefully and mark any corrections in the margin of the proof or compile them as a separate list*. This form should then be returned with your marked proof/list of corrections to VTEX.

Disk use

In some instances we may be unable to process the electronic file of your article and/or artwork. In that case we have, for efficiency reasons, proceeded by using the hard copy of your manuscript. If this is the case the reasons are indicated below:

□ Disk damaged □ Incompatible file format □ Non-LaTeX file
□ Virus infected □ Discrepancies between electronic file and (peer-reviewed, therefore definitive) hard copy
□ Other:

We have proceeded as follows:

□ Manuscript scanned □ Manuscript keyed in □ Artwork scanned
□ Files only partly used (parts processed differently:)

Bibliography

If discrepancies were noted between the literature list and the text references, the following may apply:

☒ The references listed below were noted in the text but appear to be missing from your literature list. Please complete the list or remove the references from the text.

☒ *Uncited references*: This section comprises references that occur in the reference list but not in the body of the text. Please position each reference in the text or delete it. Any reference not dealt with will be retained in this section.

Queries and/or remarks

Proof page/line	Query / remark	Author's response
206/23,29;213/3;214/14;215/43	There are references missing in the literature.	
216/18	There are references missing in the literature.	
214/3;221/19	Cunningham et al. (1999) is changed to Cunningham and Holmes (1999). Please check.	
220/20	(http:// is changed to (see http://. Please check	
224/10	Uncited references. Please check Fayyad et al. (1996a): P, S. is changed to Smyth, P. Please update if possible reference Spate and Jakeman (in press).	

Many thanks for your assistance

*In case artwork needs revision, guidance can be found at http://authors.elsevier.com/artwork.