© <2018>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/

Which method to Use? An assessment of Data Mining Methods in Environmental Data Science

Karina Gibert[a,b]**,** Joaquín Izquierdo[d], Miquel Sànchez-Marrè[a,c], Serena H. Hamilton[e,f], Ignasi Rodríguez-Roda[g,h], Geoff Holmes[i]

*[a]Knowledge Engineering and Machine Learning Group, Universitat Politècnica de Catalunya-Barcelona**Tech**. Barcelona, Catalonia; [b]Department of Statistics and Operation Research, Universitat Politècnica de Catalunya-Barcelona**Tech,** Barcelona, Catalonia; [c]Software Department, Technical University of Catalonia. Barcelona, Catalonia; [d]Fluing-IMM Universitat Politècnica de València, Valencia, Spain; [e]School of Sciences, Edith Cowan University, Joondalup, Australia; [f]Fenner School of Environment and Society, Australian National University, Canberra, Australia; [g]Laboratory of Chemical and Environmental Engineering (LEQUIA), University of Girona, Catalonia; [h]Catalan Institute for Water Research (ICRA), Girona, Catalonia; [i]Department of Computer Science, University of Waikato, Waikato, New Zealand*

Summary

Data Mining (DM) is a fundamental component of the Data Science process. Over recent years a huge library of DM algorithms has been developed to tackle a variety of problems in fields such as medical imaging and traffic analysis. Many DM techniques are far more flexible than more classical numerial simulation or statistical modelling approaches. These could be usefully applied to data-rich environmental problems. Certain techniques such as artificial neural networks, clustering, case-based reasoning or Bayesian networks have been applied in environmental modelling, while other methods, like support vector machines among others, have yet to be taken up on a wide scale. There is greater scope for many lesser known techniques to be applied in environmental research, with the potential to contribute to addressing some of the current open environmental challenges. However, selecting the best DM technique for a given environmental problem is not a simple decision, and there is a lack of guidelines and criteria that helps the data scientist and environmental scientists to ensure effective knowledge extraction from data. This paper provides a broad introduction to the use of DM in Data Science processes for environmental researchers. Data Science contains three main steps (pre-processing, data mining and post-processing). This paper provides a conceptualization of Environmental Systems and a conceptualization of DM methods, which are in the core step of the Data Science process. These two elements define a conceptual framework that is on the basis of a new methodology proposed for relating the characteristics of a given environmental problem with a family of Data Mining methods. The paper provides a general overview and guidelines of DM techniques to a non-expert user, who can decide with this support which is the more suitable technique to solve their problem at hand. The decision is related to the bidimensional relationship between the type of environmental system and the type of DM method. An illustrative two way table containing references for each pair Environmental System-Data Mining method is presented and discussed. Some examples of how the proposed methodology is used to support DM method selection are also presented, and challenges and future trends are identified.

Keywords: Data Mining, Data Science, Method Selection, Multidisciplinarity, Environmental Systems.

1 Introduction

Environmental problems, including the degradation and depletion of natural resources, biodiversity loss and climate change, among others, represent some of the most critical challenges of our world today. To effectively address environmental problems, understanding how these system components affect one another is needed. Data Science (DS) is an emergent research field that helps better understand the complex mechanisms behind environmental phenomena. In Gibert *et al.* (2018) an overview of what can be called the field of Environmental Data Science is provided. The paper describes the DS process, and the role of Data Mining (DM) methods, which is identified as one of the most critical to transform data into added value and new knowledge for ESs. DM describes the search for hidden patterns or associations in data to aid understanding of systems and/or their processes. These patterns may help, for example, determine the strength of relationship between variables, or predict future outcomes. DM methods include cluster analysis, factorial methods, decision trees, statistical modelling, time series forecasting, Bayesian networks, among others. There has been increasing interest in applying DM to environmental problems in recent years. However, its full potential in the environmental sciences has yet to be realized.

In Gibert *et al.* (2018), the main challenges of Environmental Data Science are identified and discussed to promote research in the area. One of these main challenges is the lack of guidance in choosing the right analytics method for a given problem. In fact, selection of the proper methods for effective process is difficult, and not much work has been done to establish consensus about which analytics methods are effective and appropriate for specific applications (Gibert *et al.*, 2010). This paper tries to move one step forward in this direction, providing an innovative methodology to help non-expert data scientists identify DM methods suitable to properly analyze a certain kind of data when addressing a specific type of environmental question. It has been observed how several analyses of the same dataset can provide contradictory conclusions when analyzed by two independent data scientists without a common set of guidelines for conducting the analysis in the proper way (Baeza-Yates, 2017; Silberzahn *et al.*, 2015).

This papercto contributes to generate greater awareness of the capabilities of DM in DS processes for a better understanding of ESs, and to extract valuable information from data to support decision-making. Information on the good practices required to ensure DM is correctly used in Environmental Data Science is also provided. Additionally, the paper aims to make DS and DM more accessible to a wider audience, in particular researchers and practitioners in the environmental sciences, and foster discussion of the ways in which DS could be used and encouraged in these science fields.

The paper will briefly introduce the main concepts of DS and the role of DM in the whole process, and illustrate how many DM methods can be valuable tools in the environmental and natural resource science fields. A major conceptualization effort was carried out to organize both environmental systems and DM methods, and to analyze the framework of the application and applicability of DM methods to solve environmental problems. A major contribution of the paper is the proposal of a new methodological framework to guide data scientists or practitioners in identifying the most suitable DM method for a given problem. The proposed methodology is based on two steps each one using a tool presented in the paper. The first tool is the DM methods conceptual map (DMMCM), which organizes main families of DM methods according the kind of to questions they can address. A second tool is the DM methods templates (DMMTs), which provides detailed information on the specific methods from a given branch of the DMMCM, and guide the selection of the most appropriate technique for the particular case in hand. In the first step of the proposed methodology, target environmental questions lead the decision. In the second step, specific implantation of results in the specific environmental situation adressed comes into consideration, as seen in Section 6. Also, an additional effort has been done to identify which kind of questions arises in each type of ES, and how the DMMCM and the DMMTs can be used to choose the DM technique that best fits a given real-world environmental application. Specific examples from the literature are also provided to illustrate the use of various DM methods for a variety of environmental problems, and a discussion regarding how these two elements interact (DM methods and target ESs) is raised. Finally, a reduced set of illustrative case studies show how the DMMCM and DMMTs can be used select a DM method for a specific problem. The paper ends with some conclusions, future work, and open challenges for better DM method selection.

2 Types of environmental systems

ESs encompass the complex interaction of natural units (water, vegetation, animals, athmosphere, etc.), human activities (agriculture, fishing, water treatment, etc.) and natural phenomena that occur within their boundaries. Most of our activities are directly or indirectly related to natural resources (both biotic like forest, animals and fossil fuels; and abiotic like water, air, atmosphere and land), of which their quality and availability are severely affected by climate (including natural phenomena), and human activities (Figure 1). Human activities can also influence the climate. Some of the largest critical problems now affecting the world are related to air pollution (global warming, ozone depletion, acid rain, smog), water pollution (from both point and distributed sources), hazardous waste, and rain forest destruction. There is no single way to face these environmental problems, which require multidisciplinary approaches and strategies by governments, industry and citizens across the globe.

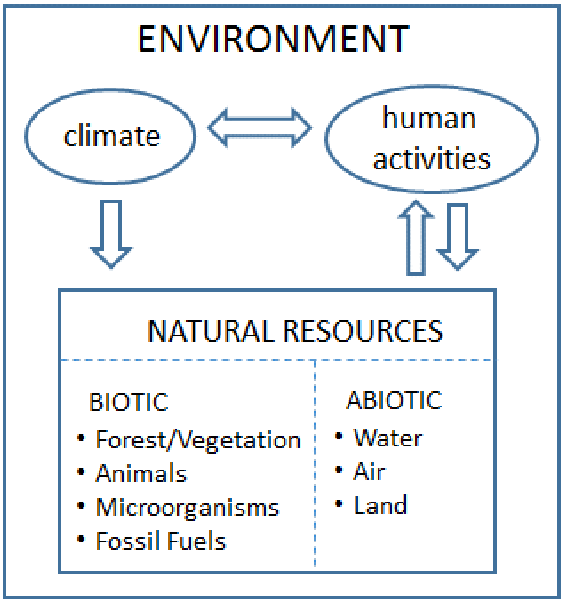

Figure 1: Environmental Systems

ESs and their interdisciplinary processes are highly intricate, with confounding complexities stemming from various sources. ESs are often characterized as ill-structured, non-linear, dynamic, and uncertain, with multiple drivers and system feedbacks (Table 1). These features make the analyses required for effective management natural resources especially difficult.

These characteristics affect all levels of decision, from planning environmental policies, to agricultural-related applications, water and wastewater treatment, storm management, landscape analysis, cloud screening, etc. Large amounts of data, with a high degree of uncertainty, are generated from many sources. These data are not enough to fully describe the ES, but will at least cover partially the domain. Gathering data is often the most difficult, time-consuming and resource-intensive step in conducting measurement activities.

Therefore, selecting the data collection tool or method that will provide the best information is critical (EPA 305-R-07-001). Recently there have been major advances in environmental monitoring technologies – i.e., automated sensors and remote sensing data (e.g. Elarab et al. 2015) – that have greatly reduced sampling costs for some data (e.g. hydrological data), and in the provision of access to data (e.g. web-based data repositories).

Table 1. Some key characteristics of ESs

Characteristics	**Description**
Interdisciplinarity	A variety of biophysical, economic and social factors are at play. This requires data, knowledge and analytical techniques from multiple disciplines to be integrated.
Heterogeneity of data source	ESs are characterized by a high level of heterogeneity in data, since it comes from numerous sources, with different formats, resolutions and qualities. Qualitative and subjective information is often very relevant.
Multiple drivers	ESs are affected by a web of multiple, and often diverse, drivers with both direct and indirect pathways of effect. It is therefore very difficult to control these systems. Also, these drivers can interact leading to cumulative or synergistic effects.
Ill-structured	ESs are poor or ill structured. High order interactions between animal, vegetal, human and climatic system components coexist and often involve processes which are not well known yet, which adds complexity to the system, and makes it difficult to be clearly formulated with a mathematical theory or deterministic model. This implies that solutions might be neither unique nor permanent, nor transferable to other places.
Nonlinearity	Environmental processes are often highly nonlinear, and can exhibit stochastic, dynamic, cyclic or abrupt behaviour.
System feedbacks	System components can interact reciprocally, forming feedback loops. Positive feedbacks can accelerate/amplify effects of stressors, whereas negative feedbacks diminish effects.
Multi.objectives	Typically managing ESs involves multiple and often conflicting objectives to be simultaneously considered, and additional constraints that can disprove model solutions.
Spatio-temporal dynamics	ESs are, in general, non-static, i.e. evolve both over time and space. The assumption of stationarity cannot be justified since interactions among the various factors involved in ESs change over space and time (Guariso and Werthner 1989), and also involve complex seasonal auto-correlated behaviours.
Uncertainty	A high number of causes of uncertainty play together in ES, producing incomplete, imprecise and uncertain data (measurement errors, lack of precision of instruments used, defective calibration of instruments, how the data are read, the frequency of measured data, and the transmission and storage of data, among others (Alferes et al, 2016)). Data generally does not fully capture the system behaviour across space and time, due to the typically high cost of sampling and/or limited understanding of the spatial structure of variables. Each sensor or sampling point is affected by numerous different factors, many of which are unmeasured. Managing uncertainty in the proper way is critical in ESs (Brugnach et al., 2008).
Multiple spatio-temporal scales	As dynamic systems under multiple stressors, environmental resources or assets are affected by processes taking place at multiple spatial and temporal scales. Cause and effect are not always related in time and space (e.g. time lags, spatially disjoint processes) and harmonization of these two dimensions is delicate and critical,

Even in cases where data describing ESs has become largely available, datasets are often not examined in depth and much of their information content still remains unexplored (Hammond, 2007; Gibert, 2016b; Burns, 2017). Reasons for this under-exploitation of data include: the effort required to manage large volumes of data, the inability of traditional statistical approaches to handle the complexities of ESs in global models, the practitioner's lack of awareness of the capabilities of DM, and the combination of skills required for getting the capacity to extract value from data Gibert et al, (2018). DS can be efficient to deal with this complexity. In DS processes, DM is in charge of detecting the patterns from environmental databases that will lead to useful information extraction, as well as helping identify the key parameters managing and controlling these complex ESs. DS is an emerging field that provides a wide opportunity to advance understanding of these systems.

However, inappropriate use and misunderstanding of the DM methods in DS processes should be avoided. Indeed, incorrect application of a DM method or misinterpretation of its results can inhibit the advancement of a science or lead to poor decision-making, which can potentially result in serious consequences for the environment. The risk of misapplication and misinterpretation of multivariate statistical data analysis methods was described by James and McCulloch (1990), and has been scaled to the wider and more complex DM field. In Baeza-Yates (2017) it is seen that various DM analyses on the same data can lead to contradictory results, which makes it critical to ensure that the proper DM method is selected to extract the right and not the wrong value from data. This paper provides an overview of how DM methods can be used to contribute to a better understanding of environmental systems, and how to correctly use them in real applications.

DM has its roots in the fields of statistics and artificial intelligence (AI). Too often, descriptions of DM methods found in the literature are highly technical, and use a language difficult for those without a strong statistical or machine learning background to fully understand. The end-user or practitioner (e.g. environmental scientist) is often more interested in understanding which types of problems can be solved with the method, what information is required as input, and how to interpret outputs, rather than how the actual method or algorithm internally works. One crucial issue for practitioners is how to select the right DM method for a particular problem. To serve this need, this paper provides end-user oriented descriptions of the most popular DM methods, together with guidelines on how to properly use them in DS processes, including advice on data representation, post-processing, and validation, interpretation and future use of the DM results.

3 Data Science and Data Mining

In Gibert *et al.* (2018) a discussion on the origins and nature of the DS field is presented. In that work, the authors provided their own viewpoint on DS, and describe it as "*the multidisciplinary field that combines data analysis with data processing methods and domain expertise, transforming data into understandable and actionable knowledge relevant for informed decision-making*, thus contributing to bridge Hammond's Fact Gap (Hammond, 2007), related to the disconnection between data and decisions.

It is well accepted that DS processes follow the main steps of data gathering, pre-processing, data mining, post-processing, and knowledge production, as original KDD, with a wider scope of embracing new information sources as inputs, like data streams, texts, images, videos… and the evenctual use of bigdata technologies when required. Currently, we are still far from having computational systems and software packages that follow this DS scheme in its entirety. Most commercial systems provide collections of several preprocessing, DM and support-interpretation tools, which have to be properly combined by the data scientist to build a correct DS process for each application. In fact, there are many difficult, technical decisions that the data scientist has to face in order to obtain the best outcome for a given dataset and objective. In fact, prior and posterior analysis require great effort when dealing with real applications. Pre-processing (data cleaning), transformation, selection of DM techniques and optimization of parameters (if required) are often time consuming and difficult steps, mainly because the approaches taken should be tailored to each specific application, and require interaction with experts or other stakeholders than the DM analyst. Once the data gathering (increasingly complex with new information sources like images, web pages, satellite data, IoT or social networks) and the pre-processing tasks have been accomplished (Gibert *et al.* (2016) propose a methodology that covers these steps), the application of DM methods can be relatively trivial and can be automated. Provided the right method is properly chosen, the application of the DM algorithm requires only a small proportion of the time devoted to the whole DS process. Interpretation of results is also often time consuming and requires much human guidance. If any of the selection, preprocessing or transformation steps are performed inadequately, the findings from the DM step may be erroneous or misleading.

There are several advantages of DS (including the application of DM techniques), which make it particularly appealing for the environmental domain. For example, DS supports:

- *Systematic and objective exploration and visualization of data*. For this purpose, DM techniques are an engaging alternative for several activities of the environmental scientist, when analytical/traditional methods fail, are too slow, or simply do not exist.
- *Improvement of data quality*. Through the preprocessing step, data deficiencies emerge and can be properly managed.
- *Identification of variables and characteristics of an ES* that are important to the problem of concern.
- *Modelling and system analysis activities.* Discovery of meaningful patterns in data can improve understanding of the system, and provide inputs to models for better simulation and prediction in ESs. This will also contribute to building more reliable intelligent environmental decision support systems (Poch *et al.*, 2004).
- *Discovery of patterns contained in large time series* can provide insight into how environmental systems have responded to changes through time, and may indicate how they may respond to future changes.
- *Integration of various knowledge sources and expertise*. DS paves the way to engage with domain experts and stakeholders during the whole process, including goal setting, data identification, and interpretation of results.

- *Production of new validated and transferrable knowledge* that can be shared and rapidly re-used among domain experts, due to the objective and formal nature of DS models for *communication about the ES*. The output of DS can include graphs or plots that help convey information about the environment in a clear and efficient manner to different audiences, from environmental experts or data scientists to the general population.

These are some of the most important contributions where the DS approach can help environmental scientists or managers understand or address real-world problems. Key reading material introducing the reader to essential points of DS are in Han and Kamber (2011), Whitten *et al.* (2011), Hastie *et al.* (2001), Larose (2004) and Parr Rud (2001).

4 A two-step methodology to determine an appropriate DM method for a given environmental problem

There is a considerable and increasing number of DM techniques available. However, not all of them are suitable for a given real-world problem. As said before, one of the most critical parts of the DS process is the selection of the appropriate DM technique.

Available commercial software packages provide researchers with access to a wide selection of techniques. However, these software tools rarely include intelligent assistance for addressing the decision about the kind of DM method to be used, or tend to do it in the form of rudimentary *wizard-like* interfaces, which make hard assumptions on the level of technical background of the user. There are not many works in the literature addressing these issues. Charest and Delisle (2006) have a first work in this direction. The authors are not aware of other works trying to solve this task, although Serban, (2013) describes state of the art in and desirable characteristics of "intelligent assistants" which help the user through the DS workflow. However, it is known that the end users tend to use a reduced set of well-known tools, and that data exploitation could be significantly improved by making a wider range of DM possibilities available to non-expert users (Hammond 2007; Cukier 2010). In this work, a proposal is presented to contribute to this issue.

The mechanism for choosing a DM method for a given problem is related to good knowledge of method properties. In fact, the output of the method must match the target question to be answered, the goals, whereas the inputs required by the method must be well aligned with the structure and properties of the available data. Figure 2 shows the idea that these two matches play a relevant role in the election of the DM technique. The proposed methodology first introduces the DMMCM as a reference decision map to browse through big groups of methods answering similar questions, and, at a second level, the DMMTs, with more specific information about those groups of methods, helping to narrow down to a box of the DMMCM containing the suitable type of technique.

The selection of the DMMCM is guided as follows:

1) Determine the main branch of the DMMCM: The questions associated to the decision nodes in the DMMCM will lead to one of the main branches of the map.
2) Identify the appropriate technique within the selected branch of the DMMCM: find the DMMT associated with the selected branch and identify a particular box in the map as the best DM method for the target environmental problem.

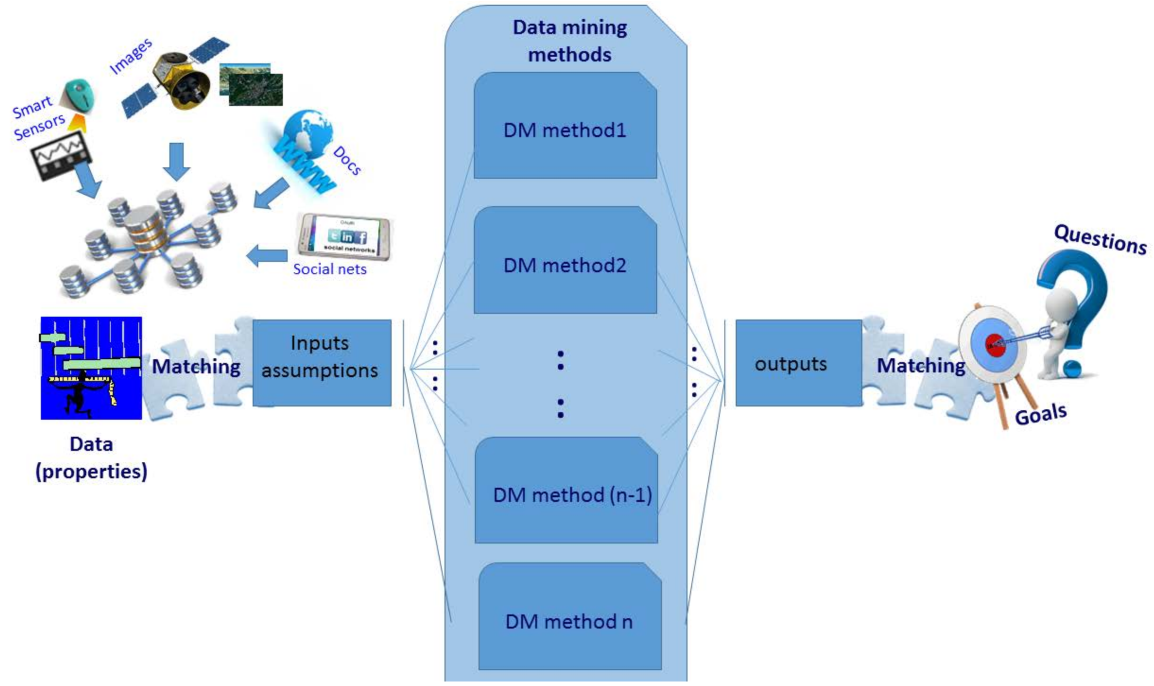

Figure 2. Choosing a DM method given a problem: Input assumptions of the method must match data structure, while output must match target goals and answer right questions.

5 Step one: choosing one group of DM methods by browsing the DMMCM

In Gibert *et al.* (2008b) a high level description of a number of most popular DM techniques was presented. The aim was to open black-boxes of DM methods to data miners and scientists to help them select the most suitable DM method for a given problem. The techniques presented in Gibert *et al.* (2008b) were grouped by technical proximity, as commonly done in the literature. In Gibert *et al.* (2010), a case-based reasoning (CBR) intelligent recommender was introduced for choosing the right DM technique, given the kind of question to be answered from the data. The CBR intelligent recommender re-used the past experiences in the GESCONDA tool (Sànchez-Marrè et al, 2010) to suggest the appropriate DM technique to apply.

However, through many years of experience on real KDD and DS applications, we have observed that both experts and data scientists do not consider the technical origin of the DM when choose the DM technique to use. Rather, we found that the final choice of technique primarily depends on the two following parameters:

- The main goal of the target problem;
- The structure of the available data set.

As a consequence, references providing long catalogues of DM methods organized by technical proximity among methods are not the best support to make this decision.

Thus, in Gibert *et al.* (2010) the first version of the DMMCM was introduced as a tool providing an overview of available DM techniques, organized according to the above criteria. The DMMCM intended to help non-expert data scientists to select the more suitable techniques for a particular application. Indeed, not all DM methods have been included in the DMMCM. Figure 3 shows a new version of the DMMCM, updated according to a detailed review of the most popular DM methods used in real environmental applications (Gibert and Sànchez-Marrè, 2012). The map organizes DM methods into four main branches that address four generic DM tasks. Each branch is suitable for answering a different type of question. Lighter cyan boxes contain DM methods coming from the AI field; cyan boxes contain methods from Statistics; and darker cyan boxes contain multidisciplinary methods. As it can be seen, methods from the same field are spread across different branches of the map, and viceversa, each branch includes methods from different fields. Thus, the DMMCM provides a new problem-oriented organization of methods, based on the kind of problem addressed, instead of on technical proximity of DM methods.

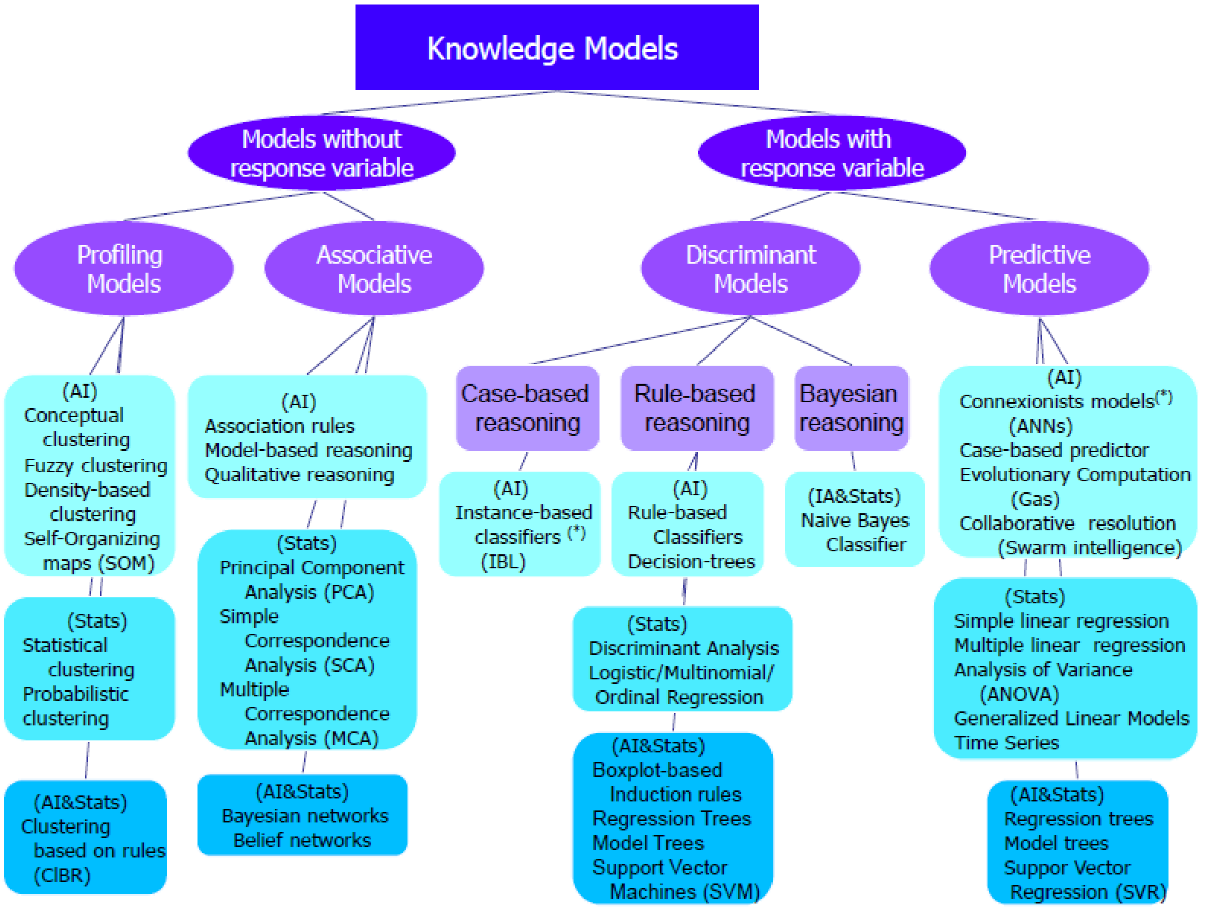
Knowledge Models
Models without response variable
Models with response variable
Profiling Models
Associative Models
Discriminant Models
Predictive Models
(AI) Conceptual clustering Fuzzy clustering Density-based clustering Self-Organizing maps (SOM)
(Stats) Statistical clustering Probabilistic clustering
(AI&Stats) Clustering based on rules (CIBR)
(AI) Association rules Model-based reasoning Qualitative reasoning
(Stats) Principal Component Analysis (PCA) Simple Correspondence Analysis (SCA) Multiple Correspondence Analysis (MCA)
(AI&Stats) Bayesian networks Belief networks
Case-based reasoning
Rule-based reasoning
Bayesian reasoning
(AI) Instance-based classifiers (*) (IBL)
(AI) Rule-based Classifiers Decision-trees
(IA&Stats) Naive Bayes Classifier
(Stats) Discriminant Analysis Logistic/Multinomial/ Ordinal Regression
(AI&Stats) Boxplot-based Induction rules Regression Trees Model Trees Support Vector Machines (SVM)
(AI) Connexionists models(*) (ANNs) Case-based predictor Evolutionary Computation (Gas) Collaborative resolution (Swarm intelligence)
(Stats) Simple linear regression Multiple linear regression Analysis of Variance (ANOVA) Generalized Linear Models Time Series
(AI&Stats) Regression trees Model trees Suppor Vector Regression (SVR)

Figure 3: Data Mining Methods Conceptual Map (DMMCM). Lighter cyan cells correspond to Artificial Intelligence methods; cyan cells to Statistical methods; darker cyan cells to hybrid methods.
(*) means main use of the method; adapted from Gibert and Sànchez-Marrè, (2012).

The DMMCM intends to guide the selection towards the most appropriate DM method for a real environmental problem. In the ***first step***, decision is focused on determining the main DM family of methods to be addressed. This is mainly related with the two criteria mentioned before: the problem goals and the structure of the available data. The identification of the proper DMMCM branch is done by answering two of the following three questions:

1. *Is there any response variable?*
   Response variables are those variables of interest to be explained by the DM model. They are determined through the environmental question to be answered, established in the project goals. Response variables are also referred to as dependent variables or target variables. Their behavior has to be described in terms of the other variables. This question determines the first split in the DMMCM and the next node inDMMCM to visit. The possible answers are:
   - **NO:** This means that a non-supervised scenario is faced, without response variable, where all the variables in the dataset play a 'symmetrical' role; the main goal is better understanding (*cognition*) of the target phenomenon, and getting a description of the global interactions is enough as a result.

     DM methods suitable for exploring interactions in these scenarios are on the left of the DMMCM.

     If the answer is NO, go to question 2.

   - **YES:** This means a supervised scenario is faced, and the main goal is related to *re-cognition* of the specific behavior of the response variable, which has to be explained by other variables. The result is a model expressing the response variable in terms of other variables (often named explanatory or independent variables).

DM methods suitable for this modelling are on the right half of the DMMCM

If the answer is YES, go to question 3.

2. *Is the interest in relationships among variables or in relationships among objects?*
   Objects, usually placed on the rows of the data matrix, are the units to be analyzed (also named instances or cases). They might be samples, locations, individuals, timestamps, etc., depending on the application. Variables, usually placed on the columns of the data matrix, represent the characteristics used to describe the objects.

   This question helps decide the second level of division of the left hand side of the DMMCM (node labelled 'Models without response variable'). Again, the answer to this question regards to problem goals. Two answers are possible:

   - **Variables:** Choose the branch labeled as *associative methods:* this branch contains DM methods describing global associations among variables.
   - **Objects:** Choose the branch labeled as *descriptive methods:* this branch contains DM methods that identify groups of related or similar objects, and provide the underlying concepts characterizing these groups.

3. *Is the response variable numerical or qualitative?*
   Numerical variables are measures on objects that can be continuous or discrete, whereas qualitative variables qualify the object and sometimes are referred as categorical. Nominal, ordinal or binary variables are particular cases of qualitative variables. In the particular context where one or more response variables exist, in this paper we will use the term *example* to refer to a row of the data matrix. An example is thus an object plus the value of the response variable (a classified object when the response variable is qualitative, an object plus some forecast when the response variable is quantitative).

   This question helps decide the second level of division on the right hand side of the DMMCM (node labelled as 'Models with response variable'). The question relates to the nature of the response variable itself, determined by the dataset inner structure. Two answers are possible:

   - **Numerical response variable:** Choose the branch labelled *predictive* methods. This branch contains methods that permit to predict the value of a numerical response variable under various formalisms and conditions.
   - **Qualitative response variable:** Choose the branch labelled *discriminant methods*. This branch contains classifier methods that permit to classify new instances in a set of predefined groups expressed in the qualitative response variable (often called class variable).

These three questions places the data scientist in one of the four main branches of the conceptual map. The next step involves identifying a specific method inside the selected branch (Figure 4: Illustrate the process)

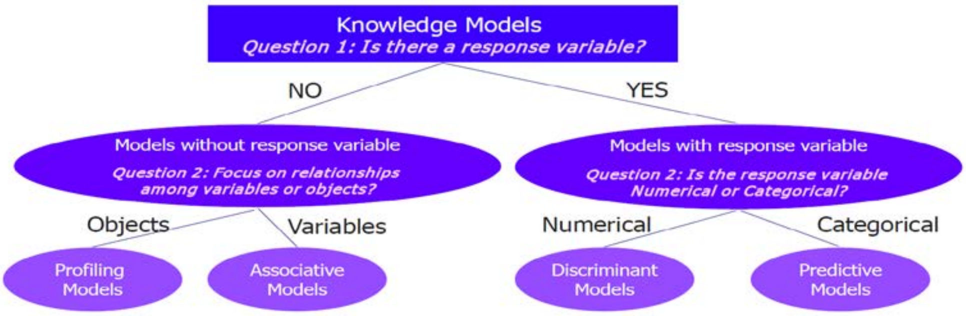
Knowledge Models
Question 1: Is there a response variable?
NO
YES
Models without response variable
Question 2: Focus on relationships among variables or objects?
Models with response variable
Question 2: Is the response variable Numerical or Categorical?
Objects
Variables
Numerical
Categorical
Profiling Models
Associative Models
Discriminant Models
Predictive Models

Figure 4: First Step of the methodology based on DMMCM map

6. Describing DM methods in a systematic way to assess decisions

Once a big group of DM methods has been selected, the most suitable method in the branch has to be identified. The methods in a given branch of the DMMCM provide different technical solutions to answer a single type of question. In a previous work, authors realized that the selection of a particular method among those available in the branch depends on the match between two aspects (Sànchez-Marrè et al, 2010).

1. *The dataset structure and the technical requirements of specific DM methods*. Input variables may be numerical, variables with normal distributions, independent variables, etc. Among all the available methods in the branch, select a method that does not include technical requirements violated by the data;
2. *The expected use of the obtained results and the characteristics of the output provided by the DM method*. Some methods generate outputs that are difficult to understand (e.g. complex equations), whereas others generate more intuitive results (e.g. trees). Depending if one wants to prepare a public report for general population, to support decision-making of a manager, to feed an automatic recommender system, etc., one might have preferences for a certain form of output.

At this level of decision, the properties of the DM methods have to be taken into consideration. It is important to know which kind of input is expected by each method and which kind of output is produced.

To help in this analysis, three categories of method requirements are proposed:

i) Technical requirements: These are critical requirements of the methods related to the intrinsic dataset structure accepted as input. Some examples include the following requirements:

- *Only numerical explanatory variables are accepted;*
- *Only ordinal (ordered qualitative) explanatory variables are accepted;*
- *Only nominal (non-ordered qualitative) explanatory variables are accepted;*
- *Only numerical/qualitative response variable is accepted.*

If data violates some of the technical requirements, the method will provide wrong results and should not be applied.

ii) Non-restrictive technical properties: These properties also refer to the data structure expected by the method; however, they are non-critical. Some examples include:

- *Recommended data size*
- *Variable's independence required*
- *Normal distribution of variable's required*
- *Linearity*
- *Requirement of no outliers*

If data violates some of the non-restrictive technical properties, the method loses performance rather than providing incorrect results. This means, for example, that algorithms suitable for small data sets are less suitable for very large datasets, but they still can be used. Similarly, algorithms that require independent variables will perform worse with highly correlated datasets, but they still can be used as well.

iii) Non-restrictive preference properties: These are method properties related to user preferences or project goals. Some examples include:

- *Speed of execution*
- *Interpretability of results provided*
- *Machine readable results provided*

In this case, mismatch between these characteristics and the user preferences or project goals indicates loss of usability of the discovered knowledge, but results provided by the method may be correct.

In the following section, the DMMTs are introduced for the DM methods included in each branch of the DMMCM, together with a discussion of the three types of method requirements introduced above, which will guide the reader towards identification of the DM method(s) appropriate for his/her problem.

7 Step two: identifying the appropriate technique within the selected DMMCM branch through the DMMTs

The descriptions presented are oriented towards the environmental scientist point of view. They intend to help a non-DM-expert user to discriminate, among a set of possible available methods, which one is the most appropriate to approach the target environmental problem. Thus, we present a set of structured templates, each one related to a branch of the DMMCM.

The structure of the templates includes:

- the main goals of a family of methods;
- a brief discussion of the main principles of the family;
- clear information on the kind of input required;
- technical assumptions to assess on data for data structure/method requirements match; and
- the type of output expected from the method.

We intend to disseminate knowledge about a broad range of DM methods, so that environmental scientists can better decide where to search for answers to their questions. Generally, there is a natural trend to use those methods better known, even though forced and intricate procedures may be required to adapt to the intrinsic nature of the problem. This research advocates the use of simpler and more appropriate alternatives when available.

Also, specific references are provided in each template where real applications of those methods are used to address environmental problems. In this paper, the classical use of each technique is provided, despite other possible uses in specific situations, which is out of the scope of this paper.

For every family of methods, a number of algorithms are available, each one with its own parameters. We do not intend to be exhaustive, nor to detail all the algorithms, methods or parameter settings involved in any family. An extensive review of DM tools for environmental science is given in Gibert *et al.* (2008b), and references to specific papers are given throughout the text.

The following information intends to help the final user identify the families of methods useful for a certain environmental problem. Once the DM technique that is suitable for the target problem is identified, specific search in the literature might be oriented to get more details on available algorithms, or accessible software suites could be inspected more in depth to search for proper options and parameter settings.

7.1 Profiling DM Methods: Clustering and Density Estimation

Response variable: No.
Main Goal: It covers the exploratory goal of finding distinct groups of homogeneous objects. These methods are suitable for discovering the underlying structure of the target domain. Thus, they belong to the group of *unsupervised learners*. They are very useful in the DM context, since the number of cases to be analyzed can be huge. Clustering can also be viewed as a density estimation technique by assuming that data was generated by a mixture of probability distributions, one for each cluster (e.g., Whitten *et al.*, 2011).

Principles: Clustering techniques are distance-based methods in which objects are compared among them and clustered together if they are close enough. Algorithms have different combinations of the distance or dissimilarity measure used for this comparison (see Gibert *et al.*, 2005; Núñez *et al.*, 2004; Jain *et al.*, 1999), and different criteria to decide how to cluster objects. In hierarchical clustering (one of the statistical clustering methods), more similar objects are clustered first. In partitional clustering, seeds of clusters are placed more or less randomly in the space, and objects are clustered to the nearest seed. Self-organizing maps, also known as SOM (Kohonen and Honkela, 2007) are a particular type of neural networks that build a 2D map where adjacency relationships among objects is preserved, and clusters correspond to subsets of homogeneous neurons. Fuzzy clustering provides a degree of belonging of objects to the clusters. Comparisons can be done directly, using distances or dissimilarities, or using more sophisticated concepts related to the quantity of information added by an object to a certain class, such as impact on the entropy of the class, and so on. Sometimes, prior expert knowledge can be introduced in the form of rules (Gibert *et al.*, 2010b) or ontologies (Gibert *et al.*, 2014) to introduce semantic information into the process, and get classes easier to interpret. Graph theory (Herrera *et al.*, 2015; di Nardo *et al.*, 2018) and social network theory (Campbell *et al.*, 2016; Brentan *et al.*, 2017a, 2018a) have also found applications in clustering. Density-based methods, like DBScan (Ester et al, 1996) or OPTICS (Ankerst et al, 1999) are computation-based methods detecting areas with higher concentration of objects and work well with non-globular clusters.

Scalable methods combine several strategies to divide the process into smaller pieces and cluster bigger datasets efficiently, like the CURE algorithm (Guha *et al.*, 2001), which hierarchically clusters an initial sample and uses class representatives to assign classes to the remaining objects in a partition-like stage. Heat maps use permutations of data matrix rows, and color coding of observations to visually find the classes (Wilkinson and Friendly, 2009).

Required Input: Data matrix with objects in rows

Standard Structure of output: Most algorithms produce a list of classes and the list of objects belonging to every class. Density estimation algorithms provide the parameters of the probability law associated to each cluster. Hierarchical algorithms can provide a dendrogram visualizing the sequence of aggregations performed.

Technical requirements:

- *Variables*: Traditional statistical clustering algorithms normally work with numerical variables. When the clustering criteria can be parameterized (distance, dissimilarity, logics or mixtures), compatibility measures open the door to cluster with heterogeneous variables (Gibert et al., 2005). Clustering algorithms coming from the AI field usually work with qualitative variables.
- *Number of clusters*: Many clustering algorithms (k-means and related) and all density estimation algorithms require the number of clusters to be known a priori, and used as input parameter. Hierarchical algorithms allow to discover the number of clusters a posteriori as a result of the analysis.
- *A priori domain knowledge*: Only required by clustering based on rules, when available. It can come in the form of knowledge bases (Gibert, 1998), or in the form of ontologies (Gibert, 2014).

Non-restrictive requirements:

- *Data size* is prohibitive with quadratic algorithms like hierarchical clustering, but they are non-order dependent. K-means or other linear algorithms can work with huge data sets. Scalable algorithms like CURE (Guha et al., 2001) are the most efficient in time. Some of them are order dependent. AI clustering methods usually work only with small datasets.
- *Metrics*: in hierarchical algorithms, dissimilarities perform worse than metrics because ultra-metric properties of the dendrograms are lost.
- Independence is not required in general.
- *Distributional requirements*: only apply to density estimation algorithms. Most require normality inside a class. Some algorithms can accept other probabilistic models. Most of them assume that all the clusters follow the same probability distribution, even though with different parameters.
- *Outliers*: Clustering methods that do not require the number of classes as an input are robust to outliers: they produce singletons integrated in the final solution. In the other cases, it is important to detect and properly treat them to avoid relevant deformations of the model, particularly in density estimation methods.
- *Missing values*: Must be previously treated unless the comparison measure accepts missing values (Gower, 1971). The alternative is to leave the incomplete data un-clustered.

Non-restrictive preference properties:

- Interpretability: Post-processing is often required to understand the meaning of the classes. Some software packages like SPAD provide helpful information about the contribution of the variables to the classes. KLASS (Gibert and Nonell, 2008; Gibert et al, 2005) provides the class panel graph (Gibert et al., 2008c), the traffic lights panels Gibert et al (2012) and the conceptual characterization by embedded conditioning (Gibert, 2014), which permit a visualization of the conditional distributions of the variables regarding the classes at a distributional or symbolic level, or a propositional description of the classes, and supports this understanding.
- Time consumption: Choose a scalable or linear algorithm if you need running speed. Quadratic algorithms can perform well for moderate datasets (with various thousands of objects and some hundreds of variables).

Further uses: clustering algorithms are descriptive methods that do not have a further predictive task associated. One of the most common uses is to associate a decision or an action to every class. Using these decisions or actions requires a mechanism to identify the class of a new object. Sometimes, the same decision associated to the class contains the conditions to be activated. In other cases, a discriminant problem has to be solved afterwards, using the class variable created by the clustering as a response variable.

Validation: For algorithms requiring the number of classes as an input, validation must ensure that the proposed classes are really distinct, and do not represent an artificial division of the domain. When a reference partition exists, this can be done using some quality index like the misclassification tax, Dun or Davies-Bouldin indexes (Brun et al, 2007)

However, being a non-supervised technique, there is no reference partition (otherwise clustering would be unnecessary), and validation is still an open problem. In that case, structural validity (Calinski and Harabasz, 1974) or the ratio of average distance within the clusters with respect to average distance between clusters may be useful where a numerical distance measure exists (Chieppa *et al.*, 2008), although it is redundant if the same criterion was used to build the clusters themselves, as is the case of using Ward's method (Ward, 1963). If density estimation has been used, the model goodness-of-fit can be objectively evaluated by computing the likelihood of a separate test set based on the mixture model inferred from the training data.

Once the structure of the clusters has been validated, stability of results may assess robustness. This can be done by performing multiple runs of the algorithm or other algorithms with slightly different parameters or initial values, and see which packs of objects keep always together (Lukačić *et al.*, 2005), or how much the results change among runs.

Finally, the decisive validation for assessing clustering usefulness is to validate understandability of results. This is usually done manually under expert guidance, but some research is being carried out to automatically induce concepts from classes which can support the detection of meanings of classes (Pérez-Bonilla and Gibert, 2007).

Applications and References: The use of clustering algorithms has been reported in various application fields for dimensionality reduction in stream flow time series (Zoppou *et al.*, 2002;), wastewater treatment plants (Gibert, 2010b and Zhao et al, 2012 use hierarchical clustering; Liukkonen, del al (2013) use SOM), cyclone paths identification (Camargo *et al.*, 2004), surface temperatures (Friedel, 2012), water quality in aquifers (Conti, Gibert, 2014), health status of wind turbines (Blanco *et al.*, 2018), and baseline air pollution levels (Gómez-Losada *et al.*, 2018). A hybrid combination SOM+k-Means Clustering was used to improve planning, operation and management of Water Distribution Systems in Brentan *et al.* (2018b), which can be easily extended to other environmental problems, etc. Graph theory (Herrera *et al.*, 2015; di Nardo *et al.*, 2018) and social network theory (Campbell *et al.*, 2016; Brentan *et al.*, 2017a,2018a) have been used to cluster a water distribution network into sectors so as to optimize these infrastructures' management. In Blanco *et al*. (2018) SCADA data is used to identify health profiles of wind turbines. Clustering has also been used in land use identification (Letourneau et al., 2012), and finding cropping patterns (Estel et al, 2016). Hybridation of case-based-reasoning and dynamic clustering was used to find spatiotemporal patterns on air pollution in Orduña et al. (2018). In Viaggi *et al*., 2013 clustering is combined with ANOVA to evaluate sustainability in farm-households. Applications in ecology include identification of behaviours in vegetation ecosystems (Tlidi et al, 2008), analyzing distribution of flora species like bromeliads (Brandão et al, 2009), finding groups of wildlife populations, like polar bears (Taylor et al, 2001), and discerning regions supporting similar assemblages of species (Hamilton *et al*. 2017).

7.2. Associative methods

Response variable: No
Main goal: Describe the relationships among the variables in a data set

Principles: Several families of methods respond to this aim. *Association rule* methods find out rules expressing regular correlation patterns among several variables hidden within the data set. The process of finding association rules could be computationally hard, especially if *brute-force* methods are used to obtain all the rules for all the possible combinations of variables and values on the right hand side of the rule. Mining algorithms seek more frequent combinations of variable-value pairs (*item sets*), overcoming a specified *minimum coverage* (or *support*). From the rules generated from an item set, those overcoming the specified minimum accuracy are kept. One of the most known methods is the *Apriori* algorithm (Agrawal and Srikant, 1994), which follows a generate-and-test methodology for finding frequent item sets, generating successively longer candidate item sets from selected shorter ones. Each different size of candidate items set requires a scan to the dataset and those under the minimum support are eliminated. Other methods try to decrease the number of scans to the dataset, like *FP-Growth* algorithm (Han *et al.*, 2004), which uses a frequent pattern tree to store a compressed version of the dataset in the main memory. Then, only two scans are needed to map the dataset into the FP-tree and, then, the tree is processed recursively to grow large item sets directly. *Bayesian and Belief networks* (Koller and Friedman, 2009) use conditional probability distributions and properties of chaining probabilities (or belief functions) to build a graph where nodes represent variables and links are annotated with the intensity of the association between nodes. They can be induced from a dataset. Most of the learning algorithms induce the skeleton of the underlying graph, and orient the rows on the basis of the conditional independences estimated over data (Needham and Bullpitt, 2007). They provide a powerful graphical model to represent very complex and highly dimensional probability distributions. Other graphical models from the same family are the *Markov random fields* or *the Hidden Markov models*. The latter are able to model dynamics from a statistical point of view. *Factorial methods* offer a completely different approach to analyze relationships among variables. They provide a low number of *factors* (linear combinations of the original variables) and project the original data set by keeping the main information and original object adjacency relationships. Factors represent a base conversion of the original variables. Variables can be projected over the factorial space themselves. Associated variables will place close in the projection, and this permits analysis of the packs of variables positively or negatively associated, or those behaving orthogonally. Principal Component Analysis (PCA) or multiple correspondence analysis (MCA) are the most popular factorial methods (Lebart *et al.*, 1984; Dillon and Goldstein, 1984).

Required Input: The data matrix with objects by rows

Standard structure of the output: The *set of association rules* mined, which have the minimum coverage/support rate and accuracy/confidence rate specified by the user. An association rule provides a set of values for some variables that appear together, and the variables on the left and right hand sides of the rules are exchangeable; every rule can be used to predict any of the variables. Each rule is shown with its coverage/support and accuracy/confidence values. Normally, rules are ordered from higher to lower coverage rules. Bayesian networks provide the visualization of the directed graph with some numerical annotation on the strength of the arrows. For factorial methods, the output is the set of factors, with the equations to build them upon original variables, as well as the graphical plane representation of pairs of factors, with the variable projections.

Technical requirements:

- *Variables:* The variables must be qualitative for association rules or simple/multiple correspondence analyses, Bayesian networks and MCA, and numerical for PCA.

Non-restrictive requirements:

- *Outliers:* Association rule mining algorithms are quite robust to the presence of outliers, because outlier values will not pass the minimum coverage cut in the generation of the association rules.
- *Missing values:* Missing values must be treated before building the models. Some software tools provide specific options to handle them, but it is better to know what the system precisely does.

Non-restrictive properties:

- *Interpretability:* The set of association rules are highly interpretable, and their meaning could be assessed by the user/expert. Bayesian networks are very intuitive and easy to understand by end-users, providing a powerful graphical model to understand the complex relationships among big sets of variables. The factors resulting from a factorial analysis are fictitious variables. The set of factors must be carefully interpreted on the basis of the original variables mainly contributing to the formation of the axes. Sometimes the interpretation becomes hard. However, visual inspection of the original variables projection over the factorial planes is very intuitive and permits to identify the associated variables very easily.
- *Time-Consumption:* Depending on the number of variables and/or the number of objects in the dataset, association-rule mining algorithms and Bayesian networks learning algorithms could show high computational times.

Further uses: associative methods are descriptive methods that are not linked to further predictive tasks. Commonly the results are used to identify packs of variables associated, and identify which variables to act to reduce (or increase) values of other variables of interest. Quantification of the impact of these decisions is called the what-if analysis and requires either further simulation or predictive models.

Validation: In association rule methods, the validation of the mined rules regarding *interpretability* could be done through the assessment of the meaningfulness of the mined associations by the user/expert. The *reliability* and *generalization* abilities of the association rules discovered can be evaluated with a new *test set* of objects. Bayesian networks are validated with experts. The goodness of the factorial methods is evaluated with the total inertia accumulated in the selected factors (directly related with the information quantity conserved), and with the contributions of the original variables to the axes.

Applications and References: Association rules have been widely used to find relationships in environmental domains. Formerly, Su *et al.* (2004; 2002) used association rules to extract relationships between environmental factors and fish distribution or fishing grounds. In geoscience and remote sensing, some works use association rules for geographic data (Rodman *et al.*, 2006), for landscape analysis (Ferrarini and Tomaselli, 2010), for finding relations between biophysical/social parameters and urban land surface temperature (Rajasekar and Weng, 2009) or adaptations to climate change (Lynam, 2016), and for image processing for urban environmental analysis (Du *et al.*, 2007). Regarding water topics there are some works that used association rules for coastal water classification (Pereira and Ebecken, 2009), lake sediments analysis (Annoni and Brüggemann, 2008), water resource management (Castelletti et al., 2007), biofilm development in water supply systems (Ramos-Martínez *et al.*, 2014), and fault detection in WWTP (Ruiz et al., 2011). Cloud screening for meteorological purposes was also investigated with *Markov Random Fields* in Cadez and Smyth (1999). In Robertson *et al.* (2003), *hidden Markov models* were used to model rainfall patterns over Brazil, producing interesting results. Air quality is analyzed in Zhu et al. (2012) with association rules mining, and with factorial methods in Sim-Siam et al. (2000). Bayesian networks are used to detect gas anomaly in coal mines (Wang et al, 2008) and risks of CO_2 storage in soil in Sousa et al. (2011). Soil quality is analysed in Khaledian et al. (2017) by combining PCA with clustering and PLS. Also, ecological applications include studies aim at better understanding vegetation (Ghosh et al., 2014; Nassr et al, 2018), nutrients (Gudimov et al., 2012) and bacteria (Liang et al., 2005); Ziliang (2002) uses association rules mining to identify security levels in a fuel plant.

Other uses: In this section it is worth noting that factorial methods are sometimes also used as a preprocessing step, where dimensionality reduction is intended. The original variables are reduced to a smaller set of factors preserving much of the information of the original dataset into a low dimension data matrix. In this situation, the interpretation of the axes becomes crucial, since DM is performed on the transformed (and reduced) matrix, and the interpretability of the final results depends on the interpretability of the factors used. Also, Bayesian networks are often used as predictive methods.

7.3. Discriminant methods

Response variable: A qualitative variable (class variable). The values or labels of this variable indicate the class of each example.
Main goal: To predict the class of a new object according to the values of the explanatory variables

Principles: A discriminant instrument, often known as *classifier*, must be induced from training data that shows the relationship between cases and the corresponding class. The nature of this discriminant instrument is quite variable, and the principles to build it vary accordingly. Most classifiers find combinations of variables with certain values that *describe the main composition* of a certain class. *Decision Tree* methods find the explanatory variables with higher discriminant power regarding the response variable and iteratively subdivide the training sample by building a tree where the internal nodes are associated with the variables, and its corresponding branches are the possible values of the variable. Every leaf contains one class identifier. Various criteria are used to select the discriminant variable at every iteration depending on the algorithm: the ID3 (Quinlan, 1986) algorithm and its subsequent version C4.5 (Quinlan, 1996), select the variable that produces a split with minimal entropy; the CART method (Breiman, 2017) uses the GINI rule. Rule-based classifiers try to build sets of classification rules with the conditions for belonging to the class. A classification rule is composed by the left-hand side, which is normally a conjunction of constraints on the values of some explanatory variables, and a class label as a right-hand side. Most algorithms, PRISM (Cendrowska, 1988), RULES (Pham and Aksoy, 1995), etc., are based on the idea of finding the variable and value that maximizes the quality (non-error rate, for instance) of the rule for a certain class, and specializing the rule with more conditions to gain precision. Others work by generalizing individual rules to sequentially cover more examples, e.g. RISE (Domingos 1996). The Bayesian classifier (Aguilera, 2011) uses the class probabilities estimated at the learning stage as *a priori* probabilities, and uses the Bayes formula to estimate the *a posteriori* probability, given the observed data. It finally assigns the most probable class. Naïve Bayes classifier is the most used Bayesian classifier, because it simplifies the computations involved by assuming some independence hypothesis among the explanatory variables.

Specific methods can combine numerical and categorical explanatory variables (e.g. CN2; Clark and Niblett, 1989), and a time component can also be introduced into the rule format. The boxplot-based induction rule method (Perez-Bonilla and Gibert, 2007) uses empirical conditional probability distributions to find the areas where ranges of variables do not overlap among classes, and combine several low coverage specific rules to provide a rule for the whole class. Bayesian classifiers use a frequentist analysis to estimate the empirical probabilities of the classes. Other methods find the algebraic equations (also a combination of variables) to identify *the frontiers* among the classes. This is the case of *Linear Discriminant Analysis (LDA)* methods (Lebart *et al.*, 1984), or *Support Vector*

Machines (SVMs) methods (Boser *et al.*, 1992; Vapnik, 1995). In the former, minimum least squares criteria are minimized to find the expression of the line(s) separating the classes. For SVMs (Christianini *et al.* 2000), criteria are based on transforming the original data matrix into a space in which classes can be linearly separable (or quasi-separable). In the transformed space, the line that better divides the space in regions (generally two) containing basically objects of a single class (generally two classes) is found. Various transformations are available in the form of kernel functions, and the user must specify which must be used in the training process. The discriminant is always a linear function of the transformed variables and the class (expressed as 1 or -1).

Required input in training phase: The data matrix containing all the examples in the training set. The examples show their explanatory variable values as well as their corresponding class label for the response variable (supervised methods).

Standard structure of the output in the training phase: The discriminant instrument (or the induced model) to be used for further predictions. These instruments could be barely different depending on the method used: decision trees produce discriminant trees, rule-based classifiers or boxplot-based induction rule methods produce a set of classification rules (probabilistic or not, depending on the algorithm). Bayesian classifiers produce a set of class probabilities. Statistical discriminant methods produce the discriminant equation and some threshold to decide the class accordingly. SVMs produce the discriminant equations or equivalently, as once the kernel functions are specified the form of the discriminant is fixed, some software tools only provide the coefficients of the linear combination constituting the discriminant equation. All those discriminants are oriented to provide a prediction for the class of a new object under different forms.

Validation: The validation of the discovered discriminant instrument or model is assessed through the goodness of the discrimination process (performance parameters), and some other parameters (size/time efficiency parameters). For a decision tree the size of the tree and the classification accuracy in the discrimination process are the two main validation parameters. In rule-based classifiers the three main parameters are size of the rule set, classification accuracy of the rules, and coverage of the rules. Coverage is the percentage of the number of examples which are classified by the rules. In Bayesian classifiers, the main validation parameter is the classification accuracy. In addition to the global classification accuracy, it is very common to assess the accuracy of each class label separately (accuracy by modalities). This way, it can be seen where errors are occurring. All the accuracy rates are computed and visualized in what is commonly known as a *confusion matrix*. For binary classifiers, as SVMs, ROC curves may be used (Hanley *et al*, 1982). The distribution of input data should also receive consideration, as many classification algorithms tend towards predicting the majority class (the one with more objects). An in-depth discussion of this topic can be found in Weiss and Provost (2001).

All the performance parameters are estimated using some unseen examples. These set of examples are known as the validation set or the test set, depending on different terminology and/or scientists. This validation scheme is known as simple validation. N-fold cross-validation if the most frequently used.

Required input in the predictive phase: A new object to be classified, where the response variable is unknown

Structure of the output in predictive phase: The output is the predicted class for the target-object, which is obtained by applying the induced discriminant model. For decision trees, the object must go through the tree, following the branches indicated by its values in each variable, till a leaf is reached, and the class label found. For rule-based classifiers, classification rules must be evaluated with object values, and the object is classified according to the right hand side of the first satisfied rule of the rule set. For boxplot-based induction rules, various criteria are applied (like maximizing coverage and/or confidence, or voting) to decide which of the satisfied rules will be used to determine the final class. Algebraic equations provided by LDA must be computed with the object values and the result compared with some thresholds to find the class. For SVMs the equation representing the linear discriminant must also be computed with the object values. The sign of the result indicates which of the two classes is predicted.

Technical requirements:

- *Variables*: the explanatory variables must be qualitative for decision trees, for most of the rule-based classifiers and for the Bayesian classifiers or discriminant correspondence analysis. They must be all quantitative for discriminant equations. Box-plot-based-induction rules can deal with both numerical and continuous variables, as well as SVMs, provided that the proper algorithm is used in this case.
- *Response variable*: Only binary for SVMs and discriminant methods.

Non-restrictive requirements:

- *Variables distributions*: LDA assumes normality of explanatory variables.
- *Homocedasticity*: LDA requires equal variances.
- Independence: The Naïve Bayes classifier assumes that the variables are conditionally independent given the class label values.
- *Outliers*: In general, these methods are quite robust to the presence of outliers, especially decision trees and rule-based classifiers. Outliers are managed as specific objects, which will generate specific classification rules or specific tree branches.
- *Missing values*: Missing values must be treated before building the models. Some software tools provide specific options to handle them, but it is better to know what the system precisely does. Classification cannot be provided if the new example contains missing values.

Non-restrictive properties:

- *Interpretability*: A decision tree model is highly interpretable and meaningful for a user. A set of rules could also be interpretable by an expert. Bayesian classifier models (i.e. probabilities) are not easily interpretable at all, as well as algebraic discriminant equations.
- *Time consumption*: Depending on the number of variables and/or the number of objects in the dataset some models could have a high computational cost in the training step, like rule-based classifiers or decision trees.

Applications and References: Classification techniques are be very popular. For example, in Spate *et al.* (2003) rainfall intensity information was extracted from daily climate data; Troncoso et al (2018) predicts monsoon; Ekasingh *et al.* (2005) discusses the classification of farmers' cropping choices using decision trees; in Sweeney *et al.* (2007) mosquito population sites are categorized, while in Stadler *et al.* (2006) decision trees are applied in a European plant life-history trait database. Agriculture-related applications include Holmes *et al.* (1998) for apple bruising, Yeates and Thomson (1996) for bull castration and venison analysis, and the Michalski and Chilausky's soybean disease diagnosis work (Michalski and Chilausky 1980), which is a classic benchmark problem in machine learning. Considerable efforts are recorded in the water-related fields, using rule-based reasoning (Zhu and Simpson, 1996; Dzeroski *et al.*, 1997; Comas *et al.*, 2003; Spate, 2005; Ramos-Martínez *et al.*, 2014), decision-trees (Kokotos et al., 2011), regression-trees (Dzeroski et al., 2003), Support Vector Machines (SVM) (Kanevski et al., 2002), case-based reasoning (Martínez *et al.* 2006 ; Wong *et al.*, 2007), regression trees (Dzeroski and Drumm, 2003) or hybrid techniques (Cortés *et al.*, 2002, Yang et al. 2012). In the study of air quality, classification has been used for air quality data assurance issues (Athanasiadis and Mitkas, 2004) and the operational estimation of pollutant concentrations (Athanasiadis *et al.*, 2003; Stebel et al, 2013; Yeganeh et al, 2012). Land use has been used with decision trees (Schneider et al, 2012), logistic regression (Li et al, 2009). Regression forest have been used to detect pesticides in fruits (Holmes et al, 2012). Bayesian networks for predicting coral bleaching (Krug et al, 2013) or fish recruitment (Fernandes et al, 2013). Rule based classifiers to predict rockburst in longwal or coal (Sikora and Wróbel, 2010).

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 in Ramesh and Ramar (2011), and other soil properties in McKenzie and Ryan (1999), which also used regression trees and other techniques 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; the potential of DM techniques has been shown in (Athanasiadis *et al.*, 2005) where statistical and classification algorithms in air quality forecasting are compared; sudden death of oak trees was modelled with SVMs in Guo *et al.* (2005).

7.3.1. Case-Based Reasoning (CBR)

Response variable: A very common use of case-based reasoning (CBR) is for *discriminant* purposes, with one qualitative response variable (class variable), although the model accepts one or several response variables, and each one could be either quantitative or qualitative. When a single qualitative response variable is used, CBR acts as a case-based classifier, commonly known as nearest neighbor classifier (NN). When one or more numerical response is/are used, CBR acts as a (multi-response-)*predictive* method, to forecast (case-based predictor).

Main goal: To recommend a solution (or a list of the best solutions) for a new problem on the basis of past experiences. It is important to highlight that CBR is much more than a simple data mining method. CBR is a general problem-solving, reasoning and learning paradigm (Kolodner, 1993) within the AI field.

Principles: The general assumption of this model is that "*similar problems have similar solutions*". CBR solves a new problem (new case or experience) by adapting the solution of one or several previous similar problem(s) (past experience or case), which are in the memory (case library or case base) of the system. This way, solutions are not built up from scratch, but taking advantage of what has been done in the past, decreasing the time of problem solving. In addition, CBR systems learn from each new experience (enlarging the case library), and its performance increases along time. The assumption is that similar cases should have similar values of the response variable(s): given a new case *(query-case)* with unknown values for some variables, similar cases are selected from the case library, and the response variable(s) of the new case is estimated according to the values of its neighbors. Additional use of domain knowledge can be used to combine the solutions of the neighbors in a new solution for the query-case. A great advantage of this paradigm is that it is robust to model changes or trend changes, since the case library updates over time and accumulates experiences on new behaviors, which can be retrieved in the future.

Required input in training phase: The training phase consists of entering into the case base a sufficiently representative set of cases, or past experiences. A data matrix with this collection of cases is required. The structure of the case library can be indexed. In this case, the structure of the index must be provided.

Standard output in training phase: CBR does not produce an explicit model describing the system behavior. In fact, it is known as a lazy learning technique. The model is implicitly composed by all the cases or experiences stored in its case library. The output of the training phase is a case library properly structured.

Validation: The system competence must be validated. It depends on the quality of the case library. Cross validation techniques can be used, but some evaluation of proposed solutions is required. For that, expert evaluation about solution quality is often required. If the proposed solution can be applied and the system can provide feedback on the goodness of the solution, automatic validation can be performed, even though this is not the common situation.

Required input in the predictive phase: Using the model means to obtain solutions for a new problem or predictions for unknown variables in a new case. The input must be the query-case, for which response variables are unknown. It is flexible enough to change the response variables from one query to another provided that the description of the case and solution variables can be modified dynamically, and the similarity works with non-solution variables.

Structure of output in the predictive phase: The system provides a proposal for the unknown variables, together with a list of more similar cases in the case base, from which the final solution is built. A relevance coefficient for the neighbor cases is also provided.

Technical requirements:

- *Variables*: Both qualitative and quantitative variables are allowed, provided that the correct similarity is used to compare cases.
- *Case library structure*: The case library structure must be good enough to allow fast retrieval processes.

Non-restrictive requirements:

- *Data size*: a flat case library must be small for good performance; for large case libraries, an indexed/ hierarchical structure is needed for reasonable computational time. Otherwise, time could be prohibitive. Performance depends more on the representativeness of cases in the case library rather than on its size. Big case libraries do not perform well if they contain lots of redundant cases.
- *Metrics:* it must consider the various types of variables used in case description. A metric structure is not strictly required, and the process can also perform well with similarity measures.
- *Distributional requirements:* irrelevant
- *Independence*: this paradigm is particularly powerful with general situations including complex interactions among variables, which are difficult to model explicitly.
- *Outliers*: robust to the presence of outliers in the case library. However, when the query-case is very different from information contained in the case library, the proposed solution must be invalid.
- *Missing values*: Missing values can be a problem for retrieval tasks, for similarity assessment (unless the similarity measure used can deal with them), and for adaptation tasks. Missing values management techniques must be applied to run properly.

Non-restrictive properties:

- *Interpretability:* There is no explicit output model. Depending on the case library structure, the implicit case library model (hierarchical structure) can be easily interpreted. Additionally, in the discrimination/prediction step, the set of similar cases is normally given as a partial output, which can give a very good interpretation of how the proposed solutions are derived.

- *Time consumption*: good for small case libraries or large/huge case libraries with an indexed case library.
- *Robustness to model changes*: CBR shows robustness to model changes as new cases implicitly introduce new behaviors into the case library, and permits adaptation of the solutions to the new behavior.

Application and references: In environmental sciences, CBR has been applied in various areas with different goals, because of its general applicability. CBR has been widely used in environmental domains such as meteorology (Riordan and Hansen, 2002; Koo et al, 2013), forest fire fighting (Avesani *et al.*, 2000), agroforestry management (Tourigny et al, 1998), rangeland pest management (Hastings *et al.*, 2002), land use (Li et al, 2009), quality air prediction (Kalapanidas and Avouris, 2001), mapping species an habitat (Remm, 2004), solution of local environmental problems (Kaster *et al.*, 2005), drilling in fossil fuels (Shokouhi et al, 2014), supervisory systems for waste water treatment plant (WWTP) management (Rodríguez-Roda *et al.*, 2002), and water management in general (Popa et al, 2011, de Araujo et al, 2004).

Cautions: The main assumption of CBR models that "*similar problems have similar solutions*" must hold in the domain. Otherwise, the accuracy and quality of the proposed solutions cannot be guaranteed. Also, the quality and the scope of the case library is critical, as well as the choice of an appropriate similarity measure and adaptation technique/s (see Núñez *et al.*, 2004).

7.4 Predictive models

Response variable: Numerical
Main goal: To find an algebraic equation relating the response variable with a set of explanatory variables, accepting a probabilistic error fitting some probabilistic distribution.

Principles: These methods try to minimize the mean square error (MSE) or some related loss function that compares fitted and observed values. They provide the coefficients of the optimal equation. The underlying algebraic structure depends on the nature of the explanatory variables and the nature of the model. Numerical variables correspond to axes in some linear space. Qualitative variables are previously decomposed into *dummies* to be entered into the models. *Linear Models* search for hyperplanes. The *General Linear Model* (Christensen, 2002) is a general formulation including most known cases as *Multiple Linear Regression, ANOVA* or *ANCOVA* being particular cases. The *Generalized Linear Model* (GLM) uses an internal link function that can have multiple forms and produce, as particular cases, *Poisson regression* or *logistic regression*, among others (Myers *et al.*, 2002). For *Time Series Analysis*, the optimized function relates a single response variable with itself in the past; this means with previous observations of the same variable. Time series models use various functional relationships among observations (Hamilton, 1994; Lerner et al, 2004). Some classical machine learning algorithms from the discriminant family, extend to models that also predict numerical variables. This is the case of Classification and Regression trees (Breiman, 2017) or Support Vector Regression (Basak et al, 2007).

Required input in training phase: the model is built by means of finding best estimates for the coefficients of the target equation. The required input parameters are in *the training* data matrix, which must contain a set of examples and the parameters of the model. In basic regression models, there is a coefficient to be estimated for every variable. More complex models include interactions between terms. Stepwise-like techniques can help decide which interactions must be considered in the model, thus avoiding the combinatorial problem of using a complete model, and having more parameters to estimate than examples in the data matrix.

Standard output in training phase: An algebraic equation (normally a linear combination of the explanatory variables or their dummies, if original variables were qualitative) relating explanatory variables with the response variable. This equation contains only the significant variables, those proved to be relevant for the response variable.

Validation: First of all, significance of the coefficients of the model are assessed by means of the corresponding statistical tests. Once the effective terms of the final equation are clear, goodness of fit indicators, like the determination coefficient (R^2), or the F statistic in the ANOVA, or deviance in the general linear model, can assess the technical global performance of the model. Graphical analysis of the residuals (Moore and McCabe, 2012) must complement this analysis by identifying violations of the technical assumptions of the models, like normality, linearity, independence, or the ones required for the different models, as well as other situations that can determine abnormal estimation of the model parameters, like outliers or influential observations. This part is rather difficult and poorly explored for some models like logistic regression. Validation of technical assumptions of the models is critical for correct interpretation of goodness of fit indicators, which are calculated under these assumptions. Wrong models can provide extremely wrong predictions and have unexpected dramatic consequences, even with high values for the determination coefficient. It is important to take care on using the right indicator for every method and type of data.

Required input of the predictive phase: Using the model involves applying the estimated equation for predicting the value of the response variable in new query-objects. The input is a query-object with unknown response variables.

Standard structure of output in predictive phase: The predicted value for the response variable. It is obtained by simply applying the equation to the values of the explanatory variables in the query-object.

Technical requirements:

- *Variables*: Various configurations are allowed depending on the model. All forms of regression work with numerical explanatory and response variables. ANOVA relates a numerical response variable with a set of qualitative variables, converted to dummy variables (Wooldridge, 2009), that is, to a set of binary variables, indicating presence or absence of a modality, one per modality. ANCOVA generalizes ANOVA to include also quantitative explanatory variables. Logistic regression relates a probability as a response with a set of numerical regressors, and is useful to predict qualitative variables, provided that the method permits to estimate the probability to belong to every modality. The GLM subsumes many particular cases like ANOVA, ANCOVA or linear regression. Users must take care to apply the correct model according to the types of variables used.

Non-restrictive requirements:

- *Data size:* These methods perform well even with a big number of objects in the dataset.
- *Dimensionality:* Some models are not feasible for a large number of variables, particularly if they are qualitative with many modalities, because running times exceed reasonable ranges.
- *Metrics:* They are not distance-based methods.
- *Distributional requirements:* Regression methods, ANOVA, and ANCOVA require conditional normality. If it does not hold, inference in the model is invalidated and significance of terms in the final equation cannot be properly assessed. Poisson regression requires Poisson conditional distribution. Every model has its own distributional requirements, and the final user must be sure that data holds these assumptions for a valid model.
- *Independence:* Most of these methods assume uncorrelated data. Multi-collinearity can be a problem in some contexts. In more general models, interaction terms can be introduced to model these correlations, but this diminishes quite a lot the interpretability of the model.
- *Linearity:* it is a condition for all linear regression models. Quadratic or polynomic models will have other technical requirements that must be satisfied by the data. Otherwise, the model can be built, but the quality of the predictions is not guaranteed at all. As an example, it makes no sense to find the best linear approximation for a quadratic model.
- *Outliers:* In general, these methods are far from being robust to the presence of outliers, which can trigger significant perturbations in the final coefficients. Outliers must be previously identified, properly diagnosed, and properly treated before building the model. Treating outliers is not synonym of eliminating them from the analysis. Proper treatment can mean enlarging the sample with more data in the area. Prediction can fail if it is asked for an object very far from the model.
- *Missing values:* Missing values must be treated before building the model. Some software packages provide specific options to handle them, but it is better to know what the system does precisely. Predictions cannot be provided if the query-object contains missing values.

Non-restrictive properties:

- *Interpretability:* The regression equation is provided. Specialists may be able to interpret the meaning of the coefficients in these equations, but this is not easy for the general user. These models are more interpretable than neural networks or genetic algorithms, but less interpretable than decision trees or induction-rules.
- *Time consumption:* Prohibitive for complex models with many coefficients and many interactions. For basic models, efficient performance in both estimation phase and predictive phase is obtained.

Applications and references: A variety of regression models have been used for environmental problems, for example to predict concentrations of pollutants in WasteWater Treatment Plants (WWTP) (Dürrenmatt, 2012), and to predict stormwater quality (Sun et al, 2012) or heatwaves (Herrera et al, 2016) or macroinvertebrate abundance in rivers (Forio et al, 2018). Spatiotemporal models based on regression were used to predict rainfall (Kamarianakis et al, 2008), bioremediation time in soil or water after petroleum contamination (Norris et al, 2018), and opal occurrence (Landgrebe et al, 2013). Fuzzy time series have been used for analyzing air quality (Domanska and Wojtylak, 2012). Support vector regression was used to predict sediment concentration in rivers (Jain et al, 2012) and classical mechanistic models were used by Cazarez-Candia et al. (2005) for temporal prediction of flow parameters in oil wells. Classification and Regression Trees have been used to predict species habitat (Fukuda et al, 2013).

7.4.1. Artificial Neural Networks (ANNs)

Response variable: the most popular use of ANNs is for a numerical response variable. However, there are ANNs prepared to work with qualitative response variables.

Main Goal: To predict a response variable. The ANN is a black-box implicit model which works as a universal function aproximator that can fit any kind of function to relate explanatory variables with the response variable, even if it is a complex non-linear combination of explanatory variables. ANNs are often used in non-linear regression and classification.

Principles: ANNs are a metaphor of brain functioning, where a network of neurons strengthens or diminishes their interconnection on the basis of new input signals provided by various biological sensors. Thus, an ANN is conceived as a graph of neurons that can be grouped into layers and adjust their weights incrementally depending on the signals received from other neurons and their own reactivity (represented by their activation function). They need to be trained with incremental inputs to converge to a stable structure usable in the long-term for predictive purposes. The architecture of an ANN determines its behavior and the kind of functions it can fit. The simplest ANN is the Perceptron, a feedforward neural network with a single neuron, which, in fact, simulates a logistic regression (Rosenblatt, 1956). Widrow and Hoff (1960) introduced the single layer neural network. Non-linear functions can be fitted with hidden layers (Bello, 1992) and non-linear activation functions (multi-layer perceptrons) or radial functions (radial basis function network). Recurrent ANNs permit dynamic prediction (Hochreiter and Schmidhuber, 1997).

Required input in training phase: The training dataset is the input for training the network. It contains a set of examples (an object with the values of the explanatory variables plus the value of the response variable). The various algorithms correspond to combinations of the architecture of the network (number of hidden layers, number of neurons per layers, a layer being a group of non-connected neurons that share input), connectivity of neurons (unidirectional (feedforward) or bidirectional (recurrent)), the activation function of the neurons (lineal, threshold, sigmoidal, cosine, Gaussian, etc.), the training algorithm itself (backpropagation algorithms, with or without optimization of neuron weights: descending gradient, based on conjugated gradients, quasi-Newton methods, Levenberg-Marquardt method, etc.), or the static or dynamic structure of the network, which determines whether structure can change over time and how new neurons or connections are added. Fuzzy neural networks proposed by Jang (1992) use fuzzy logic in the training algorithm, and there are multiple variants in this field.

Standard output in training phase: Once the network is properly trained, the weights of the connections become determined. The result of the training phase is the definite topology of the network.

Validation: As other supervised models, the validation can be done through the *misclassification tax*, or the number of examples wrongly classified by the method, preferably over a test dataset, independent of the training dataset. Cross-validation techniques are preferred to simple validation. Supervised neural networks that use an MSE cost function can use formal statistical methods to determine the confidence of the trained model. This model tends to overfit the training sample, and to provide models difficult to be generalized to other samples. Validation must ensure that the overfitting phenomenon is under control.

Required input in predictive phase: A new object with the values of the explanatory variables for which the response variable value has to be predicted.

Standard structure of output in predictive phase: the predicted value for the response variable. It will be a numerical value for numerical responses, or a binary value for binary responses. In this last case, a final threshold function is used to binarize the result. When the response is categorical with more than two modalities, arbitrary thresholds are used to discriminating among categories.

Technical requirements:

- *Variables*: ANNs usually handle numerical input data, but it is possible to find approaches that handle binary input data. Categorical variables with multiple categories are tedious to handle.

Non-restrictive requirements:

- *Data size*: The dataset used for training should be big enough to be representative to avoid biases in training. However, not too big to generate *overfitting* (a too specialized fitting of the observed cases that would not be generalizable).
- *Independence*: Not required in general
- *Distributional requirements*: Not necessary
- *Metrics*: ANNs are not distance-based methods.
- *Outliers*: ANNs are considered robust models able to handle outliers or noisy data.

Non-restrictive properties:

- *Interpretability*: ANNs are not interpretable. In the literature they are known as black-boxes since they are implicit models. This means that they can be used to obtain reliable predictions, but the genesis of the model, as well as the reasons for a certain prediction, remain unknown by the end user. Although it is possible to extract the mathematical equation implicitly used in an ANN to compute the predicted value, it is related with the topology of the network, and do not help too much to interpret the result.
- *Time consumption*: ANNs training time depends, among other things, on the architecture of the network. Time consumption increases with the number of hidden layers the ANN contains and the number of connections. Moreover, the training time depends on the training data set time. For a very large amount of data or a big network, some methods become impractical. In general, it has been found that theoretical results regarding convergence are an unreliable guide to practical application. The prediction time is very quick, even for big and complex ANNs.

Applications and References: Numerous applications have been developed for ANN; as an indication we mention the works by Kralisch *et al.* (2001) and Almasri and Kaluarachchi (2005) on nitrogen loading; Mas *et al.* (2004) on deforestation; Tasadduq et al (2002) on surface temperature in desert; Tirelli et al, (2011) on flora abundance, Carvalho et al, (2008) on biological diversity in coastal water; Bartoletti et al, (2018) on rainfall; Babovic (2005) on hydrology; Izquierdo *et al.* (2006) on detection of anomalies in water supply systems; Brentan *et al.* (2017a) on water demand forecast; Kusiak et al, (2013) on pumping in WWTP; those of Belanche *et al.* (2001), Gibbs *et al.* (2003) and Gatts *et al.* (2005) on water quality; Kurt and Oktay(2010) on air quality; and Pacifici et al, (2009) and Taghavifar et al, (2013) on land use and soil. The discussion on nonlinear ordination and visualization of ecological data by Kohonen networks; ecological time-series modelling by recurrent networks Recknagel *et al.* (2002); along with the application of Dixon *et al.* (2007) in anaerobic wastewater treatment processes. Almasri and Kaluarachchi (2005) on nitrate distribution, Recknagel et al (2002) on algal bloom and Millie et al. (2012) on miroalgal abbundance. In Huang et al, (2001) permeability in petroleum reservoirs is predicted. From multilayer perceptron, to recurrent ANN, sometimes using neurofuzzy approach, and in most of the cases combining with PCA or decision trees, or some genetic algorithm (Llanos et al 2013), or GIS.

7.4.2. Evolutionary computation

Response variable: The most common use of Evolutionary Computation (EC) (Holland, 1992) is for predicting a single quantitative response variable by means of quantitative explanatory variables. However, this model can also be used with several response variables, even qualitative and of mixed nature

Main goal: To find the optimal value of the response variable(s) together with the values of the explanatory variables producing it. In the case of multi-objective problems, the so-called Pareto front is sought.

Principles: During the last two decades, some algorithms that imitate certain natural principles have been used in various aspects of Environmental Sciences. These algorithms perform a type of search that evolves through successive generations, improving the characteristics of the potential solutions by means of mechanisms inspired by biology. EC is a bio-inspired approach mimicking natural selection or behavioral processes in biological populations.

The most popular examples are *Genetic Algorithms* (GAs), including *Genetic Programming* (GP), and *Swarm Intelligence*. A GA (Goldberg, 1989) is a biologic random search technique. It begins with a set of randomly generated individuals, called the *population*. Each *individual* is coded as a string over a finite alphabet (commonly a binary code). The next generation of the population is produced after some *genetic operators* (selection, crossover, mutation, etc.) have been applied to some probabilistically-selected individuals. Each individual is rated according to an evaluation function, named the *fitness function* which is correlated to their associated probabilities. Selected individuals for reproduction are combined using a randomly crossover point among the positions in the string. The offspring are created by crossing over the parent strings at the crossover point. Finally, each new individual is subject to random mutation of some positions with small probabilities. Since best individuals are selected and reproduced, convergence to the best individual (the optimal) is expected. GP (Michalewicz, 1996; Koza, 1992) is a subclass of evolutionary search methods, where the population is composed by individuals who are computer programs or fragments. The evolution of such a population can produce a new computer program to perform a user-defined task.

Swarm Intelligence imitates the collective behavior of a group (swarm) of individuals. Individuals are endowed with personal intelligence. In addition, some kind of collective intelligence emerges from grouping and communication among individuals, resulting in more successful performance of the whole group. There are two popular swarm-inspired methods in computational intelligence: ACO (ant colony optimization) (Dorigo *et al.*, 1996), inspired by the foraging behavior of ants, and PSO (particle swarm optimization) (Kennedy and Eberhart, 1995), inspired by the social behavior of flocks of birds or schools of fish.

Hybrid platforms that use several metaheuristics (Montalvo *et al.*, 2014), with self-adaptive abilities (Izquierdo *et al.*, 2016a) and able to exploit knowledge injected to the model (Izquierdo *et al.*, 2016b) both from the expert know-how in the field and from mining tasks performed during the evolution process itself, have also shown great interest, because of their improved search abilities.

Required Input: Data required in EC applications consists in a clear mathematical setting of the problem, including a neat distinction between targeted objectives and constraints to meet. To meet this purpose, the sometimes bulky data describing the material elements of the problem is indispensable, and may be suitably stored in appropriate databases. Sometimes, mainly to avoid frequently arbitrary constraint penalties, constraints are transformed into objectives to meet, what necessarily develops into multi-objective optimization (Montalvo *et al.*, 2014).

Structure of output: EC does not produce an explicit model describing the system behavior. The model is implicitly composed by the set of individuals (population) at each step of their evolution and their fitness function. However, frequently, only the individual of the last generation, those embodying optimal solution(s) are of interest. In the case of multi-objective optimization, those non-dominated individuals constitute the Pareto front. Reasons, frequently of non-technical character, must be *a posteriori* used to select one (or various) of those non-dominated solutions as the final solution to be implemented. Since this permits a certain trade-off between objectives of varied nature, commonly political, entrepreneurial economic, etc., reasons are used to make the last decision. However, some recent proposals suggest using multi-criteria decision-methods to select the final solution(s) (Reynoso-Meza *et al.*, 2017, 2018; Carpitella *et al.*, 2018).

Validation: Conditions under which EC algorithms perform well are not easily identified. Typically, evolutionary algorithms (including swarm intelligence) find suboptimal solutions; there is no guarantee of finding the global solution of the problem in hand. As a result, validation is performed based on either benchmarking problem available in the literature on specific repositories, or based on sound knowledge and experience regarding the problem in hand, that lead to consider acceptable a solution from a technical point of view.

Technical requirements:

- *Variables*: Both qualitative and quantitative variables are allowed. However some quantitative variables are better to thoroughly explore the search space. Instances where various types of variables coexist are frequent (Izquierdo el al. 2007b, 2012).

Non-restrictive requirements:

- *Data size*: For large/huge data size, the computation time could be prohibitive. This drawback may be somehow indirectly alleviated through various methods, such as the use of warm solutions (Brentan *et al.*, 2018), or by orienting the search through the injection of knowledge, as said before.
- *Missing values*: Missing values can be a problem both for the evolution of the population, for the fitness function assessment and for the operators' implementation. Missing values management techniques must be applied to run properly.

Non-restrictive Properties:

-
- *Time consumption*: Computation time could be highly reduced because EC is able to use parallel processing.

- *Optimization accuracy***:** EC could explore the whole search space and escape from local optimal to which other optimization techniques cannot.

Applications and References: Haupt and Haupt (2002) contains an overview of some application of GAs in Environmental Sciences. One example of fitting a model to observed data using a GA is reported by Mulligan and Brown (1998). They use a GA to estimate parameters to calibrate a water quality model. Some other works related to water quality include using GAs to determine flow routing parameters (Mohan and Loucks, 1995), solving ground water management problems (McKinney and Lin, 1994), sizing distribution networks (Simpson *et al.*, 1994), and calibrating parameters for an activated sludge system (Kim *et al.*, 2002). Aly and Peralta (1999) used GAs to fit parameters of a model to optimize pumping locations and schedules for groundwater treatment. Fayad (2001) together with Peralta used a Pareto GA to sort optimal solutions for managing surface and groundwater supplies. Another example is the use of a GA for classification and prediction of rainy days versus non-rainy days occurrences by Sen and Oztopal (2001). They used the GA to estimate the parameters in a third order Markov model. GAs are also used in Geophysics to determine the type of underground rock layers (Boschetti *et al.*, 1997), soil moisture using streamflows (Dumedah, 2013). Prediction of root zone soil water (Lü, 2011). Minister *et al.* (1995) find that EP is useful for locating the hypocenter of an earthquake, especially when combined with simulated annealing. Cartwright and Harris (1993) suggest that a GA may be a significant advance over other types of optimization models for determining the source of air pollutants given what is known about monitored pollutants, when there are many sources and many receptors. Barth (1992) showed that a GA is faster than simulated annealing and more accurate than a problem specific method for optimizing the design of an oceanographic experiment. Porto *et al.* (1995) found that an EP strategy was more robust than traditional methods for locating an array of sensors in the ocean after they have drifted from their initial deployment location. Charbonneau (1995) gives three examples of uses of a GA in Astrophysics: modelling the rotation curves of galaxies, extracting pulsation periods of Doppler velocities in spectral lines, and optimizing a model of hydrodynamic wind. PSO has also shown great potential for the solution of various optimization problems (Izquierdo *et al.*, 2007b, 2008a, 2012; Montalvo *et al.*, 2007, 2008, 2010a, 2010b; Liao *et al.*, 2007, Jin *et al.*, 2007; Janson *et al.*, 2008). In Whigham (2001) prediction of phitoplancton abundance is modelled with genetic algorithms and in Kim (2012) phitoplancton biomass is approached by combining ANN with manifold network function and an imperialist competitive algorithm. Airame (2003) uses simulated annealing to assist marine reserve design. Yu (2012) address coal production-based air pollution, Afshar (2008) storm water network design with particle swarm optimization and Ni (2012) uses geneticprogramming to forecast water storage. Specific environmental applications include cluster analysis in environmental databases, as in Herrera *et al.* (2009) and Díaz *et al.* (2008); short term scheduling for a biomass supply chain (Izquierdo *et al.*, 2008a); parameter estimation in Hydrology (Gill *et al.*, 2006); stage prediction for rivers (Chau, 2006); calibration of hydrological models (Zhang *et al.*, 2011); groundwater pump design (Mategaonkar et al 2012); run-off modeling in a basin (Kuok *et al.*, 2010); assimilation of root zone soil water predictions (Lü *et al.*, 2011); design and calibration of large and complex urban storm water management models (Muleta *et al.*, 2006); multipurpose reservoir operation (Kumar and Reddy, 2007) (Reddy et al 2007); storm water network design (Afshar, 2008); optimization in water resources management (Baltar and Fontane, 2008); among many others. In Herrera *et al.* (2018), a book devoted to hydroinformatics in water distribution systems, several chapters deal specifically with various of the previous techniques to tackle quality, energy optimization, leak reduction, etc. in those systems.

8. Identifying the proper Data Mining method for a real environmental application

In this section real environmental data mining applications are analyzed from the point of view of the kind of environmental system involved, the type of data mining method used and the environmental question responded. This will elicit a set of typical environmental problems suitable to be analyzed with some data mining methods.

On the basis of Figure 1 and the methods identified in the DMMCM, Table 2 shows a collection of real applications in the environmental field. Colors of the cells letters regards to different global environmental topics like human activity, sustainability or ecology. Shadowed cells regards to hybridation of several DM methods required to deal with the intrinsic complexity of the targeted ES.

<TABLE 2 arround HERE>

Table 2: *Applications of DM methods to ES (DIN-A3 landscape page). Red cells regards to environmental applications related to human activity; green cells refer sustainability; orange cells refer ecology. Shadowed cells refer to hybrid applications where several DM methods are combined.*

It can be seen that all kinds of DM methods can be used in all kinds of ES. The key is to properly match the environmental question that requires answer with the kind of data processing performed by the DM method and to find the method that provides an output relevant for the question.

As an example, regarding vegetation, one of the biotic natural resources identified in Figure 1, we can distinguish basically four kinds of questions:

1. Geographic or temporal distributional issues: when the focus is on discovering which geographical areas support certain plants, understanding the differences between several species of a certain family, or identifying patterns of behaviors along time, clustering methods are appropriated.
2. Characteristics relationships: when the interest is to understand how different characteristics of the context, the plants or plant communities themselves, or occurrences of certain events relate, then associative methods are appropriate. This includes questions such as: how environmental conditions relate to occurrences of species or communities, how certain parameters of a plant (chlorophyll, nutrient load, height of trees associated with environmental factors, fertilizers, etc.) can affect plant development, etc.
3. Recognizing qualitative events: When the interest is to discriminate among a well-known set of events or situations based on some measureable characteristics, then discriminant methods are useful. In this case, the following type of questions can be answered: which are the conditions that produce (or prevent) eutrophication, bleaching, the occurrence of diseases or pathogens, population declines or die offs, overpopulation, etc.
4. Predictive quantitative parameters: When the interest is to predict numerical characteristics of vegetation then predictive methods are appropriate. The methods in this branch can answer questions like quantifying abundance of a certain species, deforestation taxes, ratios of growing, reproduction of a species, heights or volumes of a certain plant, etc.

Similar analyses could be done for other natural resource fields. The interesting thing that Table 2 visualizes is that whenever a numerical variable (like abundance) has to be predicted, predictive methods from the DMMCM map will help independently of the natural resource targeted. This means, for example, that ANN methods are useful for predicting abundance of bromeliads, but also abundance of polar bears or abundance of fish in a river, or macroinvertebrates, or air pollutants in urban areas, or organic matter in water, etc., just because all of them are numerical variables that do not behave under normal distribution and, in these cases, suitable predictive models are ANNs rather than traditional statistical modelling.

So, the idea is that the selection of the DM method is the result of contrasting the target question and the kind of data available with the input and output characteristics of the candidate DM method itself.

This requires some knowledge about the DM methods properties and the environmental system to be analyzed as well. The templates given in section 7 provide the minimum information about DM methods to be able to perform this matching analysis, and to choose the right method in a real application. In the next section, we present a few case studies to illustrate how the DMMCM helps in this matter.

9. Case studies using the DMMCM and the DMMTs to identify the proper DM method for a real environmental application

In this section, the proposed methodology based on the use of the DMMCM to identify the best DM technique to solve a certain environmental problem is illustrated by means of some real-world environmental applications, with the aim to show the usefulness of the proposal to environmental scientists.

Case 1:

Case description and goal: A WWTP manager wants to identify typical scenarios to design action protocols.

Available data: Data about water quality of the influent, the effluent and the middle of the process is available for a certain period based on daily means of pollutant concentrations (organic matter, suspended solids, etc.) and some qualitative observations important in WWTP management, like the color of the bioreactor water or the existence of foam or algae.

Browsing in the DMMCM map: only 2 questions must be answered in step 1 of the proposed methodology:

> **First question to be answered**: Is there a response variable? In this case the interest is to identify states of operation of the plant. So, the answer is *no response variable*
>
> Thus, the DMMCM guides the practitioner to the left side of the first level split in the map. This determines the second question.
>
> **Second question to be answered:** *are we interested in characterizing relationships between variables or individuals*? In our case, the rows of the available data matrix provide measurements on different dates and we are interested in identifying groups of days where the quality of water is similar in all the different areas of the WWTP. So, the answer is: *we are interested in characterize relationships between rows.* And this leads the practitioner to the branch of profiling models.
>
> The recommended methods provided by the DMMCM map at step 1 of the proposed methodology are different clustering families of algorithms (SOM, Statistical Clustering, Clustering based on rules).

Step 2 of the proposed methodology guides the practitioner to the selection of one of these families based on the use of template 7.1. According to the template, as a mixture of relevant decisional variables numerical and qualitative are available, compatibility measures, like Gibert's mixed metrics or Gower similarity coefficient will be suitable. This focuses to methods that permit the use of these kind of distances. As there is no a priori information about the number of scenarios to be discovered, a method that does not require the number of clusters as an input parameter is preferred. Hierarchical clustering is suitable. It is applicable because the data size is not prohibitive. Among all hierarchical methods available, clustering based on rules is preferred since additional *a priori* domain knowledge is available and interpretability of results makes the understanding of the head of the plant easier.

Example: The work Gibert et al. (2010b) shows a real application where the clustering based on rules is used with the Gibert's mixed metrics, and the traffic lights panels are used to describe the clusters prototypically in a symbolic visual way to the head of the plant. Thus, the DMMCM map and the DMMTs permit to identify a concrete DM method that provide an answer to a certain environmental problem.

Case 2:

Case description and goal: A researcher wants to understand how vegetation is influenced by climate through time in a certain area.

Available data: georeferenced data on vegetation and climate changing (precipitation, air temperature) along time

Browsing in the DMMCM map: only 2 questions must be answered in step 1 of the proposed methodology:

> **First question to be answered**: Is there a response variable? In this case we are mainly interested in the relationships between the several variables describing vegetation and weather at a certain timestamp in a certain location. So, the answer is *no response variable*
>
> The DMMCM guides the practitioner to the left hand side of the first level split in the map. This determines the second question.
>
> **Second question to be answered:** *are we interested in characterizing relationships between variables or individuals*. In our case, the rows of the available data matrix provide measurements on different dates and locations and we are interested in analyzing the relationships between status of vegetation and climatic conditions, along space and time, i.e. between precipitation and vegetation levels and so on. So, the answer

is: *we are interested in characterizing relationships between variables.* And this leads the practitioner to the branch of associative models.

The recommended methods provided by the DMMCM map at step 1 of the proposed methodology are: association rules, factorial methods, Bayesian networks, and the like.

Step 2 of the proposed methodology guides the practitioner to the selection of one of these families based on the use of template 7.2. Associative methods. Looking at the brief descriptions of these methods in the template, one can see that association-rule mining is suitable to establish relationships between qualitative variables. Even though original data is numeric, the qualitative relationships provided by association rules are more suitable for communicating patterns to environmental decision makers. Thus, some preprocessing is applied and association-rule is selected.

Example: The work Sli *et al.* (2013) uses environmental change monitoring to detect possible trends related to vegetation and climate data correlated with geographical spatio-temporal data from north-eastern China. Detailed preprocessing is required like interpolation of weather observation data, precipitation and air temperature. Then, the geographical reference system, WGS84, was selected for registration of weather observation data and vegetation data. For the consistency of time periods, weather observation data was temporally truncated into seven years to match seven years of vegetation data. Moreover, weather observation data and vegetation data were resampled with the same resolution of 16 days in time and 500m in space. At the second stage of data conceptualization, weather observation data and vegetation data were separately clustered by the fuzzy c-means clustering (FCM) algorithm. At the third stage of the data mining process, conceptualized data were flexibly organized into transactions for extracting association rules by the algorithm APriori (Agrawal and Srikant, 1994). Massive raw data were transformed into more meaningful understandable knowledge for human decisions. The mined association rules were used to monitor and detect geographical spatio-temporal structural information (relationships) in climate and vegetation data for the detection of environmental changes in climate or vegetation.

CASE 3:

Case description and goal: A WWTP manager needs to verify the effectiveness of new treatment procedures over chemical oxygen demand (COD), Phosphorus and Nitrogen reductions and wants to conduct a what-if analysis.

Available data: Data about water quality of the influent, the effluent and the middle of the process is available for a certain period, including procedure characteristics.

Browsing in the DMMCM map: only 2 questions must be answered in step 1 of the proposed methodology:

First question to be answered: Is there a response variable? In this case we are mainly interested in the relationships between the several variables describing quality of inflow water and actions performed in the treatment process. So, the answer is *no response variable*.

The DMMCM guides the practitioner to the left hand side of the first level split in the map. This determines the second question.

Second question to be answered: *are we interested in characterizing relationships between variables or individuals?* In our case, the rows of the available data matrix provide measurements on different dates and we are interested in analyzing the relationships between effluent COD, effluent total Phosphorus (TP) concentration and effluent total Nitrogen (TN) concentration, and other parameters of the plant and inflow quality. So, the answer is: *we are interested in characterizing relationships between variables.* And this leads the practitioner to the branch of associative models.

The recommended methods provided by the DMMCM map at step 1 of the proposed methodology are: association rules, Multivariate Analysis, Bayesian Networks, etc.

Step 2 of the proposed methodology guides the practitioner to the selection of one of these families based on the use of template 7.2. Associative methods. Looking at the brief descriptions of these methods in the template, one can see that Bayesian networks is suitable for what-it analysis.

Example: The work Li et al. (2013) shows a real application where the input variables of the Bayesian network are: influent COD, influent TP concentration and influent TN concentration. There were also control variables (cycle time of sequencing batch reactor process, anoxic mixing time, aerobic aeration time) and environmental variables (pH, DO and water temperature), and output variables are effluent COD, effluent total phosphorus (TP) concentration and effluent total nitrogen (TN) concentration.

CASE 4:

Case description and goal: A farmer wants to determine the best crop to grow in the next season

Available data: The available data contains the responses to a survey to other farmers in the area regarding household characteristics: farm and household size, land type, tenure and land utilization; crop: production costs for annual crops and perennial crops including fertilizers, materials, machinery and labor use; output: product sold and income for annual or perennial crops; income for other sources and capital availability, environmental problems, past use of land, competition of annual crops, farmers' attitude, use and management of irrigation water, description of farmers' crop choice decision making. Each response is properly georeferenced with the location of the corresponding soil.

Browsing in the DMMCM map: only 2 questions must be answered in step 1 of the proposed methodology:

> **First question to be answered**: *Is there a response variable?* In this case we are mainly interested in explaining the decision of the final crop choice in terms of the other variables. So, the answer is *yes, multiple response variable is the type of crop*.
>
> The DMMCM guides the practitioner to the right hand side of the first level split in the map. This determines the second question.
>
> **Second question to be answered:** *are the response variables numerical or qualitative?* In our case, the response variable is type of crop chosen. So, the answer is: *we are interested in getting values for a qualitative response variable.* And this leads the practitioner to the branch of discriminant models.
>
> The recommended methods provided by the DMMCM map at step 1 of the proposed methodology are: rule-based classifiers, decision trees, support vector machines, etc.

Step 2 of the proposed methodology guides the practitioner to the selection of one of these families based on the use of template 7.4. Discriminant methods. Looking at the brief descriptions of these methods in the template, one can see that decision trees are good at providing results that can explain the decision, which is an interesting characteristic for the farmer.

Example: The work Ekasingh et al (2005) shows a real application where a survey on crop choices is conducted in Thailand and a decision tree is used to learn a predictive model that is included in a decision support system.

CASE 5:

Case description and goal: A city government wants to predict the air pollution levels for the next day, based on as set of relevant atmospheric parameters, to activate or not traffic restrictions for sustainability and public health purposes.

Available data: The available data contain the concentrations of SO_2, NO_2, O_3 and particulate matters of diameters up to 10 µm (PM10), the daily average value of the temperature, wind speed and direction, humidity, pressure and insolation, maximum and minimum values of temperature and pressure, maximum level of pollution for the target day and previous day.

Browsing in the DMMCM map: only 2 questions must be answered in step 1 of the proposed methodology:

> **First question to be answered**: *Is there a response variable?* In this case we are mainly interested in getting values for the concentration of air pollutants. So, the answer is *yes, multiple response variables* (SO_2, NO_2, O_3, PM10).
>
> The DMMCM guides the practitioner to the right hand side of the first level split in the map. This determines the second question.

Second question to be answered: *are the response variables numerical or qualitative?* In our case, the response variable are concentrations of pollutants in atmosphere. So, the answer is: *we are interested in getting values for a set of numerical response variables.* And this leads the practitioner to the branch of predictive models.

The recommended methods provided by the DMMCM map at step 1 of the proposed methodology are: statistical modelling (regression, time series…), ANNs, support vector regression (SVR), evolutionary computation, etc.

Step 2 of the proposed methodology guides the practitioner to the selection of one of these families based on the use of template 7.5. Predictive methods. Looking at the brief descriptions of these methods in the template, one can see that statistical methods are not suitable since most of the assumptions required by the methods fail in available data. Assuming the need of a predictive method able to learn about non-linearities and complex relationships, random forests on regression trees, SVR and ANNs are suitable. Whenever the interpretability of final model becomes important SVR and ANNs should be less interesting.

Example: The work Siwek and Osowski (2016) shows a real application where the available data contain the concentration of SO_2, NO_2, O_3 and particulate matters of diameters up to 10 µm (PM10), the daily average value of the temperature, wind speed and direction, humidity, pressure and insolation and same information for the past day, including the average, maximum and minimum values of temperature and pressure, the average and maximum pollution corresponding to the previous day, the linear trend of hourly pollution, the linear prediction of the pollution made on the basis of this trend, the season of the year (winter, spring, summer and autumn) and the type of day (weekdays and weekends). Selected hourly values of pollution of the previous day were also available.

They use a previous step for *feature selection* based on a genetic algorithm and alternatively on stepwise techniques. Then, they use two *predictive models:* In the first one, random forest (RF) of decision trees. In the second approach, ANNs: the MLP, the RBF network and the SVR with Gaussian kernel, which is a version of SVM to learn continuous functions instead of classes. The results of the air pollution predictions were used for monitoring the air quality to satisfy the European air quality directive EC/2008/50, which defines restrictions for yearly and 24h average PM10 concentrations and to diminish dangerous concentration levels, emission abatement actions have to be planned at least one day in advance. Moreover, according to EU directives, public information on the air quality status and on the predictable trend for the next days should also be provided.

Thus, the DMMCM map and the DMMTs can help identify a suitable DM method that provides answers to a certain environmental problem.

10 Conclusions

Environmental processes exhibit several intrinsic complexities that make data analysis difficult, and in these cases classical data analysis methods often do not perform well. DS is a promising approach to analyze environmental data. DS defines a new paradigm where, apart from the DM step itself, raw data preprocessing and outcome post-processing are included in the methodology. Additionally, prior expert knowledge may be used to boost the discovery of new useful results. In Gibert (2016b) a survey on preprocessing is provided. In this paper, DM is fully considered. Post-processing, which is method-specific, tries to bridge the gap between DM results and effective knowledge production. Post-processing remains an area for future research.

The main contributions of this paper include the following. First, a conceptualization of DM for environmental systems is approached. This conceptualization is intended to provide a general overview of the DM techniques regarding its use in environmental problems.

A second contribution focusses on the difficulty of choosing the right DM technique to perform genuine DS over real-world datasets. We identify the main decision elements that a data scientist has to face to make that choice, and organize them in a simple set of questions that can orient the decision from the environmental scientist point of view. This does not necessarily mean that DS must be performed autonomously by the own environmental scientist, but can enormously help him or her to find the proper expert to analyze his or her data. The guidelines are formalized under an easy-to-read conceptual map, and provides a decision support tool for the non-expert user, with a global overview of possible useful methods. This tool intends to contribute to correctly exploit data by using DM while considering the problem goals and the structure of the available data, while reducing the use of inappropriate tools in real applications. Our overview does not attempt to be exhaustive. Also, the proposal is currently limited to methods analyzing classical data matrices. When sensor data or images are the original source of information, feature extraction or signal processing operations are required to obtain descriptors of the images or signals that can be used as ordinary variables in the data matrix. Current work is in progress to explore the extension of the DMMCM to wider scenarios including other types of input data.

A third contribution includes generic synthetic templates describing the main families of DM methods trying to provide the relevant information from the point of view of the environmental scientist. Regarding this third contribution, a structured and synthetic description of the main families of methods presented in the conceptual map is provided. A major effort has been done to present the methods from the point of view of the kind of problem to be solved, and not from the classical methodological point of view, certainly interesting for data scientists and researchers, but less interesting from a practical real use of the methods point of view. The structure of the template has been carefully designed to be useful for environmental scientists who want to exploit their available data in front of a specific problem. The templates try to offer decision elements on: how to use methods correctly, which is the structure of the data that every method can properly analyze and, most importantly, how to validate and interpret the results. A specific section in the template, named *Principles*, tries to offer a high level idea of the technical principles guiding the various algorithms. These principles need not be used, but can help provide a coherent context for interested readers. Technical details on specific DM techniques are out of the scope of this paper, although a rich set of references for in-depth coverage is provided. Again, we do not attempt to be exhaustive, but to provide a global insight for the non-expert user to better select what to do in real applications.

A powerfull contribution of the paper is also the proposal of a methodology to help a non-expert data scientist or an environmental practitioner to choose the most suitable DM method for getting appropriate answers to a certain environmental problem. The proposal is based on a two-step use of the DMMCM map and DMMTs presented along the paper.

On top of this, illustration, through several case studies about the use of the proposed methodology to choose the DM technique, will certainly ease the understanding of the paper, and must be considered a valuable contribution.

This work responds to one of the main research challenge announced in the paper Gibert *et al.* (2018). The DMMCM and the DMMTs also provide background knowledge to input in an *intelligent data mining technique recommender (InDaMiTe-R),* contributing to the construction of integral DS systems, as stated in Gibert *et al*. (2012).

11 Challenges for Environmental Data Mining

Finally, as a last contribution of the paper, the enormous effort in building technical data mining templates, looking at the kind of things done or not in pre and post processing, permitted to identify the hot issues and challenging aspects in and of the interdisciplinary field of environmental DM in coming years. Achievement of the following aims would increase utility and applicability of DM methods. We summarize this in terms of specific guidelines for users and challenges for researchers.

Guidelines for users:

- Get acquainted with available methods and technical assumptions of those methods, even before data collection, if possible.
- Try to clarify the relevant question to be answered by the DS process to orient own DM. Do not lose perspective of final usefulness.
- Meta-data must be collected and clearly understood for a proper analysis and interpretation.
- Devote the required time to data pre-processing, and perform it carefully enough to guarantee the quality of the results.
- Choose the right DM method according to your goals and your available data. The conceptual map and the technical templates presented in this paper are our proposal to fulfill this step.
- Try resampling to check stability of results and to process big data sets when a particular limited method is required. Use the classical statistical sampling principles, traditionally used to get samples from real populations, to sample over a big dataset.
- Be sure to validate your results with the proper validation processes, which will depend on the DM method used. Take care of spurious results when classical statistical testing is performed over massive data sets.
- Prioritize understandability of results. Devote all required time to carefully analyze the results. Select the relevant information from the output and try to present it in an understandable way, useful for decision support.

Challenges for researchers:

- Provide support to metadata management into the DM system.
- Elaborate protocols to facilitate data sharing and data reuse.
- Need for developing methods for data fusion, dealing with situations in which data comes from different sources, with different natures, scales, granularities and formats.
- Need for developing powerful hybrid DM methods, combining technical principles for improving Data Science processess performance in highly complex domains.
- Need for improving DM techniques for on-line and heterogeneous databases.
- More research is still required to better understand the complexity of post-processing tasks in the context of Data Science processes.
- Need for formulation of tools for explicit representation and handling of discovered knowledge for greater understandability. Development of tools to bridge the gap between modelling and effective decision-making (Gibert *et al.*, 2018) can be enormously useful.
- Build MetaData Models to describe DM methods in such a way that main hypothesis and characteristics of required inputs and outputs are easy to match with real applications. Authors are working in this particular as a continuation of the proposed methodology
- Provide technical templates for new families of methods, as presented in this work.
- Develop special synthetic templates for specific methods, with particular guidelines on the use of the parameters, and standardize those templates to be used for user manuals in DS packages.
- Great need for developing DM methodologies particularly oriented to extract knowledge and models from temporal/spatial environmental phenomena, probably requiring data fusion and hybrid approaches.
- As many DM packages provide GUIs to design the workflow for a complete Data Science process, collect Data Science experiences in a workflow library and develop DM strategies to mine the workflows themselves to improve preprocessing, DM and post-processing recommenders under an evidence-based approach.
- Involvement of end-user (domain expert) criteria in algorithm design and result interpretation.
- Development of standard procedures (benchmarks) for experimental testing and validation of DM tools.
- Need to move towards the development of integral Data Science systems, with the intelligence to pursue integral approaches, covering the whole Data Science process, from problem formulation to interpretation of results, including data fusion and harmonization, metadata management, data cleaning and other preprocessing techniques, taking into account prior expert knowledge, DM method recommendation, and post-processing, and involving the expert in problem formulation, preferences on kinds of results, and interpretation of final results.

Various software packages have been analyzed from the perspective of the support they provide to the whole Data Science process. Although there are some available tools to develop Data Science in a semiautomatic way analysis, and some of them, such as KLASS, evolve towards providing intelligent support to post-processing, we believe that current packages are still far from being complete intelligent systems that could support integral Data Science, as stated above. Main limitations in current systems are the lack of support in the DM problem to be solved, and they rarely include tools to recommend the appropriate DM technique in an automatic way. The complete Data Science process is normally designed from scratch for every application and only unconnected catalogues for data selection, pre and post-processing and DM tools are provided. We strongly believe that future efforts should be given in this direction.

As a final conclusion, we could say that the Data Science approach is not always linked to massive data sets, but sometimes it is suitable to analyze non-massive datasets of high complexity where classical pure data analysis techniques do not perform well. In this cases, the integral and global approach of Data Science provides a good framework to find complex models through the combination of techniques in a global hybrid Data Science methodology.

Acknowledgments

The work presented here emanated from discussions in the DMTES workshop series, which have been held in the biennial iEMSs conference since 2006. Authors want to thank all participants in the workshop for the interesting discussions and raising questions that lead to the research presented in this paper.

References

Afshar, M.H., 2008. Rebirthing particle swarm optimization algorithm: application to storm water network design. *Canadian Journal of Civil Engineering,* **35**(10).

Agrawal, R., Srikant R 1994. Fast algorithms for mining association rules in large databases, in Bocca, Jorge B.; Jarke, Matthias; and Zaniolo, Carlo; editors, Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, September 1994, pages 487-499.

Aguilera, P.A., *et al.* 2011. Bayesian networks in environmental modelling. Environmental Modelling & Software 26.12 (2011): 1376-1388.

Airamé, S., Dugan, J. E., Lafferty, K. D., Leslie, H., McArdle, D. A., & Warner, R. R., 2003. Applying ecological criteria to marine reserve design: a case study from the California Channel Islands. *Ecological applications*, *13*(sp1), 170-184.

Alferes, J., & Vanrolleghem, P. A. (2016). Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. AI Communications, 29(6), 701-709.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. *ACM Sigmod record* 28(2):49-60).

Alferes, J., & Vanrolleghem, P. A., 2016. Efficient automated quality assessment: Dealing with faulty on-line water quality sensors. *AI Communications*, *29*(6), 701-709.

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.

Aly, A.H., Peralta, R.C., 1999. Comparison of a genetic algorithm and mathematical programming to the design of groundwater cleanup systems, Water Resources Research, 35(8), pp. 2415-2425.

Annoni, P., Brüggemann, R., 2008. The dualistic approach of FCA: A further insight into Ontario Lake sediments. Chemosphere, 70 (11), pp. 2025-2031.

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*.

Athanasiadis, I., Mitkas, P., 2004. Supporting the decision-making process in environmental monitoring systems with knowledge discovery techniques. In: *Proceedings of the Knowledge-Based Services for the Public Services Symposium, Workshop III: Knowledge Discovery for Environmental Management*. KDnet, pp. 1-12.

Athanasiadis, I., Kaburlasos, V., Mitkas, P., Petridis, V., 2003. Applying machine learning techniques on air quality for real-time decision support. In: *Information technologies in Environmental Engineering*.

Avesani P., Ricci, F Perini A., 2000. Interactive case-based planning for forest FIRE fighting. *Applied. Intelligence*, **13**, 41-58.

Babovic, V., 2005. Data mining in hydrology. *Hydrological Processes*, 19:1511-1515.

Baeza-Yates, R., 2017. Big-data or Small Data? the Correct Answer Is Both. Inside BIG-data Editorial. July 13th 2017.

Baltar, A.M., Fontane, D.G., 2008. Use of Multi-Objective Particle Swarm Optimization in Water Resources Management, *J. Water Resour. Plng. and Mgmt.*, **134**(3), 257-268.

Barth, H., 1992. Oceanographic Experiment Design II: Genetic Algorithms, Journal of Oceanic and Atmospheric Technology, 9, 1992, pp. 434-443.

Bartoletti, N., Casagli, F., Marsili-Libelli, S., Nardi, A., & Palandri, L., 2018. Data-driven rainfall/runoff modelling based on a neuro-fuzzy inference system. *Environmental Modelling & Software* 106:35-47

Basak, D., Pal, S., & Patranabis, D. C., 2007. Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203-224.

Batet, M., Valls, A., Gibert, K., 2010. Performance of Ontology-based Semantic Similarities in Clustering. Artificial Intelligence and Soft Computing. LNAI 6113: 281-288, Springer.

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

Bello, M. G., 1992. Enhanced training algorithms, and integrated training architecture selection for multilayer perceptron networks, IEEE Trans. on Neural Networks 1992; 3: 864-875.

Blanco-M, A., Gibert, K., Marti-Puig, P., Cusidó, J., Solé-Casals, J., 2018. Identifying Health Status of Wind Turbines by Using Self Organizing Maps and Interpretation-Oriented Post-Processing Tools. *Energies*, *11*(4), 1-21.

Boschetti, F., Dentith, M.C., List, R., 1997. Inversion of potential field data by genetic algorithms. Geophysical Prospecting.45, pp 461-478.

Boser, B.E., Guyon, I.M, Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: COLT ’92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York, NY, USA: ACM Press, pp. 144–152.

Brandão, S. N., Silva, W. N., Silva, L. A., Fagundes, V., de Mello, C. E. R., Zimbrão, G., & de Souza, J. M., 2009. Analysis and visualization of the geographical distribution of Atlantic forest bromeliads species. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on* (pp. 375-380). IEEE.

Breiman, L., 2017. *Classification and regression trees*. Routledge.

Brentan, B., Meirelles, G., Luvizotto Jr, E., Izquierdo, J., 2018. Joint Operation of Pressure-Reducing Valves and Pumps for Improving the Efficiency of Water Distribution Systems. J. Water Resour. Plann. Manage, 144(9): 04018055.

Brentan, B.M., Campbell, E., Goulart, Th., Manzi, D., Meirelles, G., Herrera, M., Izquierdo, J., Luvizotto Jr, E., 2018a. Social Network Community Detection and Hybrid Optimization for Dividing Water Supply into District Metered Areas. Journal of Water Resources Planning and Management, 144(5), 04018020 (1-10).

Brentan, B.M., Meirelles, G., Luvizotto Jr, E., Izquierdo, J., 2018b. Hybrid SOM+k-Means Clustering to Improve Planning, Operation and Management in Water Distribution Systems. Environmental Modelling & Software, 106, 77-88.

Brentan, B.M., Meirelles, G., Herrera, M., Luvizotto Jr, E., Izquierdo, J., 2017b. Correlation analysis of water demand and predictive variables for short-term forecasting models. Mathematical Problems in Engineering, 2017 Article ID 6343625, 10 pages.

Brentan, B.M., Campbell, E., Meirelles, G., Luvizotto Jr, E., Izquierdo, J., 2017a. Social Network Community Detection for DMA Creation: Criteria Analysis through Multilevel Optimization. Mathematical Problems in Engineering, Vol. 2017, Article ID 9053238.

Brugnach, M., Pahl-Wostl, C., Lindenschmidt, K. E., Janssen, J. A. E. B., Filatova, T., Mouton, A., ... & Gaber, N. (2008). Chapter four complexity and uncertainty: rethinking the modelling activity. Developments in Integrated Environmental Assessment, 3, 49-68.

Brun, M., C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. Dougherty, 2007. Model-based evaluation of clustering validation measures. Pattern Recognition 40(3):807–824.

Burns, E., 2017. Lack of skills remains one of the biggest data science challenges. In: Search Business Analytics Jan 11th.

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

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics – Simulation and Computation 3 (1), 1–27.

Camargo, S., Robertson, A., Gaffney, S. and Smyth, P., 2004. Cluster analysis of western north pacific tropical cyclone tracks. In: *Proceedings of the 26th Conference on Hurricanes and tropical Meteorology*. pp. 250-251.

Campbell, E., Izquierdo, J., Montalvo, I., Ilaya-Ayza, A., Pérez-García, R., Tavera, M., 2016. A Flexible Methodology to Sectorize Water Supply Networks Based on Social Network Theory Concepts and on Multi-objective Optimization. Hydroinformatics, 18(1), 62-76.

Carpitella, S., Brenta, B.M., Montalvo, I., Izquierdo, J., Certa, A., 2018. Multi-objective and multi-criteria analysis for optimal pump scheduling in water systems. Proc. HIC2018, 13th International Conference on Hydroinformatics, Palermo, Italy.

Cartwright, H.M., Harris, S.P., 1993. Analysis of the Distribution of Airborne Pollution using Genetic Algorithms, Atmospheric Environment, Part A, 27A, pp. 1783-1797.

Carvalho Pereira, G., Coutinho, R., & Ebecken, N. F. F. (2008). Data mining for environmental analysis and diagnostic: a case study of upwelling ecosystem of Arraial do Cabo. *Brazilian Journal of Oceanography*, *56*(1), 1-12.

Castelletti, Andrea, and Rodolfo Soncini-Sessa. "Bayesian Networks and participatory modelling in water resource management." *Environmental Modelling & Software* 22.8 (2007): 1075-1088.

Cazarez-Candia, O., & Vásquez-Cruz, M. A. (2005). Prediction of pressure, temperature, and velocity distribution of two-phase flow in oil wells. *Journal of Petroleum Science and Engineering*, *46*(3), 195-208.

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

Charbonneau, P., 1995. Genetic Algorithms in Astronomy and Astrophysics, The Astrophysical Journal Supplement Series, 101, 309-334.

Charest, M., Delisle, S., 2006. Ontology-guided intelligent data mining assistance: Combining declarative and procedural knowledge. Artificial Intelligence and Soft Computing 2006: 9-14

Chau, K.W., 2006. Particle Swarm Optimization Training Algorithm for ANNs in Stage Prediction of Shing Mun River. *Journal of Hydrology*, **329(3-4**), pp. 363-367.

Chieppa, A., Gibert, K., Gómez-Sebastià, I., Sànchez-Marrè, M., 2008. Improving pseudo-bagging techniques. Artificial Intelligence Research and Development. Frontiers in Artificial Intelligence and Applications 184:161—169, IOSPress

Christensen, R., 2002. Plane Answers to Complex Questions: The Theory of Linear Models (Third ed.). NY: Springer

Christianini, N., Shawe-Taylor, J. 2000. An Introduction to Support Vector Machines. Cambridge University Press

Clark, P., Niblett, T., 1989. The CN2 induction algorithm. Machine Learning 3, 261-283.

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.

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.

Conti, D., & Gibert, K. (2014). Discovering comprehensible hydrogeological profiles in the Margarita Island's acquifers including post-processing in a data mining process. In *Proceedings of the iEMSs 2014* (pp. G2-7).

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)*. pp. 603-607.

Cukier, 2010. Data, Data everywhere. In Special report on managing information. The Economist, February 27th 2010.

de Araújo, J. C., Döll, P., Güntner, A., Krol, M., Abreu, C. B. R., Hauschild, M., & Mendiondo, E. M. (2004). Water scarcity under scenarios for global climate change and regional development in semiarid Northeastern Brazil. *Water International*, *29*(2), 209-220

Díaz, J.L., Herrera, M., Izquierdo, J., Montalvo, I., Pérez-García, R., 2008. A Particle Swarm Optimization derivative applied to cluster analysis. *Proceedings of the iEMSs 2008 (International Congress on Environmental Modelling and Software - 4th Biennial Meeting)*, Barcelona, 2008.

Di Nardo, A., Giudicianni, C., Greco, R., Herrera, M., Santonastaso, G.F., 2018. Applications of Graph Spectral Techniques to Water Distribution Network Management, Water, 10, 45.

Dillon, W., Goldstein, M., 1984. *Multivariate Analysis*. Wiley, 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.

Domańska, D., & Wojtylak, M., 2012. Application of fuzzy time series models for forecasting pollution concentrations. *Expert Systems with Applications*, *39*(9), 7673-7679.

Domingos, P., 1996. Unifying Instance-Based and Rule-Based Induction. *Machine Learning*, 24: 141−168.

Dorigo, M., Maniezzo, V., Colorni, A., 1996. The ant system: Optimization by a colony of cooperating agents, IEEE Transactions on Systems, Man, and Cybernetics – Part B, 26(2): 29-41.

Du, P., Liu, P., Zhang, H., Zhang, H., 2007. Multi-objective processing of ASTER image for urban environmental analysis. International Geoscience and Remote Sensing Symposium (IGARSS), art. no. 4422886, pp. 675-678, (2007).

Dumedah, G., Coulibaly, P. Evolutionary assimilation of streamflow in distributed hydrologic modeling using in-situ soil moisture data. *Advances in Water Resources*, 53:231-241.

Dürrenmatt, D.J. 2012. Data-driven modelling approaches to support wastewater treatment plant operation. Environmental Modelling and Software, 30: 47-56.

Dzeroski, S., Drumm, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (holothurian 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. and 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*.

Elarab, M., Ticlavilca, A. M., Torres-Rua, A. F., Maslova, I., McKee, M. (2015). Estimating cholorophyll with thermal and broadband multispectral high resolution imagery from an unmanned aerial system using relevance vector machines for precision agriculture, International Journal of Applied Earth Observation and Geoinformation, 43, 32-42, https://doi.org/10.1016/j.jag.2015.03.017.

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.

EPA 305-R-07-001. Guide for Addressing Environmental Problems: Using an Integrated Strategic Approach. EPA 305-R-07-001, March 2007.

Estel, S., Kuemmerle, T., Levers, C., Baumann, M., & Hostert, P., 2016. Mapping cropland-use intensity across Europe using MODIS NDVI time series. *Environmental Research Letters*, *11*(2), 024015.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Fayad, H., 2001. Application of neural networks and genetic algorithms for solving conjunctive water use problems, Ph.D. Dissertation, Utah State University, 152 pp.

Fernandes, J.A., Lozano, J.A., Inza, I., Irigoen, X., Pérez, A. and Rodríguez, J.D., 2013. Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting. *Environmental Modelling and Software*, 40: 245-254.

Ferrarini, A., Tomaselli, M., 2010. A new approach to the analysis of adjacencies: Potentials for landscape insights. *Ecological Modelling*, 221(16):1889-1896.

Forio, M. A. E., Goethals, P., Lock, K., Asio, V., Bande, M., & Thas, O., 2018. Model-based analysis of the relationship between macroinvertebrate traits and environmental river conditions. *Environmental Modelling and Software*.

Friedel, M.J. 2012. Data-driven modelling of surface temperature anomaly and solar activity trends. Environmental Modelling and Software, 37: 217-232.

Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J. and Mouton, A.M. 2013. Habitat prediction and knowledge extraction for spawning European grayling (Thymallus thymallus L.) using a broad range of species distribution models. Environmental Modelling and Software, 47: 1-6.

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, GC, 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*. pp. 789-794.

Gibert, K., 2014. Automatic generation of classes-interpretation as a bridge between clustering and decision-making. *International Journal of Multicriteria Decision Making*, *4*(2), 154-182.

Gibert, K., Conti, D., & Vrecko, D., 2012. Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. *Environmental Engineering and Management Journal*, *11*(5), 931-944.

Gibert, K., & Cortés, U. (1998). Clustering based on rules and knowledge discovery in ill-structured domains. Computación y Sistemas Vol. 1 No. 4 pp. 213-227, IPPN, México

Gibert K, A. García-Rudolph, G. Rodríguez-Silva, 2008c. The role of KDD Support-Interpretation tools in the conceptualization of medical profiles: An application to neuro-rehabilitation. Acta Informatica Medica 16(4): 178-182

Gibert, K., J. Horsburgh, I. Athanasiadis, G. Holmes (2018). "Environmental Data Science." Environmental Modelling & Software 106: 4-12. (DOI:10.1016/j.envsoft.2018.04.005)

Gibert, K., Nonell, R., 2008. Pre and post-processing in KLASS. Proc. of the iEMSs 4th biennial meeting: International Congress of Environmental Modeling and Software (DM-TES'08 Workshop) iEMSs 2008, vol. III 1965-1966.

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

Gibert K., Rodríguez-Silva, G., Rodríguez-Roda, I., 2010b. Knowledge Discovery with Clustering based on rules by States: A water treatment application. Environmental Modelling and Software 25: 712-723

Gibert, K., Sànchez-Marrè, M., 2012. A picture on Environmental Data Mining Real Applications. What is done and how? In R. Seppelt *et al.* eds. Proceedings of iEMSs'2012: 1612-1619, Leipzig. Germany. ISBN 978-88-9035-742-8

Gibert, K., Sànchez-Marrè, 2011. Outcomes from the iEMSs Data Mining in the Environmental Sciences Workshop Series. Environmental Modelling and Software, 26:983-985

Gibert, K., Sànchez-Marrè, M., Codina, V., 2010. Choosing the right data mining technique: classification of methods and intelligent recommenders. Proc. of the iEMSs Firth Biennial Meeting: Int'l Congress on Environmental Modelling and Software vol. I 2448-2453 (Swayne, D. *et al.* eds.). Ottawa University, CA..

Gibert, K. Sànchez-Marrè, M., Izquierdo, J., 2016. A Survey on Pre-processing Techniques in the Context of Environmental Data Mining. Artificial Intelligence in Communications v29 (in press), IOSPress. Amsterdam, NL

Gibert, K., Sànchez–Marrè, M., & Izquierdo, J., 2016b. Preface to the special issue on Environmental Data Science: Air quality and water cycle applications. *AI Communications*, *29*(6), 1-4.

Gibert, K., Sànchez-Marrè, M. and 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.

Gibert, K., Sierra, C., 2014. Preface, Artificial Intelligence Communications, 28(1):1-3 (DOI 10.3233/AIC-140642), IOSPress, Amsterdamm, NL http://content.iospress.com/articles/ai-communications/aic642

Gibert, K. and 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., Spate, J., Sànchez-Marrè, M., Athanasiadis, I., Comas, J., 2008b. Data Mining for Environmental Systems. In Environmental Modeling, Software and Decision Support. State of the art and New Perspectives IDEA Series (Jackeman, A. J., Voinov, A., Rizzoli, A., and Chen, S. eds.) v 3: 205—228, Elsevier

Gibert, K., Valls, A., & Batet, M., 2014. Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. *Knowledge and information systems*, *40*(3), 559-593.

Gill, M.K., Kaheil, Y.H., Khalil, A., McKee, M., Bastidas L., 2006. Multi-objective particle swarm optimization for parameter estimation in hydrology. *Water resources research*, **42**, W07417, 14pp.

Goldberg, D., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley

Gómez-Losada, Á., Pires, J. C. M., & Pino-Mejías, R. (2018). Modelling background air pollution exposure in urban environments: Implications for epidemiological research. *Environmental Modelling & Software*.

Ghosh, S., Graf, U., Ecker, K., Wildi, O., Küchler, H., Feldmeyer-Christe, E. and Küchler, M., 2014. Dimension reduction and data sharpening of high-dimensional vegetation data: An application to Swiss mire monitoring. Ecological Indicators, 36(0), pp. 242-253.

Gower, J.C., 1971. A General coefficient of similarity and some of its properties. Biometrics 27, 857--874

Guariso, G., Werthner, H., 1989. Environmental Decision Support Systems. John Wiley & Sons, New York.

Gudimov, A. Eavan O'Connor, Maria Dittrich, Hamdi Jarjanazi, Michelle E. Palmer, Eleanor Stainsby, Jennifer G. Winter, Joelle D. Young, and George B. Arhonditsis, Continuous Bayesian Network for Studying the Causal Links between Phosphorus Loading and Plankton Patterns in Lake Simcoe, Ontario, Canada, Environmental Science & Technology 2012 46 (13), 7283-7292, DOI: 10.1021/es300983r

Guha, S., Rastogi, R., Shim, K., 2001. "CURE: An Efficient Clustering Algorithm for Large Databases". *Information Systems* **26** (1): 35–58. doi:10.1016/S0306-4379(01)00008-4.

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.

Hamilton, J., 1994. Time Series Analysis, Princeton: Princeton Univ. Press

Hamilton, S.H., Pollino, C.A., Walker, K.F., 2017. Regionalisation of freshwater fish assemblages in the Murray-Darling Basin, Australia. Marine and Freshwater Research 68(4), 629-649.

Hammond, M., 2007. The Fact Gap: The Disconnect Between Data and Decisions, *Business Objects*, 2007.

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

Han, J., Pei, J., Yiwen, Y., Mao, R., 2004. Mining frequent patterns without candidate generation, Data Mining and Knowledge Discovery 8:53-87.

Han, J., M. Kamber, J. Pei,Data Mining: Concepts and Techniques, Kaufman, 2011.

Hanley, J. A., & McNeil, B. J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

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

Hastings, J.D., Branting, L.K., Lockwood, J.A., 2002. CARMA: A Case-based Rangeland Management Adviser. *AI Magazine*.

Haupt S.E., Haupt, R.L., 2002. Genetic algorithms and their applications in Environmental Sciences. 3rd Conference on Artificial Intelligence Applications to the Environmental Science, 23, 49-62.

Herrera, M., Ramos-Martínez, E., Izquierdo, J., Pérez-García, R., 2015. Graph constrained label propagation on water supply networks. AI Communications, 28, 47–53.

Herrera, M., Ferreira, A. A., Coley, D. A., & de Aquino, R. R., 2016). SAX-quantile based multiresolution approach for finding heatwave events in summer temperature time series. AI Communications, 29(6), 725-732.

Herrera, M., Izquierdo, J., Montalvo, I., García-Armengol, J., Roig, J.V., 2009. Identification of surgical practice patterns using evolutionary cluster analysis. *Mathematical and Computer Modelling* **50**, pp. 705-712.

Herrera, M., Meniconi, S., Alvisi, S., Izquierdo, J., (Eds.), 2018. Advanced Hydroinformatic Techniques for the Simulation and Analysis of Water Supply and Distribution Systems. MDPI publishers, Pages: VIII-370.

Hochreiter, S., & Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Holland. J., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan, 1975 [MIT Press, 1992].

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.

Holmes, G., Fletcher, D., & Reutemann, P., 2012. An application of data mining to fruit and vegetable sample identification using Gas Chromatography-Mass Spectrometry. iEMSs.

Huang, Y., Gedeon, T. D., & Wong, P. M., 2001. An integrated neural-fuzzy-genetic-algorithm using hyper-surface membership functions to predict permeability in petroleum reservoirs. *Engineering Applications of Artificial Intelligence*, *14*(1), 15-21.

Izquierdo, J., Pérez, R., López, P. A., Iglesias, P. L., 2006. Neural Identification of Fuzzy Anomalies in Pressurized Water Systems, Summit on Environmental Modelling and Software, 3rd Biennial meeting of the International Environmental Modelling and Software Society, Proceedings, Burlington (VT), USA.

Izquierdo, J., Montalvo, I., Pérez, R. & Fuertes, V.S., 2007b Design optimization of wastewater collection networks by PSO. Computer & Mathematics with Applications, 56(3), pp.777–784.

Izquierdo, J., Minciardi, R., Montalvo, I., Robba, M., Tavera, M., 2008a. Particle Swarm Optimization for the biomass supply chain strategic planning. Proceedings of 4th Biennial Meeting, iEMSs 2008: International Congress on Environmental Modelling and Software, pp. 1272-1280, Barcelona, Spain

Izquierdo, J., Montalvo, I., Pérez-García, R., Herrera, M. 2012. Particle swarm optimization. In: Heuristic Optimization in Hydrology, Hydraulics and WR Management. WIT Press - Ashurst Lodge, Ashurst, Southampton, S040 7AA, UK.

Izquierdo, J., Montalvo, I., Campbell, E., Pérez-García, R., 2016a. A hybrid, auto-adaptive, and rule-based multi-agent approach using evolutionary algorithms for improved searching. Engineering Optimization, 48(8), 1365-1377.

Izquierdo, J., Campbell, E., Montalvo, I., Pérez-García, R., 2016b. Injecting problem-dependent knowledge to improve evolutionary optimization search ability. Journal of Computational and Applied Mathematics, 291, 281-292.

Jain, S.K. 2012. Modeling river stage-discharge-sediment rating relation using support vector regression. Hydrology Research, 43(6), 851-861.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data Clustering: A Review. ACM computing surveys V. 31(3)264-323

James, F., McCulloch, C., 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's Box? Annual review of ecology and systematics 21:129-166

Jang, J.-S. R., 1992. Neuro-Fuzzy Modeling: Architectures, Analyses, and Applications. Ph.D. Dissertation, EECS Dep, University of California at Berkeley

Janson, S., Merkle, D., Middendorf, M., 2008. Molecular docking with multi-objective Particle Swarm Optimization. *Applied Soft Computing* **8**, pp. 666–675.

Jin, Y.X., Cheng, H.Z., Yan, J.I., Zhang, L., 2007. New discrete method for particle swarm optimization and its application in transmission network expansion planning. *Electric Power Systems Research*, **77(3-4)**, pp. 227-233.

Kalapanidas, E., Avouris, N., 2001. Short-term air quality prediction using a case-based classifier. *Environmental Modelling and Software.* **16**, 263–272.

Kamarianakis, Y., Feidas, H., Kokolatos, G., Chrysoulakis, N. and Karatzias, V. 2008. Evaluating remotely sensed rainfall estimates using nonlinear mixed models and geographically weighted regression. Environmental Modelling and Software, 23(12): 1438-1447.

Kanevski, Mikhail, *et al.*, 2002. "Support vector machines for classification and mapping of reservoir data." *Soft Computing for Reservoir Characterization and Modeling*. Physica, Heidelberg. 531-558.

Kaster, D.S., Medeiros, C.B., Rocha H.V., 2005. Supporting modeling and problem solving from precedent experiences: the role of workflows and case-based reasoning. *Environmental Modelling and Software* **20**, 689-704.

Kennedy, J., Eberhart, R.C., 1995. Particle swarm optimization. Proc. of the IEEE International Conf. on Neural Networks, Piscataway, NJ, pp. 1942-1948.

Khaledian, Yones, *et al.* "Assessment and monitoring of soil degradation during land use change using multivariate analysis." *Land Degradation & Development* 28.1 (2017): 128-141.

Kim, S., Lee, H., Kim, J., Kim, C., Ko, J., Woo, H., Kim, S., 2002. Genetic algorithms for the application of Activated Sludge Model No. 1, Water Science and Technology, 45 (4-5), pp. 405-411.

Kim, D.K., Jeong, K.S., McKay, R.I.B., Chon, T.S., Joo, G.J. 2012. Machine Learning for Predictive Management: Short and Long Term Prediction of Phytoplankton Biomass using Genetic Algorithm Based Recurrent Neural Networks. INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH, 6(1), 95-108.

Kohonen, T., Honkela, T., 2007 Kohonen network. *Scholarpedia*. http://www.scholarpedia.org/article/Kohonen_network.

Kokotos, D.X. and Linardatos, D.S., 2011. An application of data mining tools for the study of shipping safety in restricted waters. Safety Science, 49(2), pp. 192-197.

Kolodner, J., 1993. *Case-Based Reasoning*. Morgan Kaufmann (Kolodner, 1993)

Koller, D., Friedman, N., 2009. Probabilistic graphical models, MIT Press.

Koo, Choongwan; Hong, Taehoon; Lee, Minhyun; *et al.* (2013) Title: Estimation of the Monthly Average Daily Solar Radiation using Geographic Information System and Advanced Case-Based Reasoning Author(s):. Source: Environmental Science & Technology Volume: 47 Issue: 9 Pages: 4829-4839

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press

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

Krug, L.A., Gherardi, D.F.M., Stech, J.L., Leão, Z.M.A.N., Kikushi, R.K.P., Hruschka Jr, E.R.H. and Suggett, D.J. 2013. The construction of causal networks to estimate coral bleaching intensity. Environmental Modelling and Software, 42: 157-167.

Kumar, D.N., Reddy M.J., 2007. Multipurpose Reservoir Operation Using Particle Swarm Optimization. *Journal of Water Resources Planning and Management, ASCE*, **133(3)**, pp. 192-201.

Kuok, K.K., Harun, S., Shamsuddin, S.M., 2010. Particle Swarm Optimization Feedforward Neural Network for Hourly Rainfall Runoff Modeling in Bedup Basin. *Malaysia. International Journal of Civil & Environmental Engineering IJCEE*, **9(10)**

Kurt, A. and Oktay, A.B., 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Systems with Applications, 37(12), pp. 7986-7992.

Kusiak, A., Zeng, Y. and Zhang, Z., 2013. Modeling and analysis of pumps in a wastewater treatment plant: A data-mining approach. Engineering Applications of Artificial Intelligence, 26(7), pp. 1643-1651.

Landgrebe, T.C.W., Merdith, A., Dutkiewicz, A. and Müller, R.D., 2013. Relationships between paleogeography and opal occurrence in Australia: A data-mining approach. Computers & Geosciences, 56(0), pp. 76-82.

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, USA.

Lerner, A., Shasha, D., Wang, Z., Zhao, X., & Zhu, Y. (2004, June). Fast algorithms for time series with applications to finance, physics, music, biology, and other suspects. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data* (pp. 965-968). ACM.

Letourneau, A., Verburg, P.H. and Stehfest, E. 2012. A land-use systems approach to represent land-use dynamics at continental and global scales. Environmental Modelling and Software, 33: 61-79.

Li, D., Yang, H. Z., & Liang, X. F. (2013). Prediction analysis of a wastewater treatment system using a Bayesian network. *Environmental modelling & software*, *40*, 140-150.

Li, Xia; Yeh, Anthony Gar-On; Qian, Jun-ping; *et al.*, 2009. A Matching Algorithm for Detecting Land Use Changes Using Case-Based Reasoning Source: Photogrammetric Engineering and Remote Sensing Volume: 75 Issue: 11 Pages: 1319-1332 Published: NOV 2009

Liang, Z., Xinming, T., Lin, L., & Wenliang, J., 2005. Temporal Association Rule Mining based on T-Apriori Algorithm and its typical application. In *Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion*

Liao, C.J., Tseng, C.T., Luarn, P., 2007. A discrete version of particle swarm optimization for flow shop scheduling problems. *Comput. Oper. Res.*, **34(10)**, pp. 3099–3111.

Liukkonen, M., Laakso, I. and Hiltunen, Y. 2013. Advanced monitoring platform for industrial wastewater treatment: Multivariable approach using the self-organizing map. Environmental Modelling and Software, 35: 192-193.

Llanos, J., Rodrigo, M.A., Cañizares, P., Furtuna, R.P., and Curteanu, S. 2013. Neuro-evolutionary modelling of the electrodeposition stage of a polymer-supported

Lü, H., Yu, Z., Horton, R., Zhu, Y., Wang, Z., Hao, Z., Xiang L., 2011. Multi-scale assimilation of root zone soil water predictions. Article first published online: 15 MAR 2011. DOI: 10.1002/hyp.8034

Lukačić, Z., Kern, J., Gamberger, D., 2005. Do Various Machine Learning Systems Extract the Same Attributes as Relevant Strong Attributes? European Notes in Medical Informatics 1:1104-1109.

Lynam, T, 2016. Exploring social representations of adapting to climate change using topic modeling and Bayesian networks. Ecology and Society, Vol. 21, No. 4.

Martínez, M., Sànchez-Marrè, M., Comas, J., & Rodríguez-Roda, I., 2006. Case-based reasoning, a promising tool to face solids separation problems in the activated sludge process. *Water Science and Technology*, *53*(1), 209-216.

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.

Mategaonkar, M. and Eldho, T.I. 2012. Groundwater remediation optimization using a point collocation method and particle swarm optimization. Environmental Modelling and Software, 32: 37-48.

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

McKinney, D. C., & Lin, M. D. (1994). Genetic algorithm solution of groundwater management models. Water Resources Research, 30(6), 1897-1906.

Michalewicz, Z., 1996. Genetic Algorithms + Data Structures = Evolution Programs. 3erd edition, Springer-Verlag

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.

Millie, D.F., Weckman, G.R., Young II, W.A., Ivey, J., Carrick, H.J. and Fahnenstiel, G.L. 2012. Modeling micro-algal abundance with artificial neural networks: Demonstration of a heuristic 'Grey-Box' to deconvolve and quantify environmental influences. Environmental Modelling and Software, 38: 27-39.

Minister, J-B.H., Williams, N.P., Masters, T.G., Gilbert, J.F., Haase, J.S., 1995. Application of evolutionary programming to earthquake hypocenter determination, in Evolutionary Programming: Proc. of the Fourth Annual Conference on Evolutionary Programming, pp. 3-17.

Mohan, S., Loucks, D.P., 1995. Genetic algorithms for estimating model parameters, Integrated Water Resour. Plng. For the 21st Century, Proc. Of the 22nd Ann. Conf., ASCE, Cambridge, MA.

Montalvo, I., Izquierdo, J., Herrera, M., Pérez-García, R., 2014. Water Distribution System Computer-aided Design by Agent Swarm Optimization. Computer-Aided Civil and Infrastructure Engineering, 29(6), 433-448.

Montalvo, I., Izquierdo, J., Pérez, R., Tung, M.M., 2007. Particle Swarm Optimization applied to the design of water supply systems. Computer & Mathematics with Applications, 56(3), pp.769–776.

Montalvo, I., Izquierdo, J., Pérez, R., Iglesias, P.L., 2008. A diversity-enriched variant of discrete PSO applied to the design of Water Distribution Networks. Engineering Optimization, 40(7), pp. 655-668.

Montalvo, I., Izquierdo, J., Schwarze, S., Pérez-García, R., 2010a. Multi-objective Particle Swarm Optimization applied to water distribution systems design: an approach with human interaction. Mathematical and Computer Modelling, 52, pp. 1219-1227.

Montalvo, I., Izquierdo, J., Pérez, R., Herrera, M., 2010b. Improved performance of PSO with self-adaptive parameters for computing the optimal design of Water Supply Systems. Engineering Applications of Artificial Intelligence, 23(5), pp. 727-735.

Moore, D. S., Craig, B. A., & McCabe, G. P., 2012. *Introduction to the Practice of Statistics*. WH Freeman.

Muleta, M.K., Boulos, P.F., Orr, C.H., Ro, J.J., 2006. Using Genetic Algorithms and Particle Swarm Optimization for Optimal Design and Calibration of Large and Complex Urban Stormwater Management Models. *ASCE Conf. Proc.* 200, 113, DOI:10.1061/40856(200)113.

Mulligan, A.E., Brown, L.C., 1998. Genetic algorithms for calibrating water quality models, J. of Environmental Engineering, pp. 202-211.

Myers, R.H., Montgomery, D.C., Vining, G., 2002. Generalized linear models with applications in engineering and the sciences. Wiley

Nassr, Mohammed S. and Abu Naser, Samy S., Knowledge Based System for Diagnosing Pineapple Diseases, 2018. International Journal of Academic Pedagogical Research (IJAPR), 2(7), 12-19, July 2018. Available at SSRN: https://ssrn.com/abstract=3219802

Needham, B., Bullpitt, N., 2007. A primer on learning in Bayesian Networks for Computational Biology, PLOS Comp Bio

Ni, Q., Wang, L., Zheng, B., & Sivakumar, M., 2012. Evolutionary algorithm for water storage forecasting response to climate change with small data sets: The Wolonghu Wetland, China. *Environmental Engineering Science*, *29*(8), 814-820.

Norris, D., Mitchell, V., Cameron, D., & Carter, J. (2018, June). The petroleum Prospectivity of the Central Ridge to Flemish Cap Region, Offshore Newfoundland and Labrador, Canada. In 80th EAGE Conference and Exhibition 2018.

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.

Orduña Cabrera, Fernando, and Miquel Sànchez-Marrè, 2018. Environmental data stream mining through a case-based stochastic learning approach. *Environmental Modelling & Software* 106: 22-34.

Pacifici, Fabio, Marco Chini, and William J. Emery. "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification." *Remote Sensing of Environment* 113.6 (2009): 1276-1292.

Parr Rud, O., 2001. *Data Mining Cookbook- Modelling data for marketing, risk, and CRM*. Wiley.

Pham, Aksoy, 1995. RULES: a simple ruler extraction system. Expert Systems with Applications 8 (1).

Pereira, G.C., Ebecken, N.F.F., 2009. Knowledge discovering for coastal waters classification. Expert Systems with Applications, 36 (4), pp. 8604-8609.

Pérez-Bonilla, A., Gibert, K., 2007. Automatic generation of conceptual interpretation of clustering. Progress in Pattern Recognition, Image analysis and Applications. LNCS 4756: 653-663, Springer

Poch et al. 2004 M. Poch, J. Comas, I. Rodríguez-Roda, M. Sànchez-Marrè y U. Cortés. (2004) Designing and Building real Environmental Decision Support Systems Environmental Modelling & Software 19(9):857-873. Septiembre de 2004.

Popa, Andrei; Wood, William (2011) Application of case-based reasoning for well fracturing planning and execution. Journal of Natural Gas Science and Engineering Volume: 3 Issue: 6 Special Issue: SI Pages: 687-696 DOI: 10.1016/j.jngse.2011.07.013

Porto, V. W., Fogel, D. B., & Fogel, L. J., 1995. Alternative neural network training methods. *IEEE Intelligent Systems*, (3), 16-22.

Quinlan, J. R., 1986. Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Quinlan, J. R., 1996. Improved use of continuous attributes in C4. 5. *J. of artificial intelligence research*, *4*, 77-90.

Ramesh, V., & Ramar, K., 2011. Classification of agricultural land soils: a data mining approach. *Agricultural Journal*, *6*(3), 82-86.

Ramos-Martínez, E., Herrera, M., Izquierdo, J., Pérez-García, R., 2014. Ensemble of naïve Bayesian approaches for the study of biofilm development in drinking water distribution systems. International Journal of Computer Mathematics, 91(1), 135-146

Rajasekar, U., Weng, Q., 2009. Application of association rule mining for exploring the relationship between urban land surface temperature and biophysical/social parameters. *Photogrammetric Engineering & Remote Sensing*, *75*(4), 385-396.

Reddy, M. J., & Nagesh Kumar, D. (2007). Multi-objective particle swarm optimization for generating optimal trade-offs in reservoir operation. *Hydrological Processes: An International Journal*, *21*(21), 2897-2909.

Recknagel, F. *et al.*, 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. Journal of Hydroinformatics 4.2 (2002): 125-133.

Remm, K. (2004). Case-based predictions for species and habitat mapping. *Ecological Modelling*, *177*(3-4), 259-281.

Reynoso-Meza, G., Carreño-Alvarado, E.P., Montalvo, I., Izquierdo, J., 2017. Water pollution management with evolutionary multi-objective optimisation and preferences. Congress on Numerical Methods in Engineering CMN2017, SEMNI, Valencia, Spain, July 3-5.

Reynoso-Meza, G., Alves Ribeiro, V.H., Carreño-Alvarado, E.P., 2018. A Comparison of Preference Handling Techniques in Multi-Objective Optimisation for Water Distribution Systems. In: Advanced Hydroinformatic Techniques for the Simulation and Analysis of Water Supply and Distribution Systems, Eds.: Herrera, M., Meniconi, S., Alvisi, S., Izquierdo, J., MDPI, Basel, Switzerland (doi:10.3390/w9120996).

Riordan, D., & Hansen, B. K., 2002. A fuzzy case-based system for weather prediction. *Engineering Intelligent Systems for Electrical Engineering and Communications*, *10*(3), 139-146.

Robertson, A. W., Kirshner, S., & Smyth, P., 2003. Hidden Markov models for modeling daily rainfall occurrence over Brazil. *Information and Computer Science, University of California*.

Rodman, L. C., Jackson, J., Huizar III, R., & Meentemeyer, R. K., 2006. An Association Rule Discovery System for Geographic Data. In *2006 IEEE International Geoscience and Remote Sensing Symposium, Denver, CO, Jul*.

Rodriguez-Roda, I. R., Sànchez-Marrè, M., Comas, J., Baeza, J., Colprim, J., Lafuente, J., ... & Poch, M., 2002. A hybrid supervisory system to support WWTP operation: implementation and validation. *Water science and technology*, *45*(4-5), 289.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832-837.

Ruiz, Magda; Sin, Guerkan; Berjaga, Xavier; *et al.*, 2011. Multivariate Principal Component Analysis and Case-Based Reasoning for monitoring, fault detection and diagnosis in a WWTP. WATER SCIENCE AND TECHNOLOGY Volume: 64 Issue: 8 Pages: 1661-1667 DOI: 10.2166/wst.2011.517

Sànchez-Marrè, M., Gibert, K. and Sevilla, B., 2010. Integral support to environmental decision making throug GESCONDA. Proc. of the iEMSs Firth Biennial Meeting: Int'l Congress on Environmental Modelling and Software vol. I. Swayne, D. et alt. (Eds.): 2448-2448, Ottawa University, Ottawa, CA. Jul 2010. ISBN:978-88-903574-1-1}

Sen, Z., Oztopal, A., 2001. Genetic algorithms for the classification and prediction of precipitation occurrence, Hydrological Sciences, 46(2), pp. 255-268.

Serban 2010 Serban F., Vanschoren, J. Kietz, J.-U. and Bernstein, A, 2013. A survey of Intelligent Assistants for Data Analysis. *ACM Computing Surveys*, Vol.45(3), Article 31.

Shokouhi, S. V., Skalle, P., & Aamodt, A., 2014. An overview of case-based reasoning applications in drilling engineering. *Artificial Intelligence Review*, *41*(3), 317-329.

Silberzahn, R., Uhlmann, E.L., Martin, D.P., 2015. Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results PsyArXiv. http://doi.org/10.17605/OSF.IO/QKWST.

Simpson, A.R., Dandy, G.C., Murphy, L.J., 1994. Genetic algorithms compared to other techniques for pipe optimization, J. Water Resources Planning and Management, 120(4), pp. 423- 443.

SLi, D., Yang, H.Z. and Liang, X.F., 2013. Prediction analysis of a wastewater treatment system using a Bayesian network. Environmental Modelling and Software, 40: 140-150

Sikora, M., & Wróbel, Ł., 2010. Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines. *Archives of Mining Sciences*, *55*(1), 91-114.

Sim-Siam, Manuela, Palmira Carvalho, and Cecilia Sérgio, 2000. Cryptogrammic epiphytes as indicators of air quality around an industrial complex in the Tagus valley, Portugal. Factor analysis and environmental variables. *Cryptogamie Bryologie* 21(2):153-170.

Siwek, K., Osowski, S., 2016. Data Mining Methods for Prediction of Air Pollution. Int. Journal of Applied Mathematics and Computer Science 26(2):467-478.

Schneider, A., 2012. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. Remote Sensing of Environment, 124(0), pp. 689-704.

Sousa, L. R., and R. L. Sousa. "Risk associated to storage of CO2 in carboniferous formations. Application of Bayesian networks." Int. Workshop on CO2 Storage in Carboniferous Formations and Abandoned Coal Mines. 2011.

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

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

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 & Software*, *21*(11), 1536-1543.

Stebel, K., Espinosa, G., Giralt, F., Kindler, A., Rallo, R., Richter, M. and Schlink, U. 2013. Modeling airborne benzene in space and time with self-organizing maps and Bayesian techniques. Environmental Modelling and Software, 41: 151-162.

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.

Su, F., Zhou, C., Liu, B., Du, Y., Shao, Q., 2002 A new model to extract environmental pattern for fishing ground. Acta Oceanologica Sinica, 21 (4), pp. 483-493.

Sun, S. and Bertrand-Krajewski, J.-L. 2012. On calibration data selection: The case of storm water quality regression models. Environmental Modelling and Software, 35: 61-73.

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.

Taghavifar, H, Mardani, A., Taghavifar, L., 2013. A hybridized artificial neural network and imperialist competitive algorithm optimization approach for prediction of soil compaction in soil bin facility. MEASUREMENT 46(8), 288-2299.

Tasadduq, I., Rehman, S., & Bubshait, K. (2002). Application of neural networks for the prediction of hourly mean surface temperatures in Saudi Arabia. *Renewable Energy*, *25*(4), 545-554.

Taylor, M. K., Akeeagok, S., Andriashek, D., Barbour, W., Born, E. W., Calvert, W., and Stirling, I., 2001. Delineating Canadian and Greenland polar bear (Ursus maritimus) populations by cluster analysis of movements. *Canadian Journal of Zoology*, *79*(4), 690-709.

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.

Tirelli, F., T. and Pessani, D., 2011. Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example. Ecological Informatics, 6(5), pp. 309-315.

Tlidi, Mustapha, René Lefever, and Andrei Vladimirov. "On vegetation clustering, localized bare soil spots and fairy circles." *Dissipative Solitons: From Optics to Biology and Medicine*. Springer, Berlin, Heidelberg, 2008. 1-22.

Tourigny, N., Diallo, O., & Simian, G., 1998. Using a case-based reasoning approach to manage Sahelian agroforestry projects. *AI applications (USA)*.12: (1-3) 60-64

Troncoso, A., Ribera, P., Asencio-Cortés, G., Vega, I., & Gallego, D., 2018. Imbalanced classification techniques for monsoon forecasting based on a new climatic time series. *Environmental Modelling & Software*, *106*, 48-56.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.

Viaggi, D., Raggi, M., Gomez y Paloma, S. 2013. Modelling and interpreting the impact of policy and price scenarios on farm-household sustainability: Farming systems vs. result-driven clustering. Environmental Modelling and Software, 43: 96-108.

Wang, X. Rosalind, *et al.* "Spatiotemporal anomaly detection in gas monitoring sensor networks." *Wireless Sensor Networks*. Springer, Berlin, Heidelberg, 2008. 90-105.

Ward, J., 1963. *Hierarchical Grouping to Optimize an Objective Function*.

Weiss, G. and 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

Weiss, G., Provost, F., 2003. Learning when training data are costly: The effect of class distribution on tree induction, *Journal of Artificial Intelligence Research* 19, 315-354.

Widrow, B., Hoff, M. E., 1960. Adaptive switching circuits. IRE WESCON Convention Record, New York IRE 1960; 4:96-10.

Whigham, Peter A., and Friedrich Recknagel. "An inductive approach to ecological time series modelling by evolutionary computation." *Ecological Modelling* 146.1-3 (2001): 275-287.

Wilkinson, L., Friendly, M., 2009. The American Statistician 63(2): 179-184. doi:10.1198/tas.2009.0033

Whitten, I., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers. 3rd ed.

Wooldridge, J.M., 2009. Introductory econometrics: a modern approach. Cengage Learning. ISBN 0-324-58162-9.

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.

Yang, Qinli, *et al.* "Multi-label classification models for sustainable flood retention basins." *Environmental modelling & software* 32 (2012): 27-36.

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*. pp. 495-496.

Yeganeh, B., Motlagh, M.S.P., Rashidi, Y. and Kamalan, H., 2012. Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine Yu, L. & Wang, J.-q., Analysis of Water Resources Planning Based on Particle Swarm Optimization Algorithm. *2nd International Workshop on Intelligent Systems and Applications (ISA)*, Wuhan, pp. 1-3, 2010.

Yu, S. and Wei, Y., 2012. Prediction of China's coal production-environmental pollution based on a hybrid genetic algorithm-system dynamics model. Energy Policy, 42(0), pp. 521-529.

Zhang, J., Wu, Z., Cheng, C.-T., Zhang S.-Q., 2011. Improved particle swarm optimization algorithm for multi-reservoir system operation, *Water Science and Engineering*, **4(1)**, pp. 61-73.

Zhao, J., Chen, S., Wang, H., Ren, Y., Du, K., Xu, W., Zheng, H. and Jiang, B., 2012. Quantifying the impacts of socio-economic factors on air quality in Chinese cities from 2000 to 2009. Environmental Pollution, 167(0), pp. 148-154.

Zhu, W., Wang, J., Zhang, W. and Sun, D., 2012. Short-term effects of air pollution on lower respiratory diseases and forecasting by the group method of data handling. Atmospheric Environment, 51(0), pp. 29-38.

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

Ziliang, H. (2002). STUDY ON THE SUPERVISORY INFORMATION SYSTEM OF PLANT LEVEL IN FOSSIL FUEL POWER PLANT [J]. Automation of Electric Power Systems, 15, 000.

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