International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Tools for Environmental Data Mining and Intelligent Decision Support

Karina Gibert[a,b], Miquel Sànchez-Marrè[a,c], Beatriz Sevilla[a]
[a]Knowledge Engineering and Machine Learning Group, Universitat Politècnica de Catalunya-BarcelonaTech. Barcelona, Catalonia; [b]Department of Statistics and Operation Research, Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, Catalonia; [c]Software Department, Universitat Politècnica de Catalunya-BarcelonaTech, Catalonia;

Abstract: The joint workshop in Data Mining for Environmental Scientists and Intelligent Environmental Decision Support Systems tries to provide a common discussion forum to communicate environmentalists with data miners and intelligent decision support systems developers. As environmentalists are the consumers of both products, Data Mining (DM) and Intelligent Decision Support Systems (IDSS) are complementary disciplines that must properly link to permit the maximum benefits of data analysis at highest levels. Here, a preliminary analysis of the state of the software tools in environmental DM and IDSS is presented.

***Keywords*:** Data Mining; Knowledge Discovery of Data; Multidisciplinarity; Environmental Systems.

1 INTRODUCTION

In fact, the IDSS are complex systems including several components: data interpretation level, diagnosis step, decision support step, strategy planning level and actuation step. One of the most important ones are the data monitoring and data analysis modules (which include data-driven models). Both intensively exploit data, in our context, environmental data. An Intelligent Environmental Decision Support Systems (IEDSS) can integrate the expert knowledge, stored by human experts through years of experience, in the environmental process operation and management. In addition, some knowledge can be obtained through the intelligent analysis of large databases, coming from historical operation of the environmental process. Thus, data mining and knowledge acquisition, as well as reasoning over the acquired models, are key steps to build reliable IEDSS.

This picture shows how strong is the link between data mining and IEDSS. Here, a preliminary analysis of the software tools available in both Data Mining and IEDSS is presented.

Section 2 provides a first picture on the main data mining tools which could be useful for environmental scientists. Section 3 describes some IEDSS, some of them commercial packages that can provide general decision support in a certain envrironmental area, like forest fighting or waste water treatment plants. Anyway, authors are not aware in the literature of any IEDSS tool for a general environmental purpose. Some efforts are being made in this direction by some research groups, trying to set-up IEDSS tools, starting from some data mining tools, as is the case of GESCONDA.

2 DATA MINING SOFTWARE

In this section, some software tools available to perform Data Mining on real data are described. These software tools or packages include the most commonly used, based on recent polls among users undertaken by some institutions. Both commercial packages and free software packages are briefly described, although the authors do not pretend to be exhaustive.

2.1 Commercial Data Minig Tools

Among the *commercial data mining software* packages, the following ones should be mentioned:

SAS Enterprise Miner *(http://www.sas.com/technologies/analytics/datamining/miner/*: SAS Enterprise Miner streamlines the entire data mining process from data access to model deployment by supporting all necessary tasks within a single, integrated solution, all while providing the flexibility for efficient workgroup collaborations. It provides tools for graphical programming, avoiding manual coding, which makes easy to develop complex data mining processes. It was designed for business users, and provides several tools to help with preprocessing data (descriptive analysis, advanced statistical graphics) together with advanced predictive modeling tools and algorithms, including decision trees, neural nets, auto-neural nets, memory-based reasoning, linear and logistic regression, clustering, association rules, time series, among others, as well as facility for direct connection with data warehouses. It also offers tools for comparing the results of different modeling techniques. It is integrated with other tools from the wider SAS statistical framework, which at present is one of the most powerful statistical packages commercially available. Both SAS and Enterprise Miner are delivered as a distributed client-server system. Both are especially well suited for large organizations.

IBM SPSS Modeller (formerly Clementine) *(http://www-01.ibm.com/software/analytics/spss/)* was one of the first commercial tools oriented to Data Mining. Later absorbed by the firm SPSS, which also commercializes a very popular and widely used statistical package (SPSS). Clementine is designed to support the CRISP-DM, the *de facto* standard data mining methodology. It provides a visual interactive workflow interface supporting the data mining process and has an open architecture for integrating with other systems and all SPSS predictive analytics. It includes facilities for database access, text, survey and web data preparation, model management, automatic version control, user authentication, etc. From the point of view of data mining techniques, it provides neural networks, decision trees, rule induction, association rules, classification, data visualization and the statistical functionalities of SPSS (any kind of statistical modellind and multivariate analysis techniques). Recently has been acquired by IBM which commercializes it under the name IBM SPSS Modeller. This software probably replacing the previous Data Mining tool from IBM, Intelligent Miner, which is not supported by IBM anymore.

Salford Systems Predictive Modeling Mining Suite (SPM) *(http://www.salford-systems.com/).* The Salford Predictive Modeling Suite (SPM) includes CART (classification and regression trees), MARS (multivariate adaptative regression splines), TreeNet, RandomForests, as well as powerful new automation and modeling capabilities. SPM is a highly accurate and ultra-fast platform for developing predictive, descriptive, and analytical models from databases of any size, complexity, or organization. Salford SPM automation accelerates the process of model building by conducting substantial portions of the model exploration and refinement process for the analyst. While the analyst is always in full control, SPM optionally anticipates the analyst's next best steps and package a complete set of results from alternative modelling strategies for easy review. CART generates clear, easy-to-understand decision trees in the form of flow charts. MARS produces 2D and 3D plots of detected variable transformations and interactions. TreeNet constructs dependency plots of a target vs. individual predictors, as well as 3D interaction displays. RandomForests features include clusters and segments, anomaly tagging, and multivariate class discrimination.

Angoss Knowledge Studio *(http://www.angoss.com/predictive-analytics-software/overview/)* is a powerful predictive analytics software with a robust suite of desktop, client-server and in-database software products, Angoss delivers optimized recommendations that use information to help in making actionable and effective sales-driven and risk mitigation business decisions. The software is flexible, agile and visual— making predictive analytics accessible and easy to use for technical and business users. The Angoss Knowledge Studio is composed by the KnowledgeSEEKER – a market-leading business intelligence software solution with data mining and predictive analytics capabilities; the KnowledgeSTUDIO – advanced modeling and predictive analytics capabilities for high-performance business users and quantitative analysts; the StrategyBUILDER – a module available in KnowledgeSEEKER and KnowledgeSTUDIO that uniquely allows organizations to design and deploy predictive strategies using Strategy Trees; the Market Basket Analysis – an advanced modeling technique available in KnowledgeSTUDIO used to find associations between items or events; and the In-Database Analytics – an add-on module used to perform DM and predictive analytics directly on data stored in a database.

DBMiner *(http://www.dbminer.com/)*. DBMiner solutions are server applications providing powerful and highly scalable association, sequence and differential mining capabilities for Microsoft SQL Server Analysis Services platform, and they also provide market basket, sequence discovery and profit optimization for Microsoft Accelerator for Business Intelligence. DBMiner uses intelligent and automated processes to analyse large volumes of detailed data from relational databases, data warehouses and web data with exceptional ease of use and high versatility. DBMiner's products are based on over 10 years of innovative research and development.

GhostMiner *(http://www.fqs.pl/business_intelligence/products/ghostminer)* is a DM software from Fujitsu that not only supports common databases (or spreadsheets) and mature machine learning algorithms, but also assists with data preparation and selection, model validation, multimodels like committees or k-classifiers, and visualization. All is available in one package - a large range of data preparation techniques, a broad scope of selection of features methods and a choice of DM algorithms and visualization techniques are integrated. This means that only one data format (project) is needed, and so trying out and comparing different approaches becomes extremely easy. The package also comes with an intuitive interface, easy for non-technical users.

2.2 Free-software Data Mining tools

Among the *free software data mining packages*, the most popular data mining suites, due to the capabilities and collection of methods and experimentation that they offer, are the following ones:

Rapid Miner *(http://rapid-i.com/content/view/181/190/)* is an open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. A lot of applications of RapidMiner have been constructed. Its main outstanding functionalities are: GUI interface, Data Integration, Analytical ETL, Data Analysis, and Reporting in one single suite; Powerful but intuitive graphical user interface for the design of analysis processes; Repositories for process, data and meta data handling; Only solution with meta data transformation: forget trial and error and inspect results already during design time; Only solution which supports on-the-fly error recognition and quick fixes; Complete and flexible: Many formats of data loading, data transformation, data modeling, and data visualization methods.

Weka *(http://www.cs.waikato.ac.nz/ml/weka/index.html)* [Hall 2009] contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. Weka (Waikato Environment for Knowledge Analysis, current version Weka 3.7.5) supports several standard DM tasks: data preprocessing, clustering, classification, regression, visualization, and feature selection. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.
Weka contains several components. The main user interface is managed by the component named Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the component Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets rather than a single one. Weka is a general purpose package, freely available on the Internet and it became rather famous in the Artificial Intelligence community.

R *(http://www.r-project.org/)* is not exactly a DM tool, but a well-supported plattform, open source, command line driven, specialized in statistics and related methods. There are hundreds of extra *packages* freely available, which provide all sorts of data mining, machine learning and statistical techniques. It has a large number of users, particularly in the areas of bio-informatics and social science; also a large number of developers that continuously enlarge its functionalities providing new packages. It is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It is a very used free software package for DM. It includes:

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,

- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

R can be extended (easily) via *packages*. There are about eight packages supplied with the R distribution and many more are available covering a very wide range of modern statistics.
It is the most flexible tool presented in this paper, although it do not provides a GUI interface and some programming skills are required to use it properly.

KNIME *(Konstanz Information Miner*) (*http://www.knime.org/*) is a user-friendly and comprehensive open-source data integration, processing, analysis, and exploration platform. KNIME is a modern data analytics platform that allows performing sophisticated statistics and data mining on data to analyze trends and predict potential results. Its visual workbench combines data access, data transformation, initial investigation, powerful predictive analytics and visualization. KNIME also provides the ability to develop reports based on information or automate the application of new insight back into production systems. KNIME Desktop is open-source and available under GPL license. It can be extended to include professional support and large enterprise functionality, providing the best of both worlds. KNIME was developed (and will continue to be expanded) by the Chair for Bioinformatics and Information Mining at the University of Konstanz, Germany.
The KNIME base version already incorporates hundreds of processing nodes for data I/O, preprocessing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others. It integrates all analysis modules of the well known Weka data mining environment and additional plugins allow R-scripts to be run, offering access to a vast library of statistical routines.
KNIME is based on the Eclipse platform, which permits easy building and delivery of integrated tools. Through its modular API, it is easily extensible. When desired, custom nodes and types can be implemented in KNIME within hours thus extending KNIME to comprehend and provide first-tier support for highly domain-specific data. This modularity and extensibility permits KNIME to be employed in commercial production environments as well as teaching and research prototyping settings. KNIME is released under a dual licensing scheme. The open source license (GPL) allows KNIME to be downloaded, distributed, and used freely.

ADaM *(http://datamining.itsc.uah.edu/adam/)* The Algorithm Development and Mining System (ADaM) developed by the Information Technology and Systems Center at the University of Alabama in Huntsville is used to apply data mining technologies to remotely-sensed and other scientific data. The mining and image processing toolkits consist of interoperable components that can be linked together in a variety of ways for application to diverse problem domains. ADaM has over 100 components that can be configured to create customized mining processes. Preprocessing and analysis utilities aid users in applying data mining to their specific problems. New components can easily be added to adapt the system to different science problems.
The 4.0 release of ADaM is a significant architectural paradigm shift from previous versions. The latest version (4.0.2) (see release note) provides a solution that easily supports the integration of 3rd party algorithms and the reuse of ADaM components by other systems. ADaM 4.0.2 provides this support through the use of autonomous components in a distributed architecture. Each component is provided with a C, C++, or other application programming interface (API), an executable in support of generic scripting tools (e.g. Perl, Python, shell scripts) and eventually web service interfaces to support web and grid applications. ADaM 4.0.2 components are general purpose mining and image processing modules that can be easily reused for multiple solutions and disciplines. These components are well positioned to address the needs for distributed mining and image processing services in web and grid applications. This DM and Image Processing Toolkit is getting very used in Environmnental sciences requiring image processing abilities.

GESCONDA *(http://kemlg.upc.edu/projects/gesconda-1)* (Sànchez-Marrè et al. 2010; Gibert et al. 2004;) is the name given to an Intelligent Data Analysis System developed with the aim of facilitating KD and especially oriented to environmental databases. On the basis of previous experiences, it was designed as a four level architecture connecting the user with the environmental system or process: *Data Filtering* (Data cleaning; Missing data analysis and management; Outlier data analysis and management; Statistical one-way analysis; Statistical two-way analysis; Graphical visualization tools; Attribute or Variable transformation), *Recommendation and Meta-Knowledge Management* (Method suggestion (Gibert et al 2010); Parameter setting; Attribute or Variable Meta Knowledge

management), *Data Mining techniques* (Clustering techniques; Decision tree induction; Classification rule induction; Statistical Modelling), and *Knowledge Management and Reasoning* (Integration of different knowledge patterns; Validation of the acquired Knowledge pattern; Rule-based reasoning; Case-based reasoning; User interaction).
Central characteristics of GESCONDA are the integration of statistical, AI and mixed methods into a single tool for extracting knowledge contained in data. All techniques implemented in GESCONDA can share information among themselves to best co-operate for extracting knowledge. It also includes capability for explicit management of the results produced by the different methods. Portability of the software between platforms is provided by a common Java platform.
Initially, GESCONDA was conceived as a system for knowledge discovery and Data Mining. Currently, the system supports two new functionalities: A case-based reasoning engine and a rule-based reasoning shell are provided. These new skills of GESCONDA makes it a suitable prototype tool for the deployment of Intelligent Decision Support Systems, including all main steps like data preparation and filtering, data mining, model validation, reasoning abilities to generate solutions, and predictive models to support final users

KLASS *(*Gibert et al 2005; 2005b; 2008) KLASS is a software developed by the group of K. Gibert which was originally conceived for profiles discovery in ill-structured domains (a special kind of real domains with complex structure including heterogeneous data matrices with numerical and qualitative variables). It includes a high performance basic descriptive statistics (where the user has control to many parameters, like the intervals of the histograms, or the limits of the axes in plots), data cleaning and transformation (either using mathematical expressions, logical filters or via recodification or discretization, including proper methods), basic missing data treatment, knowledge management for classical or probabilized rules, hierarchical clustering and some mixtures of statistical and artificial intelligence tools to support knowledge discovery, like Clustering based on rules, which can introduce prior expert knowledge into the clustering process, or boxplot-based induction rules, which can be used as a pure classifier or to provide conceptual interpretations of the classes discovered by the clustering. Particularly interesting the possibility to include both numerical and qualitative variables in the data matrix and to use both for the clustering, as KLASS includes several compatibility distances or dissimilarities that permits to work with heterogeneous data matrices, like Gower's similarity coefficient, or Gibert's mixed metrics. Specific tools are oriented to support the interpretation of classes, like the visualization of the dendrogram, the Class Panel Graphs or the CCCS methodology, which finds concepts associated to final classes by taking into account the hierarchical structure of the clustering (Pérez-Bonilla 2007). One of the particularities of the system is that the output, either graphical or numerical, are produced in LaTeX font files, which are directly processed by the kernel of KLASS and automatically sent to the LaTeX viewer. From the final user point of view, there is no difference with other systems, since graphical representations are directly displayed on the screen. However, when those results are to be included in a scientific document, the user can choose among transforming it into a PostScript or PDF file, and manage as usual, or to get the original LaTeX font file to be included into a LaTeX document in native code. This provides quite a complete support to reporting phase and also permits to include KLASS results in technical papers.

3 INTELLIGENT ENVIRONMENTAL DECISION SUPPORT SYSTEMS

The available tools for intelligent environmental decision support are still difficult to locate. Some scientific works can be found in the literature, and there is a wide community developing intelligent environmental decision support systems. However, it is still early to find a completele developed background of software tools. Authors are not aware of the existence of general IEDSS developers, similar to general purpose data mining packages presented in previous section. In the following we mention and briefly describe some domain-oriented tools, specific for some environmental area, but with enough generality to be used in different applications.
CARMA [Branting 98] *http://carma.johnhastings.org/index.html*: *CA*se-based *R*angeland grasshopper *M*anagement *A*dvisor, it produces advice about the most economical responses to *rangeland grasshopper infestations* in the western U.S. CARMA also includes a prototype cropland advising module. This advisory system for grasshopper infestations has been successfully used since 1996. CARMA employs a variety of artificially-intelligent (AI) techniques to provide advice about the most environmentally and economically effective responses to grasshopper infestations.In the process, CARMA illustrates an approach to providing advice concerning the behavior of a complex biological

system by leveraging multiple, individually incomplete, knowledge sources including the introduction of a technique known as *approximate-model based adaptation* which integrates case-based reasoning with model-based reasoning for the purpose of prediction within complex physical systems.

atl_EDAR [Sànchez-Marrè 2004] *http://sisltech.net/en* Intelligent Decision Support System to monitor, control, *supervise and manage a Wastewater Treatment plant (WWTP) in real time*. Currently commercialized by SISLTech, a spin-off from the Universitat de Girona and Universitat Politècnica de Catalunya. The system improves the automatic control systems of the WWTP. It can generate action plans against problematic situations and provides optimum operating mangement proposals for daily management. Intelligently monitors the processes using expert knowledge, experiential knowledge and some fuzzy logic control module. The system can generate daily, weekly or monthly reports. It can also be used as training tool.
It is an interactive, flexible and adaptive platform, useful for WWTP from both urban and industrial environments. It allows full integration with PLC, SCADA and databases, and allows remote communication and/or accessibility. It can:

- Perform on-line Data acquisition and data processing, monitoring the information provided by the wastewater treatment processes
- Diagnose the operation states in the treatment processes, and detect abnormal situations
- Identify the causes of the detectect problems
- Propose action plans: To manage the problematic periods; to optimize the treatment and regeneration processes; to optimize energy consumption
- Manage, automatically and in real time the control links between treatment and regeneration processes

To do that, the system integrates mathematical models classical for WWTP control, rule-based reasoning and case-based reasoning inference engines, uncertain data, qualitative information, and prior expert knowledge.

QnD™ [Kiker et al., 2005]: The Questions and Decisions ™ screening model system was created to provide an effective and efficient tool to integrate *ecosystem, management, economic and socio-political factors into a user-friendly model/game framework*. The model framework is used in a larger process of stakeholder participation in order to generate questions and decisions for the management of complex environmental challenges. The model is written in object-oriented Java and can be deployed as a stand-alone program or as a web-based (browser-accessed) applet. The QnD model links spatial components within geographic information system (GIS) files to the abiotic (climatic) and biotic interactions that exist in an environmental system.
QnD can be constructed with any combination of detailed technical data or estimated interactions of the ecological/management/social/economic forces influencing an ecosystem. The model development is iterative and can be initiated quickly through conversations with users or stakeholders. Model alterations and/or more detailed processes can be added throughout the model development process. QnD can be used in a rigorous modeling role to mimic system elements obtained from scientific data or it can be used to create a “cartoon” style depiction of the system to promote greater learning and discussion from decision participants.

SIADEX [Asunción et al 2005] *http://decsai.ugr.es/siadex/* : it provides *assistance to design fighting plans for forest fires* by integrating several AI techniques. In fact, it is an AI Planning and Scheduling system that uses HTN as description language for planning domains and problem description. SIADEX is useful to assist technical staff in the decision making stages of a real forest fire fighting episode, or to train staff by reproducing past episodes and allowing them to share decisions with the planner.
It is based on four main components:

- web server, that centralizes all the flow of information between system and user
- the ontology server, that is the cornerstone of the architecture as the basis for knowledge sharing and exchange between all the components,
- the planning and the monitoring servers that are offered as intelligent services through the web server.

This perspective also allow to view SIADEX as a collaborative working environment where it provides two basic functionalities: the intelligent services offered to the user (the ontology, planning and monitoring modules) and a middleware level that interfaces these back-end services to the user front-end (a web browser) achieving several valuable goals like transparent acces to a distributed

architecture, an ubiquitous access to the services that allow the mobility of the user and his/her independence of the access device.
The basic planning process of SIADEX is a state-based forward HTN planning algorithm that, starting from the initial state and a goal expressed as a high-level task, iteratively decomposes that top-level task and its sub-tasks by selecting their decomposition methods according to the current state and following the order constraints posed in tasks decomposition schemes as a search-control strategy.
This process makes possible to know the current state of the world at every step in the planning process and, concretely, when preconditions of both methods and primitive actions are evaluated, what allows to incorporate significant inferencing and reasoning power as well as the ability to call external programs (might be the itself web services) to infer new knowledge by requesting information to external sources. For this purpose, SIADEX uses two mechanisms: deductive inference tasks and abductive inference rules

PICO [Perini & Susi, 2005] A Decision Support System for plant disease management used by technicians of the Advisory Service of Trentino region, Italy, and by the researchers in disease management techniques was developed. The aim of the system is twofold. On one side the entomologist (the expert) is given support while developing a pest model using Machine Learning techniques, given a set of biological and meteo data collected on different orchards, in different seasons. On the other side it supports the final user (agronomist and producer) in tuning the resulting model to the specific environmental characteristics of the territory under his own control.
In particular they aim to offer the possibility to follow the changes that occur in the insects behavior during the time also supporting the experts during the process of selecting the most relevant characteristics that influence the life of the insects, using some Machine Learning techniques like decision tree, that, given a set of parameters and their values for a particular territory in a certain period of time, are able to produce sets of rules induced by these data. It is agent-oriented software and they obtain the following six agents or actors:

- the actor GISP (Geographic Information Services Provider) to which the Advisor SW Agent delegates the goal use GIS techniques;
- the actor DBL (Disease Behavior Learner) which performs the plan run disease models on the basis of information extracted from the seasonal data on the disease;
- three wrapper actors, namely, the PDE-DBW (Plant Diseases Expert DB Wrapper) which takes care of retrieving meteo and orchard historical data; the wrapper of the database of the meteo service, called Meteo-DBW (Meteo Service DataBase Wrapper) which retrieves weather forecast; the Local Knowledge actor, which is the wrapper of the local database containing data relative to the orchards belonging to the area under the advisor control.
- the actor User Interface which manages the interaction between the user of the application (the actor Advisor) and the other specialized subactors of the Advisor SW Agent.

DIAGNOZA_MEDIU [Oprea 2005]: assists to *the air pollution state diagnosis and control in urban regions with industrial activity*. It is a rule-based expert system that provides qualitative information to a DSS used in an environmental protection management domain. This expert system is based on a knowledge base (which comprises data coming from environmental and meteorological database, and also forecasting data), an inference engine, an explanation module, a knowledge acquisition module, and a user interface. The system allows to handle uncertain knowledge, in the sense that a set of terms (corresponding to linguistic certainty values), is used to express the user's degree of confidence in the facts stored in the knowledge base. The system has been already applied for air pollution analysis and control in urban region.

SYRIADE [Zabeo et al., 2010**]** (Spatial decision support sYstem for Regional rIsk Assessment of DEgraded land) project is a project funded by EU Joint Research Center (JRC) in order to apply the EU Soil Thematic Strategy. The SYRADE DSS was implemented within a GIS platform and includes a set of tools which support the developed spatially resolved regional risk assessment methodology. The overall objective of the system is the ranking of potentially contaminated sites for priority of investigation, when no information on characterization and risk by site specific methodologies is available. The SYRIADE DSS is based on MCDA where the alternatives to be ranked are the different risk values related to sources generating hazard. Risk is evaluated by the aggregation of criteria related to the different elements characterizing the risk scenario (source, pathway and target), normalized by the use of fuzzy membership functions.

ETIC [Conruyt et al., 2008] is a publicly funded project, based on La Réunion Island in the South-West of Indian Ocean, whose goal is to develop innovative ideas and ICT solutions for the management of biodiversity research contents. ETIC is based on several thematic projects and a collaborative methodology, stressing partnerships between researchers, educators, decision- makers, enterprises, associations and end-users who wish to share and communicate their environmental data and knowledge off and on line. With the help of computer scientists, web designers, programmers and graphics experts, the common goal is to participate in the construction of an *Information System* (IS) for environmental management on the Internet. Content include terrestrial (i.e. Herbarium) and marine (i.e. Corals) biodiversity descriptions about specimens, their geography, ecology, photography, taxonomy and bibliography contextual information in La Réunion Island in the South-West Indian Ocean. The program was created in 2004 at University of La Réunion Island for research and knowledge enhancement of Insular Tropical Environments, by using AI techniques such as *Knowledge Engineering* for building expert systems and *Collective Intelligence* for building multi agent systems, and *Information and Communication* tools such as content management Systems.

4 CONCLUDING REMARKS

The tools presented do not pretend to be an exhaustive description of the available tools in both areas. Just to provide a first insight to the state of the art of the software tools available to environmentalists.

The first conclusion is that the development state of Data Mining software is a little bit more advanced that the one of IEDSS. This is a normal situation, since the IEDSS development hardly depends on the existence of reliable DM tools and it is a sensibly younger research area. In addition, the IEDSS development is a multi-level process, and its inherent complexity is harder than the complexity of a data mining tool, eventhough Data Mining itself involves a complex multi-step process. This means that there is a lot of work to do in this field and we might assist to an increasing activity in this area in the near future.

Regarding the DM tools, most of them provide graphical user interface to define the DM process and some of them permit to save the workflow and retrieve them in future sessions to execute with new data repeatedly. Only GESCOND includes a methodological recommender to help the end-user to decide, regarding their environmental goals, which data mining method can provide more suitable answers. Another important characteristic that R, SAS-Enterprise Miner or GESCONDA provide is the possibility to reuse the results of a certain DM process (or the induced model itself) in a subsequent part of the DM process, so reusing the mined knowledge for further data mining, which permits more powerful analysis. KLASS is very specifically oriented to clustering-based DM processes, but the integrated support to both numerical and non-numerical data is interesting, as well as the extensive choice of compatibility measures offered to perform clustering using both simoultaneously. Also interesting the postprocessing tools provided to support the interpretation of the results and the process of conceptualization of the classes.

Another interesting remark is that it seems that the current IEDSS development is intensively focused to a specific environmental problem or environmental domain. Apart from the paper [Sànchez-Marrè et al 2008] where a general architecture to develop IEDSS was established, the real existing tools are specifically oriented to waste water treatment plants, or forest fighting or specific infestations or urban air pollution. Probably, because the specific domain-knowledge is a clue in the IEDSS construction, and obviously, this is clearly depending on the problem. We think that in the near future the IEDSS will still develop for specific environmental domains because the area is still far from being able to develop designs general enough to parameterize the specific domain knowledge itself inside the IEDSS.

IEDSS are built by integrating several artificial intelligence methods, geographical information system components, mathematical or statistical techniques, and environmental/health ontologies, and some economic components. Although some architecture proposals are found in the literature to combine all this techniques and knowledge [Sànchez-Marrè et al., 2008], there is not a common framework to be taken into account as first guideline to deploy IEDSS with an easy way to integrate several AI models. Single AI techniques such as rule-based reasoning, fuzzy models, case-based reasoning, qualitative reasoning, artificial neural networks, genetic algorithms, model-based reasoning, Bayesian networks, and multi-agent systems provide a solid basis for construction of reliable and real applications, but

there is the general agreement among researchers that a *semantic interoperability of AI techniques* is one of the main open challenges in this field.
In a first approach it is difficult to get more precise knowledge about how the data mining methods are internally used in the described IEDSS. At first impression it seems that some of the models used in the IEDSS have been previously build using data mining and are integrated into the IEDSS to provide the decision support in front of a new situation. Current research is in progress to analyze how data mining is really bridged to real IEDSS. We expect to provide some insight on this issue in the extended paper after the conference.

REFERENCES

Branting, L. K. Integrating cases and models through approximate-model-based adaptation. *In Proceedings of the AAAI 1998 Spring Symposium on Multimodal Reasoning* (SS-98-04), *Menlo Park, CA, USA, AAAI Press*, 1-5, 1998.

Conruyt D., D.Sebastien, D. Payet, Y. Geynet, D. Caron, and D. Grosser. Managing Insular Tropical Environment through Data and Knowledge Bases by using Web Services: a case study on Corals and Herbarium of La Réunion Island. *In procs of 4th iEMSs'2008*, 1, 264-271, 2008.

De la Asunción, M., L. Castillo, J. Fdez-Olivares, O. García-Pérez, A. González and F. Palao, SIADEX: An interactive knowledge-based planner for decision support in forest fire fighting *AICom* 18(4): 257-268.

Gibert K. and R.Nonell, Pre and post-processing in KLASS. *Proc. of the iEMSs 4th biennal meeting: Int'l Congress of Environmental Modeling and Software (DM-TES'08 Workshop) iEMSs 2008*, 3, 1965-1966, 2008.

Gibert K. and R. Nonell, Descriptive statistics with KLASS. Supporting LaTeX documents elaboration. *In Procs 3rd World Conferencs on Computational Statistics and Data Analysis*, 90, 2005.

Gibert K., R. Nonell, J. M. Velarde and M. M. Colillas, Knowledge Discovery with clustering: impact of metrics and reporting phase by using KLASS. *Neural Network World* 4, 319-326, 2005.

Gibert K., M. Sànchez-Marrè and I. Rodríguez-Roda, GESCONDA: An Intelligent Data Analysis System for Knowledge Discovery and Management in Environmental Databases. *Env. Mod. and Software* 21(1), 115-120, 2005.

Gibert K., Sànchez-Marrè, V. Codina, Choosing the right data mining technique: classification of methods and intelligent recommenders. *In Proc. of the iEMSs Fifth Biennial Meeting: Int'l Congress on Environmental Modelling and Software*, 1, 2448-2453, 2010.

Hall M, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H.Witten, The weka data mining software: an update. *In ACM SIGKDD Explorations Newsletter*, 11(1), 10-18, 2009.

Kiker G.A. and I. Linkov, The QND model/game system: integrating questions and decisions for multiple stressors *G. Arapis et al. (eds.), Ecotoxicology, Ecological Risk Assessment and Multiple Stressors*, 203–225, 2006.

Kiker G.A., N.A. Rivers-Moore, M.K. Kiker and I. Linkov, QND: A Scenario-Based Gaming System for Modeling Environmental Processes and Management Decisions *B. Morel and I. Linkov (eds.) Environmental Security and Environmental Management: The Role of Risk Assessment*, 151–185. Springer, 2006.

Oprea M., A case study of knowledge modelling in an air pollution control decision support system *AI Communications - Binding Environmental Sciences and AI*, 18(4), 293-303, 2005.

Perinni A. and A. Susi, AI in support of Plant Disease Management. *AIComm* 18(4), 281-291, 2005.

Sànchez-Marrè M., K. Gibert and B. Sevilla, Evolving GESCONDA to an Intelligent Decision Support Tool. *5th International Congress on Environmental Modelling and Software, iEMSs'2010 Procceedings*, 3, 2015-2024, 2010.

Sànchez-Marrè M., J. Comas, I. Rodríguez-Roda, M. Poch and U. Cortés, Towards A Framework for the Development of Intelligent Environmental Decision Support Systems. *4th Int'l Congress on Environmental Modelling and Software (iEMSs'2008). iEMSs'2008 Procs*, 1, 398-406, 2008.

Sànchez-Marrè M., M. Martínez, I. Rodríguez-Roda, J. Alemany and U. Cortés. Using CBR to improve intelligent supervision and management of wastewater treatment plants: the atl_EDAR system. *Procc. of Industrial Day, 7th European Conference on Case-based Reasoning (ECCBR 2004) (Eds. Francisco Martin and Mehmet Göhker), Madrid*, 79-91, 2004.

Zabeo A., E. Semenzin, S. Torresan, S. Gottardo, L. Pizzol, J. Rizzi, S. Giove, A. Critto and A. Marcomini, Fuzzy logic based IEDSSs for environmental risk assessment and management. *5th Int'l Congress on Environmental Modelling and Software (iEMSs'2010). iEMSs'2010 Procceedings*, 3, 2073-2078. 2010.