Universidade de Aveiro
2011
Departamento de Electrónica, Telecomunicações e Informática

Joana Marques de Melo

AQUAWEB: Ferramenta de avaliação da qualidade da água

Universidade de Aveiro
2011

Departamento de Electrónica, Telecomunicações e Informática

Joana Marques de Melo

AQUAWEB: Ferramenta de avaliação da qualidade da água

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Mestrado Integrado de Engenharia de Computadores e Telemática, realizada sob a orientação científica do Dr. Carlos Manuel Azevedo Costa, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e da Dra. Maria João Feio, Investigadora Auxiliar do Instituto do Mar da Universidade de Coimbra.

Dedico este trabalho aos meus pais.

I dedicate this work to my parents.

o júri / the jury

Presidente / president

Prof. Dr. José Luis Guimarães Oliveira
Professor Associado do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Vogais / examiners committee

Prof. Dr. José Paulo Ferreira Lousado
Professor Adjunto do Departamento de Informática, Comunicações e Ciências Fundamentais da Escola Superior de Tecnologia e Gestão de Lamego do Instituto Politécnico de Viseu

Prof. Dr. Carlos Manuel Azevedo Costa
Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Dra. Maria João de Medeiros Brazão Lopes Feio
Investigadora Auxiliar do Instituto do Mar da Universidade de Coimbra

agradecimentos

Neste ano que passou adquiri muitos conhecimentos que não teriam sido possíveis sem o apoio das pessoas que me acompanharam.

Em primeiro lugar quero agradecer ao meu professor e orientador Carlos Costa por todo o apoio, disponibilidade e acompanhamento que me deu ao longo de toda a minha dissertação. À professora Maria João Feio pela paciência e disponibilidade para as minhas questões no que diz respeito à biologia. Agradeço também à equipa da Bioinformática do IEETA pela forma como me acolheram e sempre se mostraram disponíveis para ajudar no meu projecto, e em especial ao Samuel Campos pelo seu contributo no trabalho para esta dissertação.

Obrigada aos meus amigos de Aveiro, em especial ao grupo mais próximo, que acompanharam o meu trajecto Universitário.

Um enorme obrigada aos meus pais, pois sem eles nunca teria chegado até aqui nem seria quem sou, pelo seu apoio incondicional e constante durante toda a minha vida em particular na etapa universitária.

Márcio, a ti agradeço em especial por toda a paciência, dedicação, apoio e motivação em tantos momentos, estiveste sempre aqui para mim.

palavras-chave Biomonitorização, plataforma web, modelos preditivos, qualidade ecológica

resumo

A qualidade ecológica das águas dos rios é uma preocupação que tem vindo a afirmar-se cada vez mais em muitos países. Ao olharem com maior atenção para a qualidade dos rios, as entidades responsáveis podem tomar mais cedo medidas que previnam a sua degradação e que visem a sua reabilitação. Com esta necessidade, surgiram vários métodos de avaliação da qualidade ecológica dos rios, entre eles os modelos preditivos baseados em comunidades aquáticas e nas características ambientais dos rios. Estes modelos baseiam-se em tarefas complexas, lidando com extensas quantidades de dados e, por esse motivo, os biólogos recorrem a programas de computador que integram diversas ferramentas estatísticas e operam grandes quantidades de informação.

Esta dissertação apresenta uma solução integrada de todos os processos inerentes à criação e utilização de modelos preditivos baseados em dois tipos de modelos existentes: RIVPACS (River Invertebrate Prediction and Classification System) e BEAST (Benthic Assessment of Sediment). Ambos os tipos de modelos seguem a filosofia da RCA (Reference Condition Approach), criando o modelo com base nas comunidades biológicas de locais de referência (i.e., boa qualidade). Locais possivelmente afectados são avaliados quanto ao seu grau de perturbação pela diferença da sua comunidade às comunidades referência.

A solução foi concretizada numa aplicação web, construída para que os biólogos possam gerir os dados obtidos em campo sem qualquer tratamento prévio e, ao mesmo tempo, oferece uma visão próxima e simplificada de todos os resultados de cada etapa do processo, permitindo sempre ajustamentos ao longo de cada passo. Ao facilitar a verificação e edição dos dados recolhidos de uma forma intuitiva, detalhada e bastante visual, permite uma eficiente detecção de falhas nos dados que invalidariam a sua utilização. Também de forma intuitiva, são providenciados em cada passo resultados sob a forma de imagens, tabelas e texto, para melhor ajudar o utilizador final a decidir sobre as opções a tomar para o passo seguinte do algoritmo.

keywords Bioassessment, web platform, predictive modeling, ecological quality

abstract

There is a growing concern in many countries about the quality of streams and rivers ecosystems. By paying close attention to the ecological quality of rivers, the responsible entities can take early measures to prevent their degradation and for their rehabilitation. With this necessity, several assessment methods have emerged, including predictive models based on aquatic communities and environmental characteristics. These models integrate complex tasks, deal with large amounts of data, and for those reasons biologists rely on computer programs designed to perform statistical calculations on large sets of data.

This dissertation presents an integrated solution of all the steps involving the creation and use of predictive models based on two types of existing models: RIVPACS (River Invertebrate Prediction and Classification System) and BEAST (Benthic Assessment of Sediment). Both types of models follow the RCA (Reference Condition Approach), since they create the models based on reference sites (i.e., good quality sites). To determine the degradation level of disturbed sites, the models compare their biological communities to those of reference sites.

The solution was implemented in a web application, built so that biologists can manage the data obtained in the field without any prior treatment and at the same time offering a simplified and close vision of all the results produced at each step of the algorithm, always allowing adjustments throughout the process. By facilitating the verification and editing of collected data in an intuitive, comprehensive and very visual way, it allows efficient detection of data errors that would preclude the validity of this information for assessments. Also intuitively, results are provided in each step in the form of images, tables and text, to help the end user decide on the choices to make for the next step of the algorithm.

Contents

List of Figures

Acronym list

ANN	Artificial Neural Networks
ANNA	Assessment by Nearest Neighbor Analysis
ANOVA	Analysis of Variance
AUSRIVAS	Australian River Assessment Scheme
BBN	Bayesian Belief Network
BC	Bray-Curtis
BEAST	Benthic Assessment of Sediment
CABIN	Canadian Aquatic Biomonitoring Network
CSV	Comma Separated Values
CT	Classification Tree
CV	Cross-Validation
DFA	Discriminant Function Analysis
FAQ	Frequently Asked Question
HTTP	Hypertext Transfer Protocol
IDE	Integrated Development Environment
LOOCV	Leave-One-Out Cross-Validation
MAC	Macrophyte Assessment and Classification
MACPACS	Macrophyte Prediction And Classification System
MDA	Multiple Discriminant Analysis
MDS	Multi Dimensional Scaling
MEDPACS	Mediterranean Prediction And Classification System
MS	Member State
ODS	Open Document Spreadsheet
PAC	Principal Axis Correlation
PHP	PHP: Hypertext Preprocessor
RCA	Reference Condition Approach
RIVPACS	River Invertebrate Prediction And Classification System

RMSE	Root Mean Squared Error
RSSD	Replicate-Sampling Standard Deviation
SD	Standard Deviation
SVM	Support Vector Machine
SYSTAT	System Statistics
TWINSPAN	Two Way Indicator Species Analysis
WFD	Water Framework Directive
XLS	Excel Spreadsheet

1. Introduction

In this chapter, the necessity for the creation of the present platform is discussed, as well as the defined requirements to tackle this objective. Finally, the structure of this dissertation is provided.

1.1. Motivation

The evaluation of the ecological quality of water bodies is a present key concern in developed countries over the world. Scientists and governmental agencies have been working together in order to create effective methods of detecting the state of the aquatic quality in streams and rivers. Many techniques have therefore emerged for the assessment of water bodies throughout the XX century, including biotic indices, saprobic indices and predictive modeling.

Among these methods, predictive modeling became widely used. Predictive modeling is a powerful approach to evaluate the quality of water bodies, but several problems arise for the utilization of such techniques. It demands the execution of numerous heavy calculations and sequences of multivariate analysis that need time to process and are difficult to manage. The effort and time necessary to create predictive models are the main vicissitudes preventing a wider applicability of these approaches. Thereupon, these approaches are not being used as much as desired by biologists, researchers, technicians and environmental authorities.

There are already tools, some of them in form of web sites, for facilitating the assessment of streams through predictive models. However, these tools have restricted application as they force the researcher to accept the models already created and available. Therefore, the system is not applicable for countries other than the one the models were intended to, since the natural characteristics can diverge from one country to another.

Moreover, identifying aquatic species to the most specific level possible is a very difficult challenge that typically demands the researcher to perform this process manually, through the help of paper manuals with images.

The identification of the macroinvertebrate taxa (the identification of organisms) present at a stream site is the basis for the application of any bioassessment tool. Therefore, there was also the need of incorporating in the same website a digital taxonomic key, based on photographs and videos, to facilitate the identification of the Portuguese macroinvertebrate taxa, and which was previously developed at the Institute of Marine Research to be incorporated into this project.

Creating a framework that would provide all of these features to simplify the procedures that researchers must follow to obtain results is thus valuable.

A web platform was previously created with the purpose of tackling these needs: RIO. This platform provided a basic process of creating predictive models and making assessments. Despite that, a few of the capabilities were incomplete, or needed to be updated in order to fit with the recent developments in the predictive modeling approaches being used. Likewise, new requirements were introduced for this project, as new necessities emerged for the web site utilization.

This project had the RIO web platform as a starting point for improvements and additions.

1.2.Goals

To improve the previous project RIO and to manage all the required procedures to use the predictive modeling approach in a more detailed view, an enhanced integrated software tool is necessary.

The web platform for this project was developed with the purpose of:

- Allowing the creation of predictive models in a more integrated and complete logic, in which the information does not need to be processed *a priori*;
- Distributing the several calculations involving the predictive modeling studies in separated and compartmentalized steps, in which intermediate results and configuration parameters are provided to best fit necessities;
- Providing specific and detailed information regarding errors existent in tables of spreadsheets provided for assessment;
- Managing biological and environmental information onto a global “database” through the form of an editable spreadsheet within the web site;
- Showing visual final results of assessments through a map with color-based pins identifying the quality of each assessed site in its location;
- Integrating a digital taxonomic key identification tool of Portuguese macroinvertebrate taxa.

Such goals should be integrated into the web site RIO, which shall onwards be named AQUAWEB.

1.3.Dissertation structure

The dissertation consists of the following remaining chapters:

Chapter 2 presents the bibliographic research for the development of this project. It includes the origin of the needs for water quality evaluation, the main developed techniques for this matter and the previous work, RIO.

Chapter 3 provides a full description of the technologies and the architectural design of the system, providing the insight of the structure developed for the support of this work. Moreover, the description of the involved technologies utilized to attain the requirements and the elected assessment approaches are presented in this chapter.

Chapter 4 describes the organization of the features that were included in the implementation of this work, as well as the complete sequence for the utilization of these features.

Chapter 5 provides the conclusions of this research, identifying the satisfied results, remaining issues and future work.

2. State of the Art

2.1. Motivation for bioassessment

All over the world, streams and rivers ecosystems suffer from increasing risk of integrity loss due to urbanization, agriculture and energy needs for human populations. Because of this, there is a progressive concern to monitor the quality of aquatic ecosystems [1]. In cases of deterioration, the assessments are the foundation for the establishment of recovery measures.

Initially, the quality of streams and rivers was prepared strictly through the evaluation of water physics and chemistry. Over the years, the biological communities were integrated in the quality assessments (the bioassessment), as they are directly influenced by the loss of water quality and hydromorphological changes in water bodies [2].

Benthic macroinvertebrates are the most often used organisms for assessment (used in over 90% of the American assessment programs), for being relatively sedentary, widespread and easy to collect, relatively quick and easy to identify to the upper taxonomic levels, such as classes or families and for responding to common stressors [3]. They also are ecologically diverse and therefore provide a wide variety of responses to different kinds of environmental stressors, which makes the identification of the stress type easier to define [1].

Presently, other biological elements such as diatoms, fish and macrophytes are also used in bioassessment of freshwaters. These elements are complementary in their sensitivity to human pressure, since each one responds differently to various types of degradation (chemical, hydrological and morphological) [4-6].

The first step of bioassessment is the definition of ecological quality goals. Thus, the streams have to be classified according to quality classes throughout a series of calculations. By using the biological information, the evaluation of water quality can be obtained through several methods, such as saprobic indices, biotic indices, multimetric indices and predictive models.

Saprobic indices define four zones of self-purification (which means that each of them define a level of pollution, from slight or none to extremely severe), indicated by the presence of certain saprobic species. Hence, comparing the species list from a sampling site to the list of species present in each zone enables the site to be classified into a quality category [7].

Biotic indices assign a level of tolerance to organisms, from tolerant to intolerant. The scores attributed to the complete collection of organisms are summed, in order to obtain a final score. Finally, the total score is matched with a scale using intervals of scores for a given region. The result is the quality status of the water body [8].

Multimetric indices are another form of evaluation. These indices are obtained by deriving several metrics from a biological assemblage, which translate in different aspects, such as the percentage of sensitive taxa, percentage of animals in different trophic groups, among others. Each metric is weighted according to its relative importance for bioassessment and afterwards they are summed. This final value is juxtaposed to a scale to obtain a quality class [9].

Predictive modeling is a process used to build a statistical model that predicts the community composition expected on a given river site in the absence of disturbance. In the present work, this approach is used and will be further explained.

2.2.Water Framework Directive

With the purpose of achieving a good environmental quality of all water bodies, a "policy" for water monitoring and water assessment in the European Union, the Water Framework Directive (WFD) was proposed in 2000. This directive encourages the insurance of "good state" of waters, pollution control and a better management of territorial planning, like new industrial or civil sites improving water sustainability, among others [10]. For this, it establishes policies and measurement programs for all the Member States (MS), in order to attain comparable results of the status of the water quality. To tackle this goal, methods for water protection, monitoring and management emerged or were adapted by all MS [11]. This directive had an initial implementation timeframe ending in 2015 [10].

The WFD defines long-term and general objectives for the MS and specific rules for the establishment of bioassessment programs. However, each MS has a different approach to the requirements. For example, the MS may define different methods for sampling invertebrates or diatoms assemblages, different taxonomic levels and indices provided that the quality classes and respective levels of disturbance measured are comparable [11].

2.3.Reference Condition Approach

The Reference Condition Approach (RCA) is nowadays the basis of many assessment programs in Europe, USA, UK, Australia and elsewhere. It exonerates that the quality status of a site should be measured through the distance between its communities and the communities expected under non-disturbed conditions (reference conditions), for comparable environmental conditions (e.g. climate, geology, altitude) [12, 13].

To build an assessment program based on the RCA it is necessary to select a set of reference sites, which are stretches of rivers considered in very good conditions (ideally pristine) and that fully characterize the diversity of environments in the target area. Usually, these are the

best available sites in the present, even though in the absence of real good conditions, reference could be obtained from historical data or modeled. The “reference condition” will provide the goals for the recovery of degraded streams or rivers. Sites to be evaluated, test sites, can either be compared to a subset of the reference sites that are expected to be alike, or to all of them with probability weightings that translate their environmental similarity. Thus, the deviation of a test site from the reference condition indicates its quality, which in turn is a measure of the effect of stressors in the ecosystem.

Multimetric and multivariate methods (predictive models) follow the RCA and both have strong arguments to be employed. A main conceptual difference between them is that while the multimetric methods are based on an *a priori* geographic and physical data for classification, the multivariate methods are based on biological classification and *a posteriori* determination of which environmental variables distinguish the biological groups [3, 12].

2.4.The origin of predictive models

Among the various bioassessment methods, predictive models are robust statistical methods which follow the Reference Condition Approach. They are constructed based on large sets of biological and environmental data from many sites known to be in good condition (reference sites). The covariation of the biota and environmental features of healthy ecosystems enables the construction of predictive models.

Since the first predictive model developed for the bioassessment of UK streams – the RIVPACS (River InVertebrate Prediction And Classification System) [2, 14] –, different predictive approaches and models have been elaborated over the world. In this section, a review of the most common types of models is presented.

2.4.1. RIVPACS and AUSRIVAS models

The first predictive model for bioassessment, the River InVertebrate Prediction And Classification System (RIVPACS), was created from the necessity of evaluating both biological and environmental information of UK rivers [14]. The RIVPACS project started in the 1970’s and its goals were the collection of data of unpolluted UK running-waters and the assessment of rivers quality through their macroinvertebrate communities. Later, software in the programming language FORTRAN was developed to perform the sequence of statistical methods used by the predictive models [1, 2].

For the creation of such models, reference sites that have similar biota values are firstly placed in similar groups by classification analysis. The technique used was the TWINSPAN (Two Way Indicator Species analysis), which repeatedly splits the reference sites into sub-groups following a hierarchical logic based on their dissimilarities. The assortment is normally accomplished with presence/absence information of taxa.

After grouping the sites based on the previous hierarchy, the sites in each group are checked for similarities in their environmental attributes with Multiple Discriminant Analysis (MDA), a method for compressing a multivariate signal into a result of a lower dimension for a liable classification [15]. The probability of a new test site belonging to a group based on its values for the environmental predictor variables can be calculated from the Mahalanobis distance[1] of the test site from the centre of the group [15]. Also, comparing various clustering methods, it was considered that similarity in group sizes is a useful attribute [16]. In RIVPACS, a new test site is assigned to each group by weighing its probability. Then, the Expected and Observed values are computed.

The Expected value (E) for a test site is given by the sum of the probabilities of the expected taxa down to a given level, frequently defined at 50% [2]. This percentage is obtained comparing the environmental information from the reference group with the test site, supposing that the test site is not stressed [17].

The Observed value (O) at a site is a sum of taxon occurrences. This value compared to the expected fauna, an O/E (Observed/Expected) value, is calculated. Ideally, an O/E value close to unity indicates a high quality, i.e., similar to reference condition.

The quality levels in RIVPACS are called bands and correspond to intervals of O/E values. The upper band width corresponds to the distribution of O/E reference sites values down to a pre-defined percentile. So, commonly, the 90^{th} percentile of reference sites values is used to define the lower limit of band A; the next two bands, B and C have theoretically the same widths of band A. However, the last band is limited by zero.

As in UK, Australia started a growing concern with water quality, which led to the National River Health Program (NRHP) creation. This Program created the AUSRIVAS (Australian River Assessment Scheme) [18], which was based on the RIVPACS approach. Each State and Territory was defined as a different eco-region and prediction and classification systems were created for each of them.

There are some significant differences between the UK and Australia for this matter, such as:

- the size and variety of landscape;
- the number of users;
- the present taxa;
- the level of the taxonomic resolution;
- the model building and running;
- outputs;

[1] Mahalanobis distance is a measure based on correlations between variables intended to find patterns and similarities.

- the different predictors for each model;
- the taxa probabilities limit for them to be taken into consideration, among others [19].

For such reasons, some changes were made in the prediction system. For example, the establishment of reference groups was based on the unweighted pair-group mean arithmetic averaging (UPGMA), a hierarchical clustering method [20] and in reference sites Bray-Curtis pairwise similarity [19, 21], which is a statistic that measures the ratio between the species that are disjoint at the two sites and the total of species in the two sites. The groups had a minimum of 5 sites each, as less could lead to a poor representation of an habitat type or may have resulted from errors in the sampling or degradation of sites [19].

Then, as in RIVPACS, they used a stepwise multiple discriminant analysis (MDA) with cross-validation to assign sites to a reference group. Cross-validation is a method in which sites are taken from the initial reference set and, using the discriminant variables, the final group assigned to that site is compared with the originally assigned group. To verify the accuracy of a model, a validation set (a given % of the total initial number of reference sites) is set aside from model building and run later on through the model. Because the validation set is similar to the calibration set in terms of quality, the accuracy provides the correctness of the model.

The representation of quality levels in AUSRIVAS is made by the definition of bands, very similar to RIVPACS. The difference is that if the test site has an O/E ratio greater than the 90^{th} percentile of reference sites, it is considered "richer than reference" and belongs to band X (this does not necessarily imply an exceptionally good quality and needs further testing), while RIVPACS sets the A band from the 90^{th} percentile to unity, not considering a X band [18].

As the AUSRIVAS, other models based on RIVPACS-type of model were created, such as SWEPAC in Sweden in 2001 [22], PERLA in Czech Republic in 2006 [23], regional and national predictive models in Portugal in 2009 [24], MEDPACS (MEDiterranean Prediction And Classification System) in Spain in 2009 [25], the MACPACS (MACrophyte Prediction And Classification System) in Portugal in 2010 [4], among others.

2.4.2. BEAST

The BEAST (Benthic Assessment of Sediment) is also a type of predictive model, originally developed and implemented in North America for Great Lakes [26] and in British Columbia [27]. The method responds to the Canadian Aquatic Biomonitoring Network (CABIN) need of getting standardized information about biological features on water, to provide quality classifications of streams. This approach is different than RIVPACS in some points. Initially, the models were created using principal axis correlation analysis (PAC), a multiple linear regression method for describing the capacity of the environmental data to be fitted to the ordination axes created from the species matrix. It also uses a stepwise multiple discriminant analysis for the choice of environmental variables that best fit the grouping based on the biological

features and ANOVA (Analysis of Variance) to obtain the environmental variables that best distinguish the different groups [28].

Although the group definition is based on biological values, it only assigns a site to its most probable group, contrary to RIVPACS, which uses probabilities of belonging to each group. Then, the test site and the reference sites of the most probable group are compared by calculating the Bray-Curtis dissimilarities and using semi-strong hybrid multi-dimensional scaling (MDS) as an ordination technique [29]. The impairment level of the site is given by the distance of community composition of the test site to the reference sites of the assigned group, in an MDS-ordination space. Over this space, 3 Gaussian probability ellipses (90, 99 and 99.9% of the points) are added, to define intervals (bands) of increasing distance from the centroid.

Figure 1 represents an example of an MDS-ordination space with a test site (in red on the right) with its reference group (the group of points in green). The black point represents the centroid.

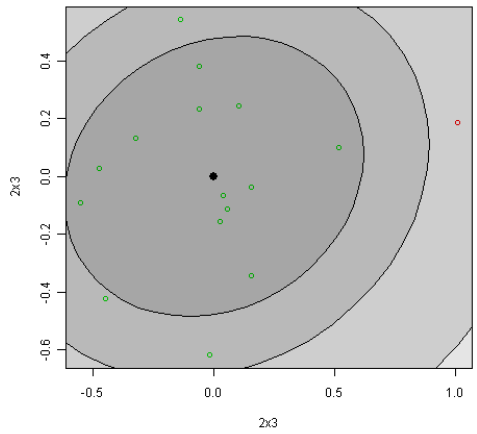


Figure 1 MDS ordination-space with probability ellipses

The positioning of the test site determines its quality band:

- The first band, meaning “equivalent from reference”, lies in the inner ellipse (up until 90%);
- the second band, “potentially different from reference”, is between the 90% and the 99% ellipse;
- the third, “different from reference”, is between 99% and 99.9%;
- and the remaining space, further from the 99.9% ellipse, is for a site “very different from reference” [30].

The BEAST method was adopted to other locations, such as Portugal, first with macroinvertebrates [6, 31, 32], and then with diatoms [6] and macrophytes [4, 32], and also in Brazil [33].

In Portugal, the method was first tested in the Mondego River catchment [31, 32] and lately the same methodology was adopted to build models for central Portugal [32] and for the entire country [4]. For establishing the reference biological groups, the Bray-Curtis coefficient and UPGMA (Unweighted Pair Group Method with Arithmetic Mean) were used. SFDA (Stepwise Forward Discriminant Analysis) was performed with cross-validation classification to elect the best available choice of predictor variables[32].

2.4.3. Other types of predictive modeling

Other types of models emerged from the need of the assessment of aquatic communities, using several techniques, such as artificial intelligence. A few of these are presented in this section.

2.4.3.1.ANNA (Assessment by Nearest Neighbor *Analysis)*

ANNA has a very similar approach to RIVPACS. The fundamental difference from these two is that ANNA skips the classification, which may be considered artificial, as it implies a division of the natural continuum, and the DFA (Discriminant Function Analysis) when comparing test sites to reference sites. ANNA searches the reference sites that best resemble the test sites based on their discriminant environmental variables. Then, it predicts the expected community at the test sites based on the communities of those nearest neighbors. This defines the reference set to compare as a continuum, opposed to the discrete groups in RIVPACS [34].

2.4.3.2.Artificial Neural Networks (ANN)

Predictive models have been created based in ANN, which are computational models inspired by the structural and functional characteristics of biological neural networks. An ANN is a connected assembling of artificial neurons that may change its structure based on either external or internal information received in the learning phase of the network. These neural networks model complex relationships between different data in order to find patterns. The most common types of ANN applied in freshwater predictive modeling have a training algorithm which learns patterns. Throughout the training, the calculated output and the actual output are compared and the connection weights of the neuron connections are altered to reduce the error. This approach has been applied to predict presence or absence of invertebrates. The observed and predicted values may then be computed to assess the

impairment of a site. Moreover, ANN has also been used to predict scores in biotic indices [35, 36].

2.4.3.3.Support Vector Machines (SVM)

This technique is inspired by characteristics of biological information processing. Multi-class problems are successively divided into two groups through pairwise classification, by extracting spatial patterns and describing highly nonlinear and complex data [37]. A generic example of an SVM with two variables with 120 points with normal distributions having a mean of 0.4 and 1.5 is given in Figure 2, where the division in two groups is defined by the blue and red colors.

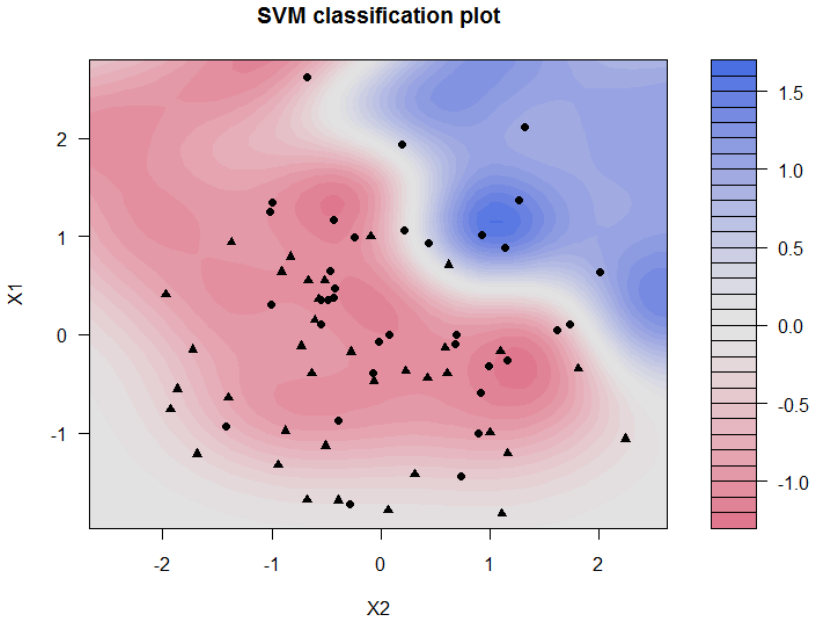


Figure 2 Example of a nonlinear SVM classification with two groups of 120 variables with normal distributions of 0.4 and 1.5

SVMs have been implemented using environmental variables to predict taxa occurrence by successively dividing the set of results in two sub-groups based on these environmental values, such as “pH above 3” or “altitude between 100 and 250”. The final value of a leaf of the SVM (the final step) is the predicted value for the taxa [37].

2.4.3.4.Classification Trees (CT)

The method of Classification Trees (CT) uses a map of observations about an occurrence in a set of data to infer the class it belongs to, by growing a binary tree (Figure 3 [38]).

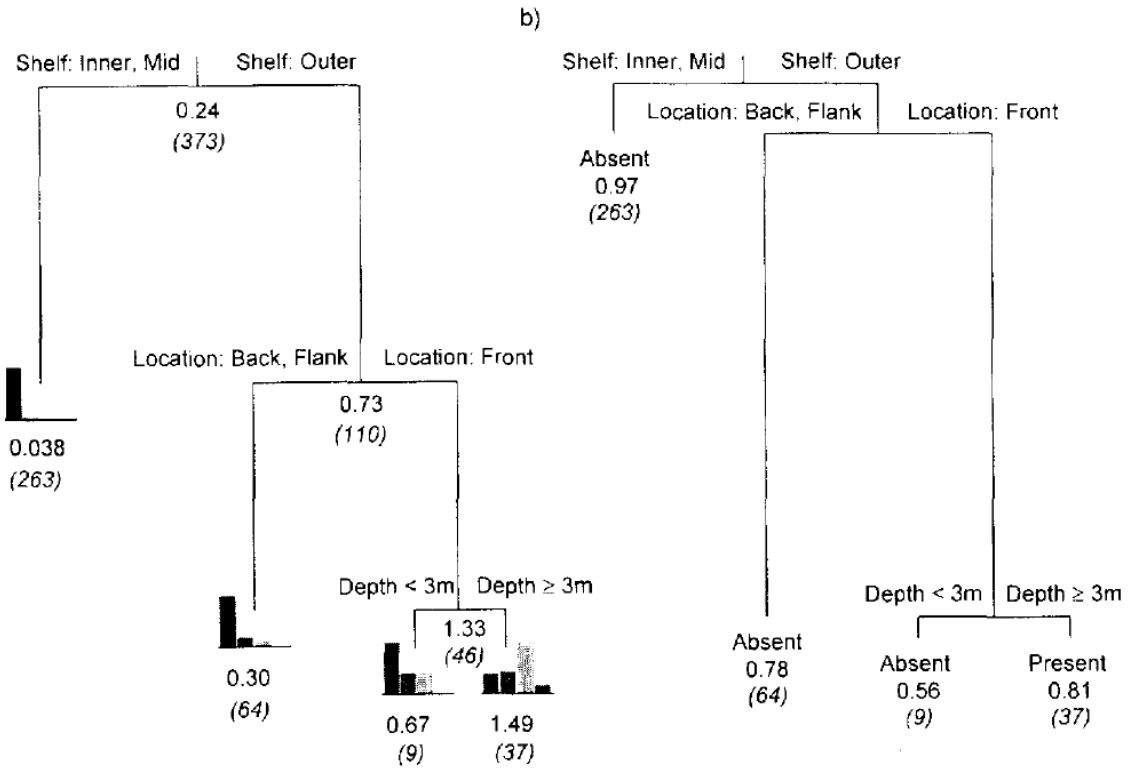


Figure 3 A Classification Tree example

The goal is to predict the response of a class to several inputs. At each node of the tree, a test is applied to one of the inputs and the answer defines the sub-node to go to. An implementation of this approach has been studied in Vietnam [37].

2.4.3.5.Bayesian Belief Networks (BBN)

BBNs are models with a network structure that are based on the assumption of causal relationships between variables and a group of conditional probability matrices, which relate each variable to its assumed causal variables. BNNs are directed acyclic graphs, which means that the relationships are oriented (have only one direction) and that starting at some node, there is not a sequence of nodes that may loop back to that node again. This technique has been developed in freshwater biological assessment with the intent of predicting taxa by creating these causal networks and assessing the predicted data and evaluating the proximity to the observed values [39].

2.5.Further developments in predictive models statistics

Along with the evolution of predictive modeling, appeared several improvements, alternatives and options on the creation and verification of performance and accuracy.

2.5.1. Stepwise vs. all-subsets

The predictor variables selection in a model is a crucial step for its performance. It is commonly executed with the help of multiple-regression methods. These methods have the goal of finding a relation between a dependent variable and a set of independent variables. Two of the possible methods for regression of the groups with the environmental variables are the stepwise regression and the all-subsets regression.

Stepwise regression, or stepwise MANOVA (Multiple Analysis of Variance), creates a data-directed search for the variables that best divide the groups and is separated in three types of approaches:

- Forward selection, where no variable is considered a predictor to start with, and they are tested individually for statistical relevance, being included if so;
- Backward selection, which does the inverse path of forward selection – considers every variable as being a relevant predictor, successively testing each one of them for relevance and deleting the ones not significant;
- The combination of both the previous methods, testing at each step the variables that are to be included or excluded [40, 41].

The decision of statistical relevance is given by cutoff limits defined by the creator.

Alternatively, the all-subsets regression method creates combinations of the possible predictors given by all possible models and tests them for statistical relevance. The best subsets of the environmental variables to predict the biological information result in the best possible models.

These analyses of variance can also be evaluated through the score of Wilk's lambda, a test of mean differences in Discriminant Analysis; the smaller the lambda for a variable, the more that variable contributes to the discriminant function. This measure indicates the group means on a combination of dependent variables [42, 43].

2.5.2. Null-model and Replicate Sampling Standard Deviation

A null model for a predictive model is used to determine the minimum precision of any model created for a given set of reference site-data, since its O/E standard deviation (SD) value is supposed to be worse than the one obtained by the predictive model, as it does not account for natural variability, i.e., the clustering created in the model and the posteriorly selected grouping is not considered in the null model.

The null-model is based on either the same set of sites for the model construction or a set of reference sites that were saved and put aside during the creation to evaluate model. The sites included in this evaluation are called validation sites.

The null-model E (expected value) is the average number of all found species above the limit defined by the model and that were observed at the references sites. The O (observed value) is the average number of all of the found species in the validation sites.

If the O/E values of the null model are closer to the O/E values of the model, it means that the model precision is low and more room exists for improvement. This limit is only achieved if a model fails to account for variability in taxon richness under reference conditions.

The O/E score of a predictive model is almost never equal to 1 even if the model has an exact prediction of E, since the Observed score is a sum of taxon occurrences (1 for present, 0 for absent), but the E score is a sum of probabilities. Moreover, if the same test site is evaluated again, the taxon occurrences will be slightly different (1s and 0s) for many taxa, as the taxon probabilities are randomly generated. Thus, collecting a large number of replicate samples of a reference site (the same sample assessed by the model repeatedly), the O/E values will vary around 1.0, despite the E score being predicted correctly.

Thereby, the Replicate Sampling Standard Deviation (RSSD) is the theoretical definition of the lower bound on the SD one should expect for any model and the maximum precision attainable. This score may be compared to the attained precision of the model through its standard deviation, and the distance amid these scores indicates how much improvement may be achieved.

Together, the null model and the replicate-sampling SDs estimate the minimum and maximum precision, respectively, reachable by any predictive model of a given set of reference values. Hence, the location of a model between these two measures defines the models attained precision and possibility of improvement [44].

2.5.3. Chi-square test and F-statistic

One of the statistics that may be applied to measure the goodness of fit of a model – i.e., if the observed values are significantly similar to the expected values or, also, if the expected values are a good fit for the observed values – is the chi-square test.

To measure the goodness of fit for a model, two values must be calculated: the chi-statistic of the model (observed and expected values) and the critical chi-square value.

The chi-statistic provides an approximately chi-squared distribution, which is a sum of squares of a group of independent standard normal random variables, based on the expected and observed values of the model.

The critical chi-square value is based on the degrees of freedom (the number of sites minus one) and the significance value, or *p* value. The significance value indicates the chance of getting an extreme value. Therefore, a *p* value will provide the value that has *p* chance of appearing. This value is the cutoff limit to whether the model is rejected. In the plot of the chi-square distribution, the *p* value provides the limit of acceptance. This probability, the *p* value, is typically 5%, 1% or 0.1%.

These two values are then compared, and if the chi-statistic value is higher than the critical chi-square, this means that the obtained value has less than a determined probability of appearing, based on the observed values.

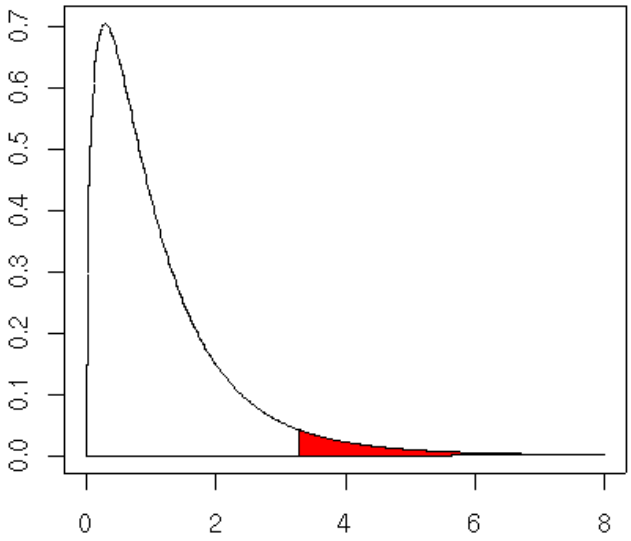


Figure 4 Chi-square distribution p value set at the beginning of the red region (0.5%)

Concluding, for a set of expected values to be a good fit for the set of observed values, the chi-statistic of the model must be lower than desired the critical chi-square value [45].

Another statistic, the F-statistic, measures the variation of data between groups of data points (for the purpose of this dissertation, this regards the groups of sites defined by the model

creation) and the variation of data within group, by calculating the chi-square distributions of both of them. The higher this F-statistic is, the more the variation between groups is significant compared to the variation within group. Knowing the F-statistic value and setting a critical value for rejection, the *p* value, or the significance value, may be calculated.

This F-statistic requires a critical value for rejection of the validity of the model, similarly to the chi-square test. The comparison of the actual F-statistic value and the value for the critical value indicates if the model is significant, according to this particular statistic [46].

2.5.4. Performance assessment

The predictive modeling approaches have inherent classification accuracies, and misclassification is a problem that must be taken into account when implementing such techniques. In this section, two of the existing performance classification methods are described.

2.5.4.1.Resubstitution estimation

Resubstitution error is the error rate on the training set of a predictive model, computed by estimating the error of the predicted values of a trained model against the observed values in the training set. In this approach, the same dataset is used to build the classifier and assess its performance. In other words, the classifier is trained using the complete learning set and then applies the selected classifiers to each observation. Finally, it annotates the number of incorrect predicted class labels based on the comparison amid the predicted and actual classifications [47].

2.5.4.2.Cross-Validation Error Estimation

Cross-Validation (CV) is a statistical method of evaluating and comparing data. This method has the goal of estimating the performance of the model created from the existent data, defining the generalizability of an algorithm or to compare the performance between multiple variants of a parameterized model.

Cross-validating a model involves dividing it into segments: the training set and the validation set. The sets defined for training are utilized to provide the classifiers (or discriminant variables) and the validation sets are computed to define the error rates. This is a method of

predicting the model to a hypothetical validation set when an explicit validation set is not provided (Figure 5).

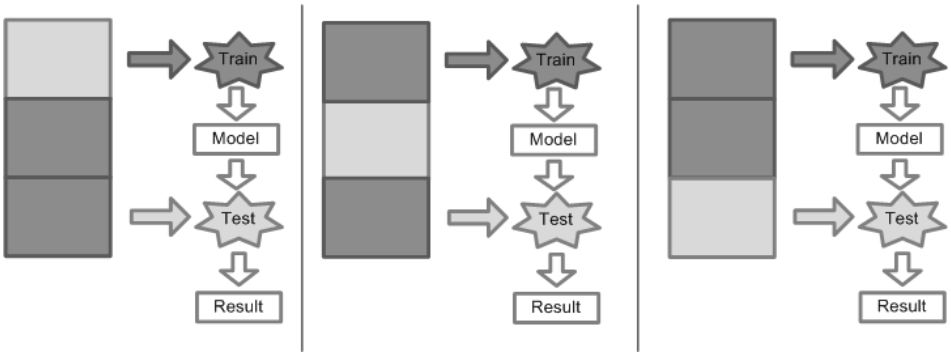


Figure 5 Cross-validation example

A typically used form of CV is leave-one-out cross-validation (LOOCV). This version of cross-validation uses a single observation of the entire set for validation, whereas the remaining observations are the training data. This is repeated for every sample of the set. This is the most common method of CV used by predictive models for bioassessment.

Other types of cross-validation include K-fold, where the original set is divided in K sub groups of samples and each of the groups is used once as validation and the remaining are used as training data; and repeated random sub-sampling validation, in which the training and validation sets are selected and divided randomly.

Cross-validation is widely used in bioinformatics especially for small datasets [47, 48].

2.6.RIO – an online platform for predictive modeling

The present project implementation is based and built from a previous work, named RIO. It was a website that allowed the creation and application of predictive models based on RIVPACS or BEAST. The service was created for the assessment of rivers ecological quality based on their communities (macroinvertebrates, algae, fish and plants). It allowed the access to predictive models to authorized users. It also allowed the creation of new predictive models for a specific region. The supported languages were Portuguese and English.

2.6.1. Existing features

2.6.1.1.Model creation

The first step for creation of these predictive models was the selection of the type of model (ellipses – BEAST – or observed/expected values – RIVPACS), the biological and the environmental tables in CSV (Comma Separated Values) extension. These tables would have to be totally verified for possible errors like empty cells or incorrect data types (since the only feedback for incorrect tables was a failure notification and a request to reintroduce the tables with correct information). The environmental table would have to include a column indicating the association of each reference site to its respective group, which had to be assigned in a previous clustering analysis external to the web site. Then, the user would have to identify which columns represented sites, groups and variables.

After this, the model would be created by an uninterrupted sequence of analyses: forward stepwise multiple discriminant analysis for predictors selection; complete discriminant analysis for calculation of probabilities of group membership for reference sites; calculation of taxa frequencies in each group; and calculation of O, E and O/E values and classification system. The final results were provided visually in the web page and in a compressed file, including, for both RIVPACS and BEAST, the observed/expected values, the bands limits, frequencies of taxa and group probabilities for each site and the original tables that the user provided in CSV. Also, the histogram plot of the O/E50 frequencies (Observed/Expected values for all taxa occurring in over 50% of the sites) was provided.

Finally, the user was prompted to fill model information, such as its name and description (English and Portuguese), sampling methods, the transformations made to the discriminant variables, the geographical location of the evaluated sites in the map (this step was optional) and the indication of whether the model should be listed. This latter had to be approved by an administrator.

2.6.1.2.Assessments of test sites

For the assessment of test sites through the existent predictive models, the user first had to select the type of model (observed/expected or ellipses – in other words, RIVPACS or BEAST), the model (from the list of previously created and listed models) and to provide the biological and environmental tables of test sites. The environmental table should contain all discriminant variables defined in the elected model. Next, the user indicated which columns should be taken into consideration in the analysis and which column referred to the sites names.

After this, the analysis was performed by an uninterrupted sequence of calculations and the provided results were, for the observed/expected type of analysis (or RIVPACS type):

- The biological and environmental tables provided by the user;
- The discriminant variables of the model;
- The definition of the bands;
- The frequencies and group probabilities of taxa occurrence;
- A table with the O/E values;
- Tables with the assigned bands of quality.

For the ellipses (or BEAST type), the results were:

- Three plots for each test site, with the reference group of sites and the probability ellipses (because normally three axes of the MDS ordination are used, each plot would provide the combinations of the three first axes, two by two);
- A table with the discriminant variables;
- The biological and environmental tables initially provided by the user.

For both types of analyses, there was a possibility of indicating and saving the location of the sites in a map and the marker would have the color corresponding to the quality class attributed (according to the European color-code system).

2.6.1.3.Other features

Apart from these main features, the site also provided authentication, registration, personal information, two languages, model visualization, taxonomic listing (for the taxa inserted in the database of the site), tutorials, general information (contact us page, about the site page) and frequently asked questions (FAQ). For users with administrative privileges, there was an administration section, with pages for:

- Editing the introduction;
- Editing and activating users;
- Editing and approving models;
- Adding and editing taxa;
- Creating taxonomic groups;
- Adding and editing FAQs;
- Adding and editing tutorials;
- Editing the administrator public information, such as e-mail and address.

2.6.1.4.Technologies

The RIO website was built in PHP and HTML/JavaScript for the pages and processing, used SYSTAT, a statistical software package, for the construction of probability ellipses, Python for the graphics, MySQL for the database, Smarty for page templating and a few libraries, such as JQuery and TinyMCE.

Regarding security, considerations were taken about the performance of the predictive modeling computations and for security with field validations for SQL operations. The maps and the identification of the location of sites were integrated using Google Maps.

The classes that were implemented were:

- *Biblioteca*, for the connection to the database, starting the main template, checking user authentication and the starting PHP session;
- *TrataFicheiros*, for file manipulation;
- *FazCalculos*, for the analysis calculations calls.

The database supported the management of predictive models, users, help and information in two languages (Portuguese and English). A Class Diagram of the database is illustrated in Figure 6.

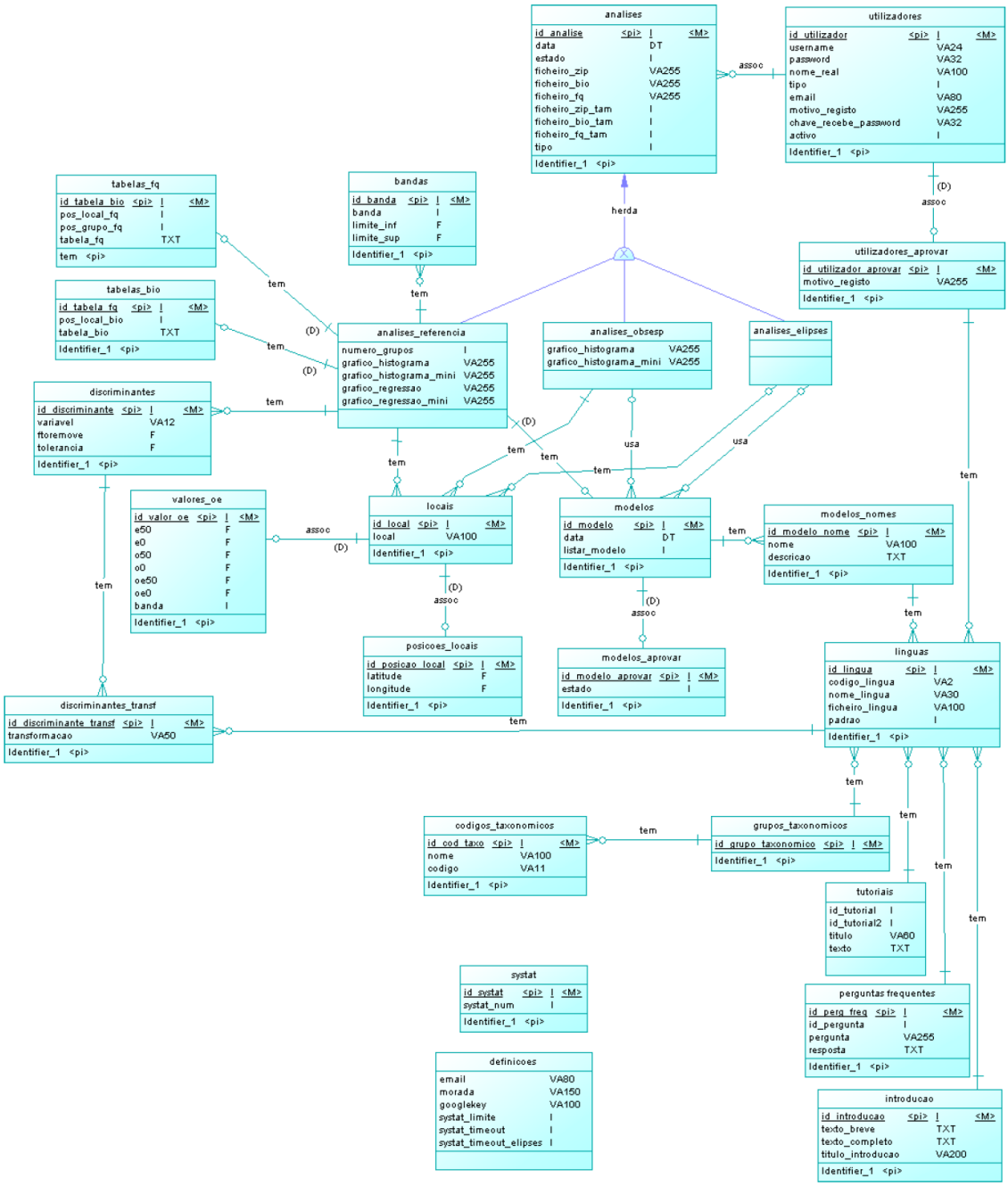


Figure 6 RIO Database model

2.6.2. Known problems

This first version of RIO had issues with the integration and range of its features, namely:

- The bands system (statistic for division of the gradient into quality classes) implemented in RIO needs to be updated to the currently applied methods for establishing according to the European directive;
- The taxa were represented by their codes, which would make less straightforward the interpretation of results, even though useful in preventing typing errors;
- The data input process was inflexible, since the only type of file allowed as input on the system was CSV. Considering that there are other common types of files for the tables, such as XLS (Microsoft Excel Spreadsheet) or ODS (Open Document Spreadsheet), this forced the user to open the file in another program, such as Microsoft Excel, and convert the file to the supported type before providing it to the website;
- There was a need to perform previous data transformation on both abiotic (for normalization) and biotic (for diminishing the weight of very frequent taxa or to reduce the information to presence absence) outside of the web site;
- The clustering step was not supported, so it had to be made in an external software and the grouping chosen previously had to be indicated in the biological table as a column;
- The amount of information describing models requirements was poor, because the website did not support fields to introduce specific information on the models;
- Recent developments in predictive models provide more robust and liable models and tests to the methods than those implemented.

These issues compromised the usability and manageability of RIO, therefore motivating the creation of an improved and updated version of that project.

3.AQUAWEB - Framework Proposal

AQUAWEB is a web platform created for the development and use of freshwater predictive modeling based on aquatic faunal communities as well as on environmental features. The users will also be allowed to manage the data with flexibility, making use of a broad range of choices supported by a set of informative data along the way.

As expressed, AQUAWEB is an improved and more complete version of a previous work, i.e., RIO. For this reason, a few of RIO's more basic characteristics were maintained, such as the general page organizational logic and the backend management. The purpose of this work is to update and enhance the existing features and broaden the number of incorporated properties.

The main initial objectives of AQUAWEB were:

- The creation of projects by users for analysis and data storage;
- Providing information to the users about the available models, their performances and the requirements of the type of model used (BEAST and RIVPACS);
- Allowing the input of data in diverse formats (other than CSV);
- Displaying model assessments with detail and providing the results for download;
- Displaying in maps the location of the reference and test sites with their respective assigned class color code, as defined by the WFD;
- Providing an electronic taxonomic key identification tool for an easier identification of fauna and flora;
- Allowing the future modification of most of the information, by an administrator.

3.1.Systems and technologies

3.1.1. Application server

The server hosting this web site is running an integrated application called WampServer, a Windows web development environment. This was the tool used to provide server-side centralized management of the web site due to its simplicity, centrality, being free software, having a constant record of every occurrence or failure and for having high reliability. It allows the creation of web applications using the following technologies:

- Apache, a free web server currently being used in many worldwide web sites;

- PHP, a server-side programming language for general purposes on web development;
- MySQL, a relational database management system running as a server and that provides multi-user access to multiple databases.

Furthermore, it integrates a web-based database manager named phpmyadmin.

3.1.2. Statistical computing language

As expressed, RIO integrated a commercial package to handle statistical computing needs. Due to integration facility and development flexibility, the R programming language was selected for statistical computing and graphics. This technology is a very important piece of this dissertation, for it is used to support the full implementation of the water quality assessment algorithms, including the creation of graphical results and tables for the predictive models, the main focus of the web site. Thereby, the choice of this programming language is based on its balance of advantages and disadvantages.

Several advantages can be identified:

- The broad variety of extension packages for all kinds of purposes;
- The quality, availability and completeness of the documentation for either the basic R packages, the normal functioning of the tool and the available additional external packages;
- The developer community is massive and helpful;
- This software is completely free;
- The performance is reported to be high, as it was the initial intention (since it was developed based on two other statistical languages – S and Scheme – to be an improved and faster version of them);
- The language syntax was created to be similar to other well known statistical software such as Matlab, S and SPSS;
- It is multi-platform, being supported in Windows, Unix and MacOS;
- It has state-of-the-art graphics capabilities.

As for the disadvantages of this piece of software:

- The R processor is not multithreaded, so if two scripts are requested, it processes the first request completely before dealing with the second;
- Since all of the variables and data are stored in memory, there is a need to wittingly save each of the desired variables in a specific type of file extension (.Rdata) for later use;
- Packages that are not loaded as a part of the basic library of R have to be downloaded and put directly in the libraries folder. Otherwise, at the end of an R session, they will

be deleted, which means that in the next session they would have to be downloaded once more [49].

The application server had to be configured to allow the loading of web pages with extended execution time, in order to be able to wait for long necessary processing cycles with the R scripts (this was made in the PHP configuration file, php.ini, in the option *max_execution_time*).

3.1.3. Other technologies

JavaScript is utilized as the scripting language on event handling of the PHP web pages. Some libraries were used throughout most pages of the web site, such as Smarty, a web template system, to separate the presentation of the web page from the back-end management and JQuery, a cross-browser JavaScript library created to facilitate the client-side scripting.

The PHPExcel library [50] was used for management of tables, since it provides a set of PHP classes that allow reading and writing from different formats of spreadsheets such as Excel and CSV.

The spreadsheet style interface is provided through the jQuery.sheet library [51]. This interface was tested in the Windows and MacOS Operating Systems on the following browsers: Internet Explorer, Mozilla Firefox and Google Chrome. The advised browser is Mozilla Firefox, since the remaining web browsers performed poorly in the table management pages, taking several minutes to completely load and also being slow after loading.

The digital taxonomic key for Portuguese macroinvertebrates was built at the Institute of Marine Research with Lucid Phoenix software developed by Lucidcentral.org at the Centre of Biological Information Technology, University of Queensland, Australia. It is a piece of software intended for dichotomous or pathway key builder and player to make identification keys in an interactive way with the help of images and videos. Taxonomic keys are a series of paired statements that describe the physical characteristics of different organisms.

Finally, the IDE (Integrated Development Environment) used in the development of the website was Netbeans 7.0. The project was implemented for the most time in a virtual machine installed in a web server running Windows XP with WampServer and R installed. The server installations and configurations were performed via Remote Connection on Windows.

3.2.System Architecture

In summary, the technology provided by this system has a working logic where, in the client side, the user uses the browser to access the web site with HTML and to communicate with the server through JavaScript, which is received in the server side by Apache. The Apache server provides the HTML pages based in the programming language PHP with integrated functionalities: Smarty for dividing the business logic of the pages from the presentation layer, PHPExcel to manage upload, download and conversion of spreadsheet file types, JQuery libraries for simplifying actions such as presenting a spreadsheet style interface within the web site and Lucid Phoenix as an embedded player to manage taxonomic keys. Moreover, PHP communicates with the database of the project through MySQL and executes R scripts for the predictive modeling calculations (Figure 7).

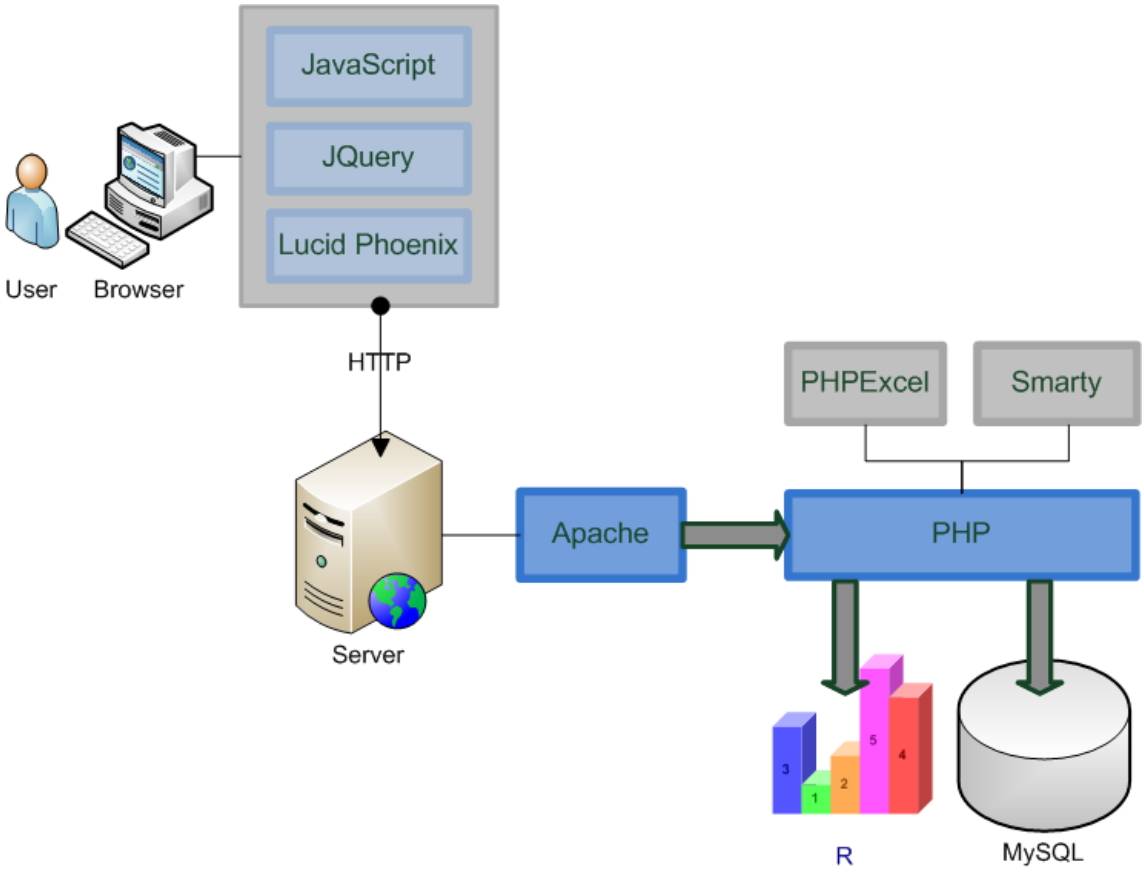


Figure 7 System architecture

3.2.1. Supported predictive modeling approaches

The predictive modeling tools implemented in this project were based on RIVPACS and BEAST models.

The starting point for the implementation was the collection of R-language scripts for RIVPACS predictive model created by John Van Sickle, from the US Environmental Protection Agency [52]. These scripts provided model creating and model testing, with fully documented code and several options in many of the steps. As the author suggests, users need to modify the scripts in order to fit their particular data sets. From this basis, both the model building and the model assessing scripts were split in smaller scripts, to be integrated in the website in a step-by-step logic with intermediate inputs and outputs. Several modifications were also made in the scripts, such as error correction, error prevention and improvement of the results.

The BEAST-type predictive model implementation had most of the steps equal to RIVPACS, thus, up until the last script of the model construction, a decision about the type of model to create is not necessary.

3.2.2. Class model

The requirements listed in the beginning of this chapter were tackled with the help of several PHP classes that created a more modular management (Figure 8).

The management of sessions and page templating is the responsibility of class *BibliotecaAdmin* or *Biblioteca* for administrators or other users, respectively.

ValidaTabelas is the class in charge of the correction and evaluation of tables, as well as the presentation of errors in these tables.

The class *PclZip* was used for handling compressed files, from gathering all the results of predictive models to uploading new versions of taxonomic keys from Lucid Phoenix.

The templating system is managed through a plug-in named Smarty, through which were created the classes *Template* and *TemplateAdmin* to create generalized methods for both the regular and the administration page layouts.

Some of the enunciated classes are based in RIO, but improvements were introduced and new functionalities were added to them in order to comprise all of the new features.

The class *TrataFicheiros* now supports all the features involving the new spreadsheet file types uploading and management, the embedded spreadsheet type interface.

The class *ValidaTabelas* now provides specific, detailed and explicit errors, not providing only the first error detected for the table to be invalid, but a list of all the detected.

The package *SuporteTabelas* includes the PHPExcel library.

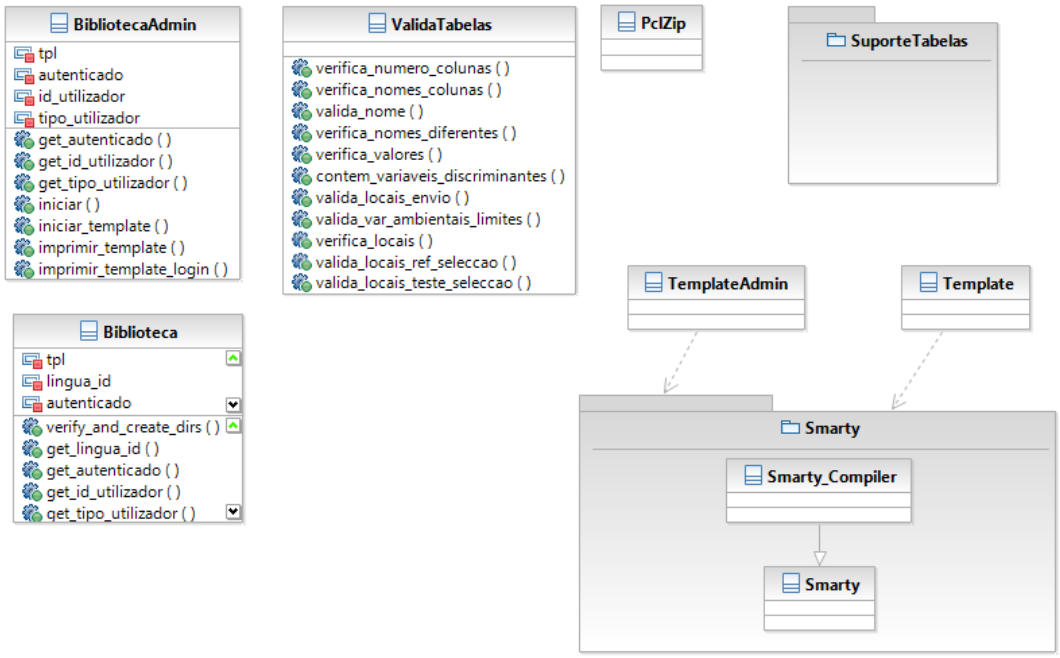


Figure 8 AQUAWEB Class Diagram

3.2.3. User actions

In a more detailed view of this system, a visitor of the web site without an account can view all helping sections, check the existent taxa in the database, alter the language of the web site, view the listed models and register in a new account, as illustrated in Figure 9.

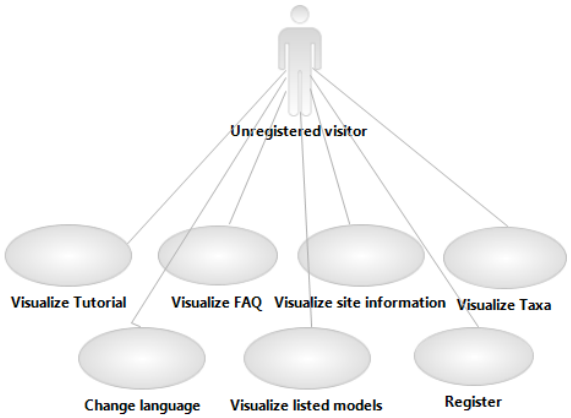


Figure 9 Use case model for an unregistered visitor

When a new account is registered, the user broadens the range of possible actions. Thus, the privileges of a regular user are now, also, using listed models, the global matrices and the taxonomic key, besides logging in and out of the system (Figure 10).

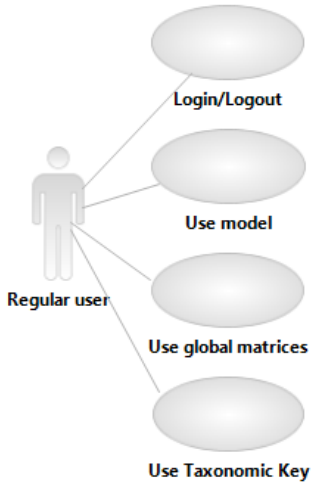


Figure 10 Use case model for a regular user

When privileged access is provided to a normal user, the range of possible actions is broadened to creation of predictive models (Figure 11).

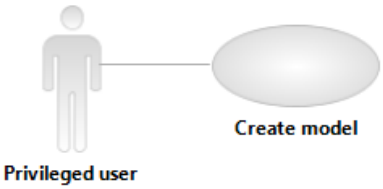


Figure 11 Use case model for a privileged user

Administrative privileges can also be granted to a user by an administrator. Apart from all the previously described actions, this type of user also has the possibility of editing the helping and information sections, adding and editing the allowed taxa for predictive models, managing

users, uploading new versions of the taxonomic key, editing the global matrices and edit and approve models (Figure 12).

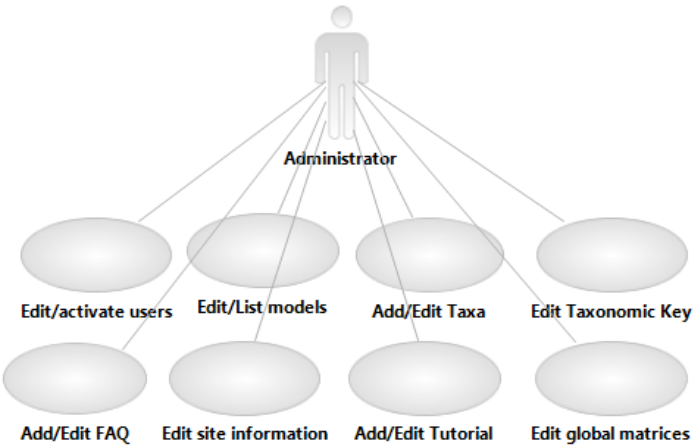


Figure 12 Use case model for an administrator

3.2.4. Database Model

The database of this project was almost entirely redesigned to fit the new features, as well as to provide more reliability, performance and scalability, as depicted in Figure 13. Appendix A provides both the database models of RIO and AQUAWEB for comparison.

The tables *analise*, *analisemodelo*, *variaveldiscriminante*, *local* and *classe* manage information directly related to the predictive models and assessments.

The table *taxon* has list of the allowed types of taxa for when a user provides a biological table.

The table *contacto* keeps general information about the administrator for any user to be able to contact them.

The tables *texto* and *tipotexto* contain the long textual fields in the web site like the introduction or tutorials, where *tipotexto* defines the type of text and *texto* defines the textual content. Every text has translated versions in Portuguese and in English; the definition of the available languages is in the table *lingua*.

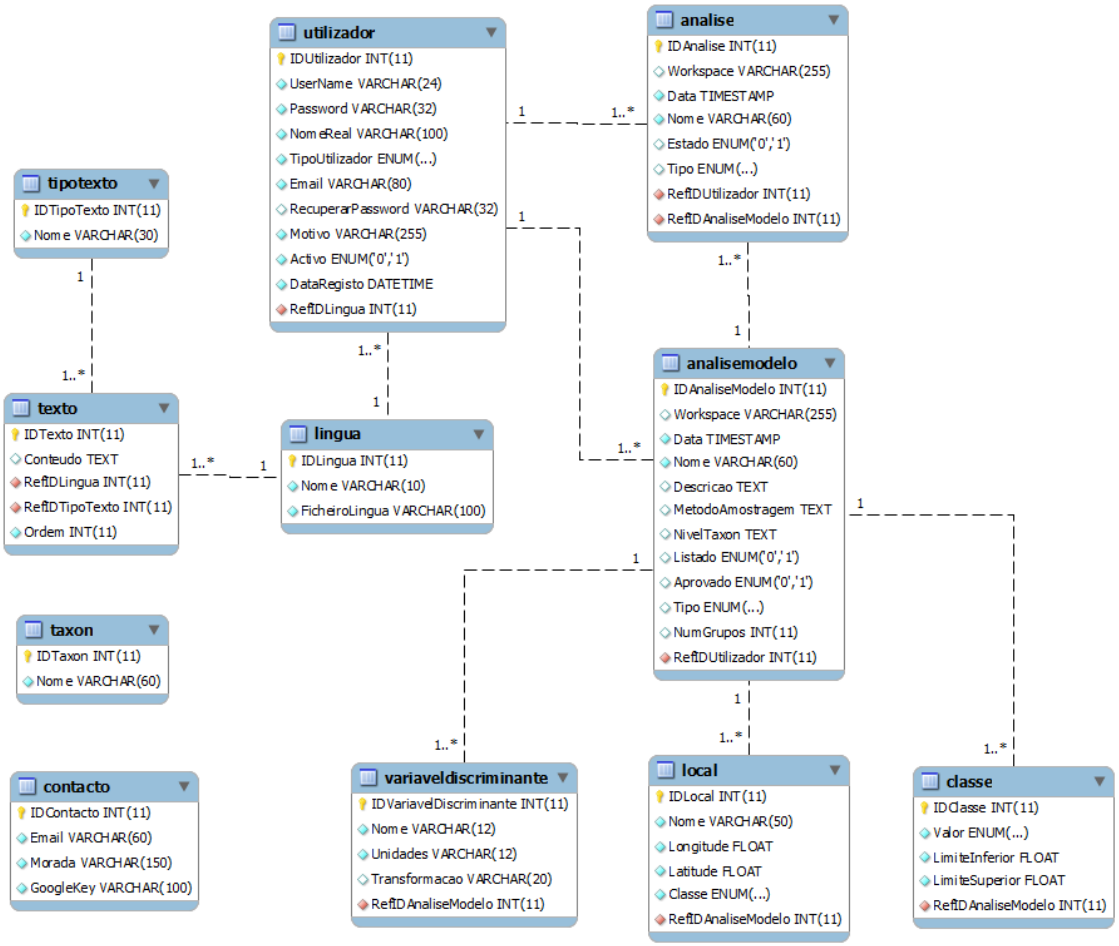


Figure 13 Internal database class diagram

4. Platform Implementation

In this chapter, the implementation of all the requirements listed in the previous chapter is described. First, the AQUAWEB graphic user interface is shown, followed by the description of new functionalities implemented, as well as the introduction of new capabilities that offer the user a more holistic tool with an integrated support in every step of the way.

4.1. Graphic Interface

The general presentation of AQUAWEB is illustrated in Figure 14. The language of the web site can be switched between Portuguese and English in the button upper right of the site. Next, the tabs in the top of the page provide the main functionalities of the web site: *List Models*, *Use Models*, *Create Models*, *Taxonomic Key*, *List Taxa* and *Info*.

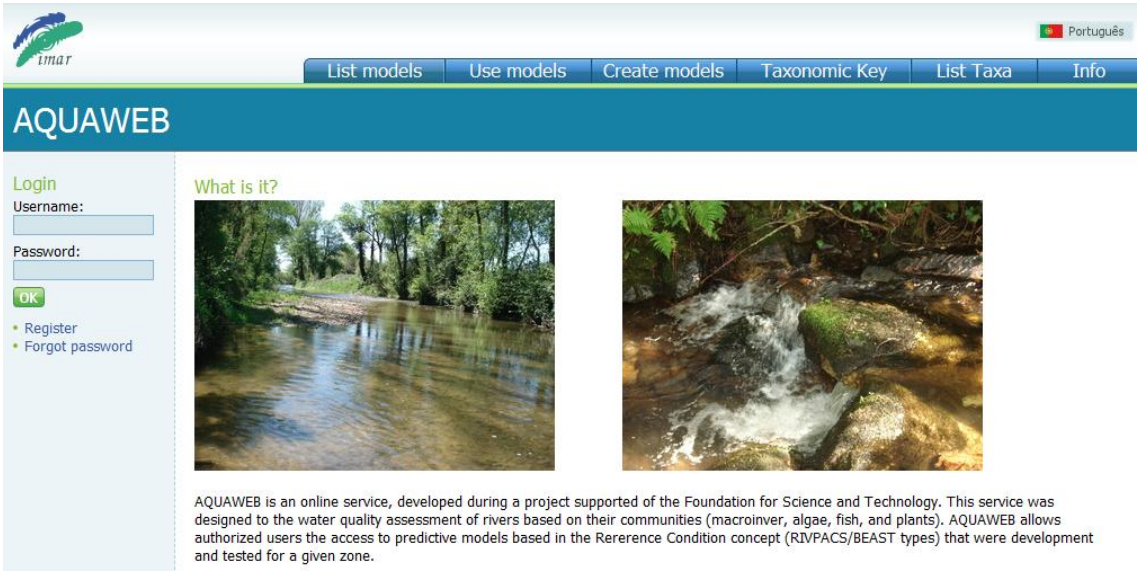


Figure 14 General appearance of AQUAWEB

The left area of the web site is reserved for login/logout and for personal management of the user. If a logged in user does not own privileged access, he main only perform the basic tasks: view his previous analysis and manage his profile (Figure 15).

Figure 15 Left area of the web site for regular users

Moreover, if the user has privileges, the global matrices management is also available (Figure 16).

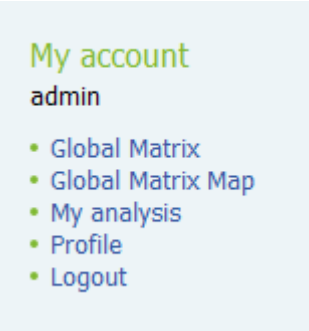


Figure 16 Left area of the web site for privileged users

The specific content for each different page, and the only section of the page that varies according to the present page, is the white central area of the web site. An example of the alteration is in Figure 17, which may be compared to Figure 14.

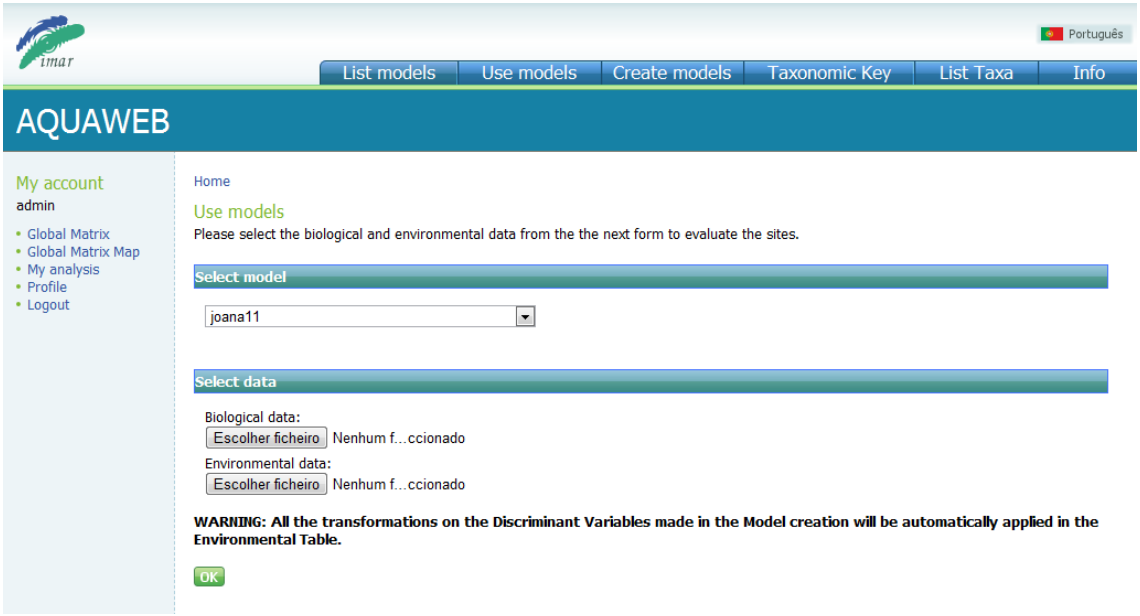


Figure 17 Example of the Use Models initial step page

The administrative section of the web site, allowed only to users with administrative privileges, is only provided in Portuguese and has the appearance depicted in Figure 18.

Figure 18 Administrative section appearance

4.2.AQUAWEB improvements

In this section, the development in features that were present in the previous work but needed improvements or extensions to the RIO web site of their capabilities is presented.

4.2.1. Spreadsheet file type

One of the goals of this project was broadening of the number of environmental and biological tables file types accepted as input. Therefore, the support for Microsoft Excel extensions (XLS and XLSX), besides the already supported simpler type, CSV (comma separated values), was implemented. Also, the initially provided tables are later returned in all of the mentioned file extensions. This was achieved with the help of the spreadsheet engine PHPExcel. Now, the user may upload the desired file type indiscriminately. Moreover, the user may download the spreadsheet tables in any of the three supported types.

4.2.2. Incorporation of precedent model steps

In the previous RIO platform, the first steps for the creation of predictive models were not supported. Hence, the user would have to edit the environmental and biological tables outside of the web site (usually in Excel) in order to make sure that there were not errors, such as empty cells or misspelled taxa names. If there were errors in the tables, the only feedback was a general and unspecified error, like indicating that there were taxa not allowed. Moreover, it was necessary to make the clustering of the sites into groups, providing these groups as an additional column in the biological table.

In the AQUAWEB implementation, the tables are graphically displayed and can be edited. The errors are clearly marked in the cells/columns/rows that have problems, with color-based types of errors (this will be further demonstrated in section 4.2.5 (Step by step algorithm).

After correcting the tables, there is a new step, which is applying transformations to environmental variables to normalize the data distribution. Also for biological data, transformations may be applied to downweight the importance of highly abundant taxa or upweight the importance of less abundant taxa.

When all the necessary transformations are made, another new step is supported: the clustering of the sites into groups, based on the biological information, based on Bray-Curtis coefficient.

4.2.3. Information on models requirements

The new platform AQUAWEB provides new fields for the user to create descriptions of the predictive model, in order to improve the information provided to a potential user of the models created. The fields available now are the model description (already existent in RIO), the community sampling methods and the taxonomic level used. This information is saved in the database, in new fields created especially for this purpose. A demonstration of this is provided in the first step of model creation in section 4.2.5 (Step by step algorithm).

4.2.4. Database

The database implemented in our project (Figure 13 in section 3.2.4) was an improvement of the previous RIO design, which had issues on the matters of scalability, reliability and performance. The number of supported tables was reduced from 25 to 11.

Previously, in RIO, each type of text in different pages had its own table, like the introduction or frequently asked questions. This made it easier to query the database but did not provide the generalization needed for possible future extensions, because new tables would have to be created. So, all types of textual descriptions found in the web site now have their content in the same table *Texto*. To identify which type of text it is, there is a connection to a table called *TipoTexto* with unique identifiers for each of the fields, such as the introduction title, or the FAQ (frequently asked question) answer.

All direct features of models are now grouped in the same table *Modelo*, whereas previously there were distinct tables for not approved models, for model descriptions, for the permissions and for the location of the files resulting from the model creation. The results and descriptions regarding assessments in test sites (using models) are all recorded in one single table, *AnaliseModelo*.

Regarding discriminant variables, the number of tables was also reduced to one, *VariavelDiscriminante*, whereas before there were two tables: one for the enumeration of these and another to identify the transformations applied to these, such as a logarithm or a squared root.

The tables that saved the content of the biological and environmental tables became obsolete, since the present table *Modelo* provides a field with the path of a folder with all the results of the analysis.

The distinction of type of predictive modeling approach in assessment went from being defined by different tables (*analises_referencia*, *analises_elipses*, *analises_obsesp*) to a single field in the generalized tables *Modelo* and *AnaliseModelo,* for model construction and test sites assessment respectively. Similarly to models, a field in this table points to the path of a folder with the results of the execution of the algorithm.

4.2.5. Step by step algorithm

One of the main concerns of this project was providing results of models and assessments that would be more comprehensible by the user, such as spreadsheets and graphical representations. RIO only provided a small generalized set of results. Therefore, along with

more detailed results, more control over the steps taken by the algorithms was required in order to make any desired adjustments to fit the requirements of the user.

The creation of predictive models in this project is implemented through several steps provided to the user as successive web pages. Each of these steps requires the user to provide additional information about how to proceed to the next one. In each of these steps, a set of graphical and textual results are shown to help the user make decisions. The summary of these steps is available in Appendix B and the description of the implementation of the steps in R scripts is documented in C.

The following contents demonstrate the series of steps that are to be taken in order to create and use models.

4.2.5.1.Model creation

The first step for creating a predictive model requires as input (Figure 19)[2]:

- The biological table;
- The environmental table;
- The transformation that the user wants to apply to the entire environmental table (none, squared root, fourth root, logarithm or presence/absence);
- The model name;
- The model description;
- The community sampling methods;
- The taxonomic level used.

[2] The file uploading and submitting buttons are displayed according to the web browser definitions and language, disregarding the language in which the web site is presented.

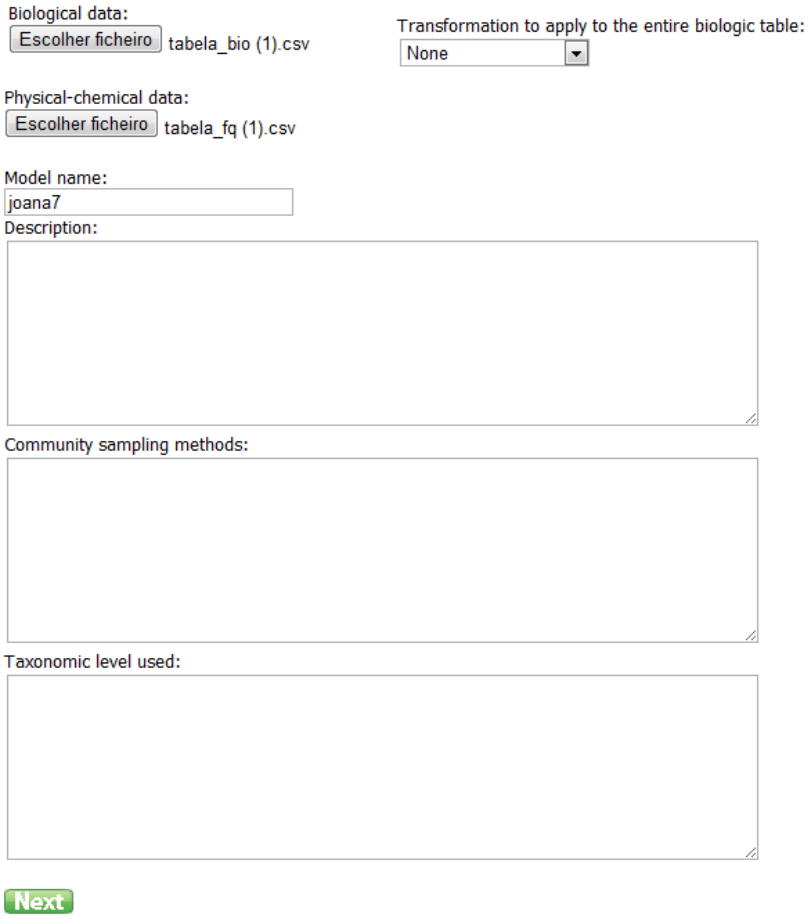


Figure 19 First step for creating models

By going to the next step, pressing “Next”, the user is prompted with the two editable tables (biological and environmental) and presents the detected errors that must be corrected (Figure 20). For each table, the user must identify which columns represent the sites (in dropdown lists), as well as the latitude and longitude columns, so that these last two may be excluded from the candidate predictors.

If there are still errors when the user presses “Next”, the same page will reload with any remaining or introduced errors. The error associated with invalid taxa is only displayed when the user chooses what column represents the sites and presses “Next”. After this, the evaluation of the listed taxa can be checked in the database, and if any of them are not in it, the error will be presented.

The types of errors are identified in the tables by their color and the background of the cell is painted in the same color. The page has a description of the available errors, as shown in Figure 20 and Figure 21.

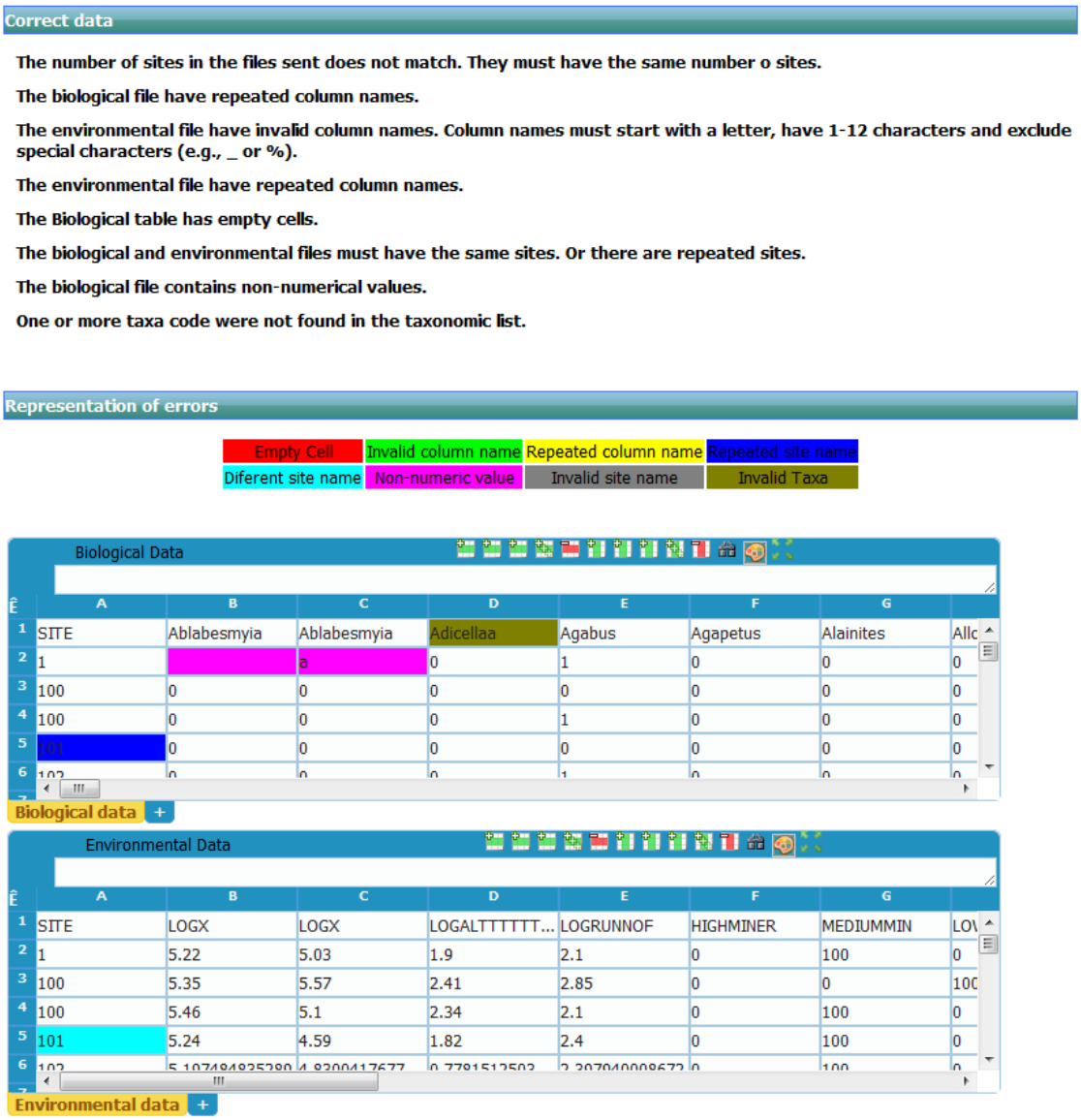


Figure 20 The interface for editing tables in the second step of model creation

Figure 21 The columns definition for the tables in the second step of model creation

After correcting all errors and pressing “Next”, the transformations page is shown. At this point, the tables are saved in CSV files and an R script is executed, performing the following tasks:

- The previously requested transformation (in the first step) is applied to the entire biological table. The table edition occurs prior to this transformation to assure that the transformation is made to well-formed data;
- The latitude and longitude columns are placed on a separate table;
- A histogram is created to show data distribution for each environmental variable.

The following model creation page (Figure 22) provides download of both tables in CSV, XLS and XLSX types. It also shows all of the environmental variables data distribution in histograms so that transformations can be chosen and applied to them (none, logarithm, squared root and 1/x), as illustrated in Figure 23. After selecting the desired transformation, a histogram of the variable with this new distribution appears in the last column (Figure 24). Both the original and the transformed histograms may be seen in full size by pressing the preview thumbnail. Every time a transformation is requested, an R script creates the new histogram, always saving both the original and the transformed environmental table. On a side note, all of the transformations performed in this step are recorded and, in model using, they are applied automatically to the predictor variables in the test tables.

At the bottom of the page, the user is asked to indicate a percentage of validation reference sites to be randomly removed from the model to not be considered for the model building, allowing only a minimum of two sites. A demonstration is depicted in Figure 25.

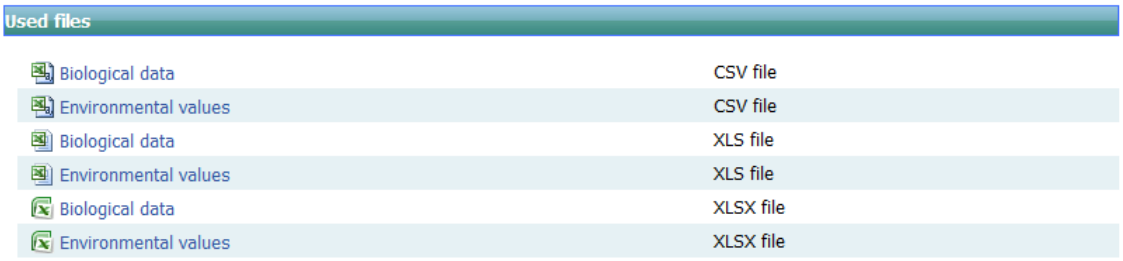


Figure 22 The top of the third step of model creation

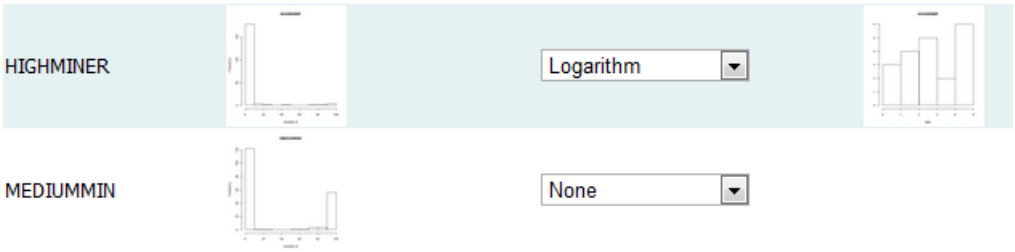


Figure 23 The appearance of the third step of model creation for histograms and transformations

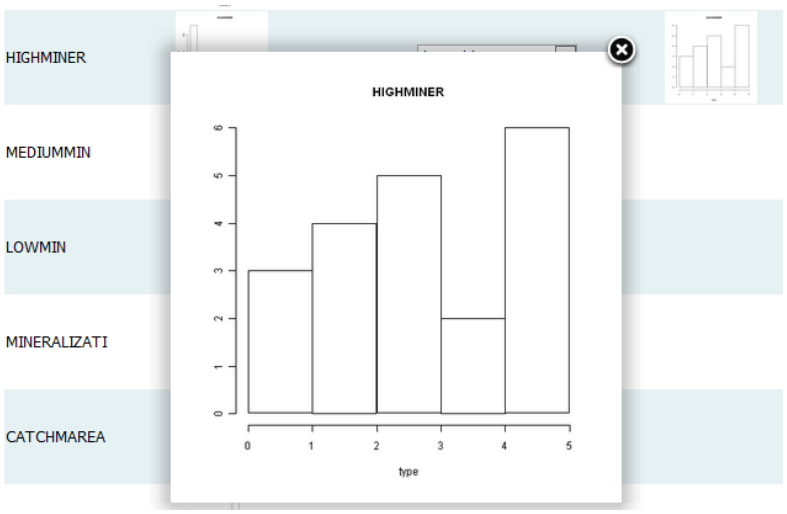


Figure 24 Visualization of a variable's histogram in full size

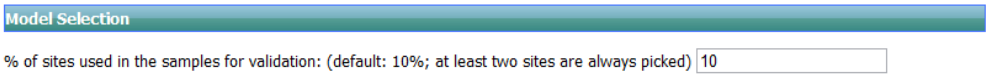


Figure 25 The definition of the percentage of validation sites

The next model creation step is the clustering (Figure 26), in which another R script is executed. The calibration sites are separated from the validation sites randomly from the elected percentage and the dendrogram only from the calibration sites is created from the Bray-Curtis similarities matrix.

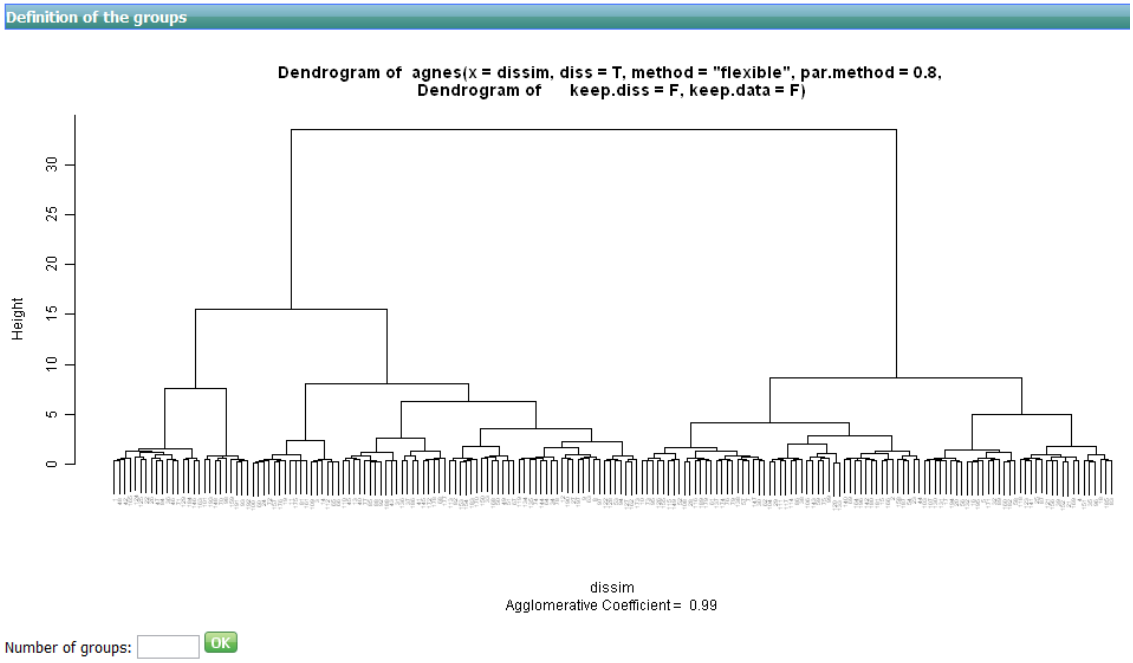


Figure 26 Dendrogram of the sites based in biological data

Then, the user indicates the desired number of groups, which are chosen by pruning the dendrogram at a level that satisfies the request (Figure 27). The result is provided from the execution of an R script created for that purpose. From this point, because only after the clustering the user is allowed to go to the next page, a new button to go to the next step appears at the bottom of the page (the "Next" button).

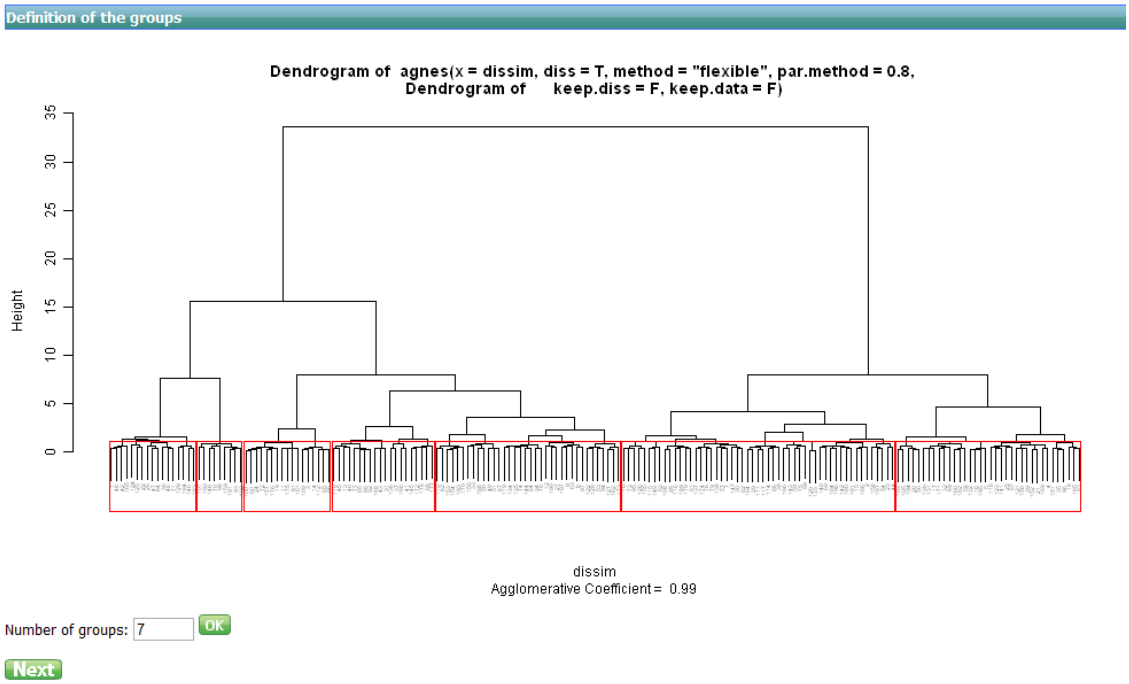


Figure 27 Dendrogram with the sites already divided into groups as indicated by the user

The next step in the predictive model creation is the selection of predictor variables. Hence, an R script is performed with the intention of finding the five models from each order of environmental variables up to 10 variables (which means that there will be shown 5 models of each amount of variables in a combination of variables starting with 2 variables and up to 10) plus a model with every environmental variable, that best describe the previous reference sites grouping based on the biological data[3].

At this step, and because the computing task is demanding, the page takes several minutes to present the results. While performing these computations, the page presents a "Please wait…"warning (Figure 28).

Figure 28 The waiting warning for the all-subsets regression step of model creation

[3] These specific values were requested by the researchers in this project because these would fit their needs and provide results in an acceptable time span, according to initial tests.

When the script, that calculates the all-subsets linear DFA regression method to determine possible predictive models, is completed, several results are presented. First, a table with the following information (Figure 29):

- The number of the model, for identification and selection;
- The order of the model (the number of discriminant variables for that model);
- The chi-square statistic and the F-statistic for the overall model;
- Wilk's lambda;
- Resubstitution and leave-one-out cross validation classification accuracies for calibration sites;
- Mean, standard deviation and root-mean-squared error for O/E at calibration sites;
- Replicate-sampling standard deviation of O/E at calibration sites;
- Mean, standard deviation and root-mean-squared error for O/E at validation sites;
- Statistics of the BC measure for calibration (.cal) and validation (.vld) sites. "MD" denotes the median and "90" the 90th percentile. BC measures the Bray-Curtis dissimilarity between observed and expected assemblages;
- The predictors in the model.

model	num	o	Fstat	Wilks	resub	cv	MNOEcal	SDOEcal	RMSEcal	SDRScal	MNOEvld	SDOEvld	RMSEvld	BCMDcal	BC90cal	BCMDvld	BC90vld	model
1	1		31.985	0.485	39.362	39.362	0.971	0.243	0.244	0.131	0.648	0.804	0.669	0.213	0.347	0.519	0.815	HYDROLOGICAL
2	1		31.871	0.486	40.957	40.426	0.992	0.241	0.24	0.138	0.61	0.73	0.647	0.204	0.343	0.515	0.79	RAIZQPRECIP
3	1		29.812	0.503	40.957	40.957	1.002	0.239	0.239	0.139	0.724	0.783	0.619	0.202	0.33	0.477	0.717	MEDIUMMIN
4	1		28.464	0.515	43.617	42.021	1.001	0.242	0.241	0.139	0.616	0.738	0.648	0.205	0.33	0.516	0.793	LOGRUNNOF
5	1		26.835	0.529	43.617	43.617	1.012	0.241	0.241	0.14	0.709	0.737	0.597	0.206	0.332	0.474	0.725	V1LOGCONDUC
7	2		23.684	0.312	51.064	51.064	0.972	0.248	0.249	0.129	0.645	0.796	0.665	0.212	0.35	0.512	0.805	HYDROLOGICAL LOGRUNNOF
8	2		23.223	0.318	50.532	50.532	0.981	0.246	0.246	0.132	0.632	0.789	0.668	0.216	0.327	0.514	0.814	HYDROLOGICAL MEDIUMMIN
9	2		23.111	0.319	43.617	43.085	0.998	0.244	0.244	0.137	0.717	0.774	0.616	0.21	0.346	0.474	0.715	MEDIUMMIN RAIZQPRECIP
10	2		21.465	0.34	45.213	45.213	1.003	0.242	0.241	0.138	0.722	0.782	0.619	0.212	0.327	0.477	0.717	LOGRUNNOF MEDIUMMIN
11	3		20.948	0.207	52.128	51.596	0.978	0.247	0.247	0.129	0.645	0.812	0.675	0.211	0.331	0.52	0.818	HYDROLOGICAL MEDIUMMIN RAIZQPRECIP COEFFVARIA

Figure 29 Example of a portion of the results of models selected by the algorithm

Below the table, a graphical representation of the root-mean-squared error of the O/E against the model order is presented. Once the user indicates a model number in the text box and presses the "OK" button below the graphic, the calibration and validation models are selected in the graphic (as demonstrated in Figure 30, where model number 20, of order 4, is chosen).

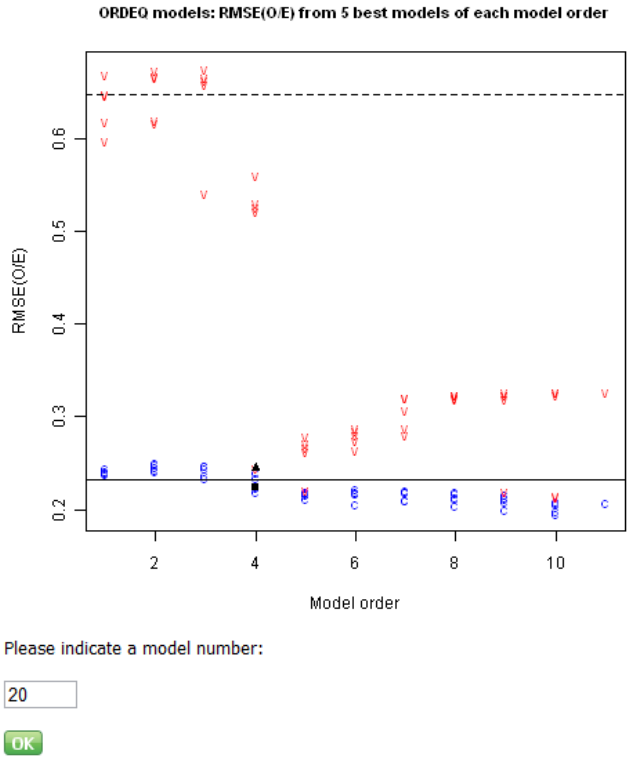


Figure 30 The root mean squared error of each model against its order

The following results of this step are the processing details and available results from the all-subsets regression and a list of all the possible predictors ordered by the percentage of best models that include each of the predictors (Figure 31).

```
[1]  New run of all-subsets DFA
[1] Count of calibration sites in each group
grps
 1  2  3  4  5  6  7
17 53 23  9 37 20 29
[1] 121669  models being screened. Please wait ...
[1] Model screening complete. Computing O/E and BC stats for subset of best models....
[1] Calibration: Mean, SD, RMSE of Null O/E =   1  0.233  0.232
[1] Calibration: Median, 90%ile of BC =   0.191  0.305
[1] Validation: Mean, SD, RMSE of Null O/E =   0.607  0.726  0.647
[1] Validation: Median, 90%ile of BC =   0.521  0.799
[1] "All-subsets function finished."
                    user.self          sys.self           elapsed
"elapsed time = "    "990.22"            "0.53"          "991.75"
       user.child         sys.child
                NA                NA
$subset.stats
   order    F.stat      Wilks cls.crct.resub cls.crct.cv  MNOE.cal  SDOE.cal
1      1 31.985320 0.48536931       39.36170    39.36170 0.9714621 0.2432689
2      1 31.871258 0.48626170       40.95745    40.42553 0.9915932 0.2405048
```

```
[1] "PREDICTOR IMPORTANCE. Calculate the percentage of best models that include each of the
predictors. Percentage is not weighted by model quality;"

  CATCHMAREA   COEFFVARIA    HIGHMINER HYDROLOGICAL  LOGALCALINI       LOGALT
        51.0         64.7          7.8         88.2         27.5         11.8
   LOGDISTSO  LOGHARDNESS    LOGRUNNOF       LOWMIN    MEDIUMMIN  RAIZQPRECIP
        13.7          2.0         19.6          7.8         80.4         80.4
  SLOPERAIZQ   TEMPERMEAN  THERMICVARI  V1LOGCONDUC
        25.5         43.1         35.3          2.0
```

Figure 31 Textual information about the executed regression

Lastly, two-dimensional plots of the following three scores are presented: the model order, the RMSE of the calibration sites and the RMSE of the validation sites (Figure 32).

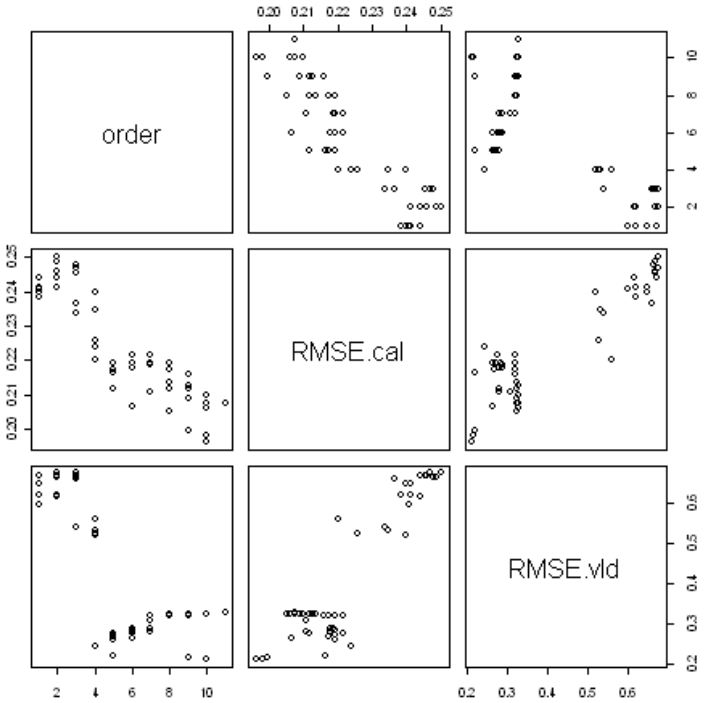


Figure 32 Graphical representation of results from the analysis

After a model is selected, access to the next step becomes available, through the button at the end of the page.

At the following step (Figure 33), the discriminant variables of the selected model are presented and the user is asked to indicate in which units these variables were collected. Also, the type of model must be selected.

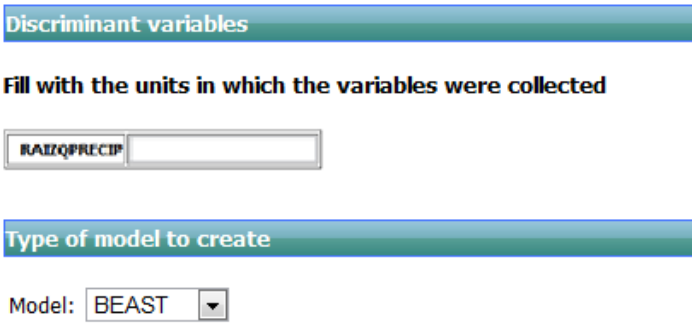


Figure 33 The penultimate step of model creation

The following and final page depends on the type of model selected, despite both models offering a ZIP file with all relevant results of the model creation.

If the selected model type is RIVPACS, several plots with different perspectives of the results and a table of the O/E values are presented in Figure 34 and Figure 35.

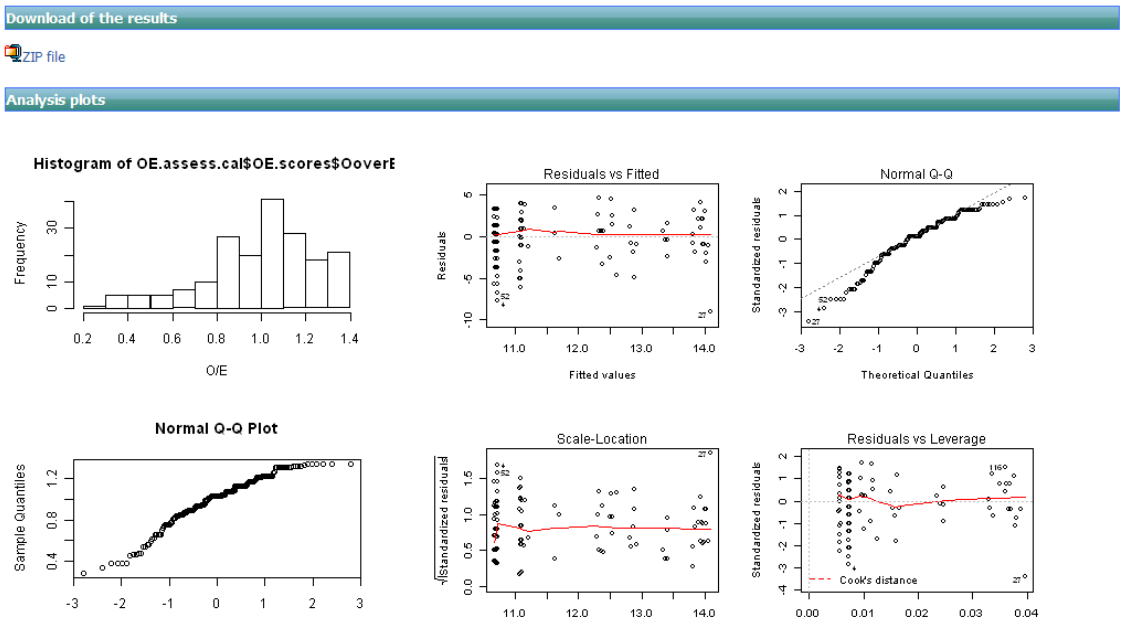
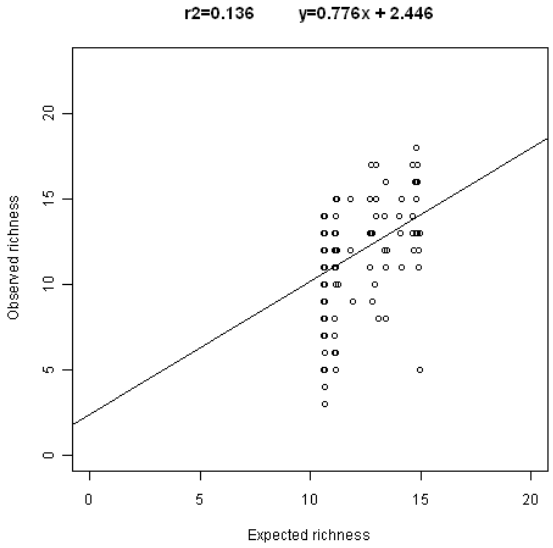


Figure 34 Illustration of a portion of the final results of the creation of a model of RIVPACS type



Site	O50	E50	OE50	O0	E0	OE0
1	8	10.667	0.75	15	31.57	0.475
10	11	10.613	1.036	32	31.071	1.03
100	14	11.151	1.255	32	31.638	1.011

Figure 35 The second part of the final results of a RIVPACS-type model

If the selected model type is BEAST, the stress value[4] (Figure 36) and Sheppard[5] (Figure 37) and MDS plots (Figure 38) are presented.

[4] A statistic to help evaluate the quality of the ordination results (MDS).

[5] The Shepard plot displays the relationship between the proximities (BC dissimilarities) and the distances (MDS) of the point configuration. Less spread in this diagram implies a good fit.

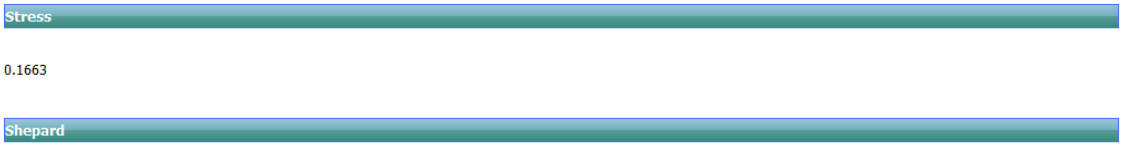


Figure 36 The stress indication in the final step of a BEAST-type model creation

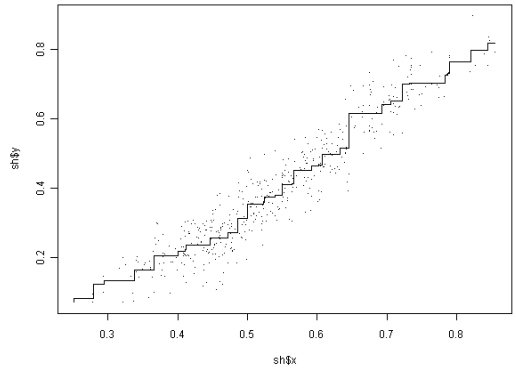


Figure 37 The Sheppard distribution of the model

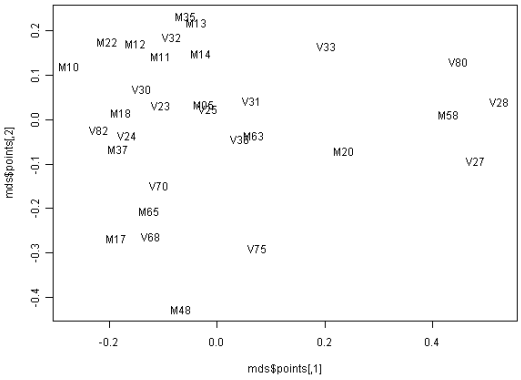


Figure 38 The MDS spatial distribution of each site

This step concludes the creation of predictive models and saves the model in the database, allowing it to be published in the web site upon administrative authorization for assessment of test sites.

4.2.5.2.Using models

In this section we will present the sequence associated with the utilization of predictive models for assessment of possibly affected sites. The process starts with the model selection and with providing the biological and environmental tables of the test sites (Figure 39). When

the user selects a model from the list, the discriminant variables and the type of model are displayed.

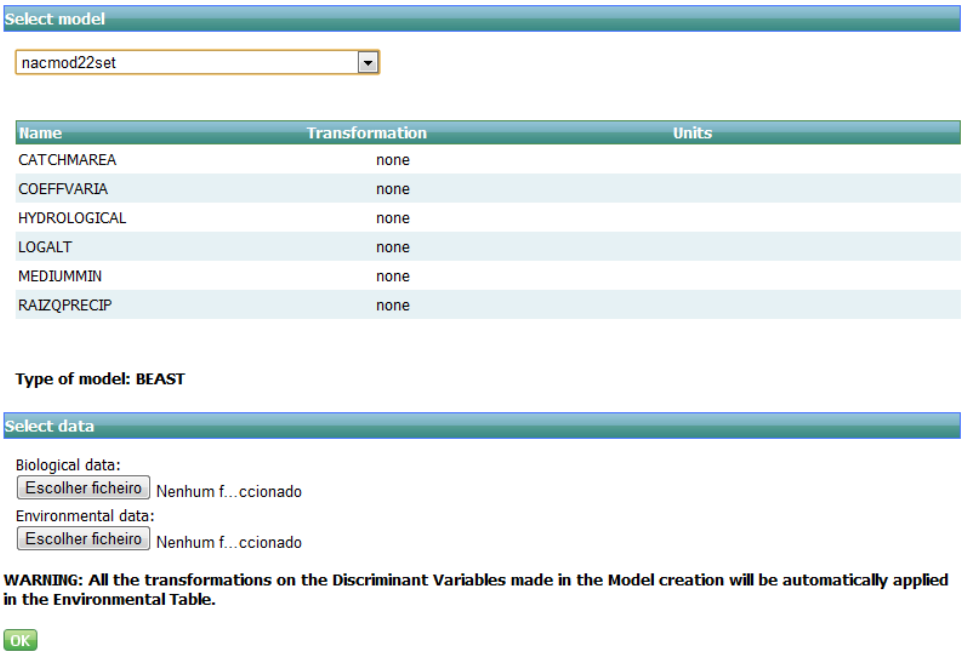


Figure 39 Illustration of the first step of using models

The following step allows the user to manage both tables and correct errors in the same way as in the creation of models (Figure 40). At the bottom of that page, the indication of the column with the sites on both tables is required. If the type of model selected is BEAST, the alternative to choose an extra ellipse defined by the user is also provided. If the user does not define a value, only the three typical ellipses are drawn in the plots (90, 99 and 99.9%).

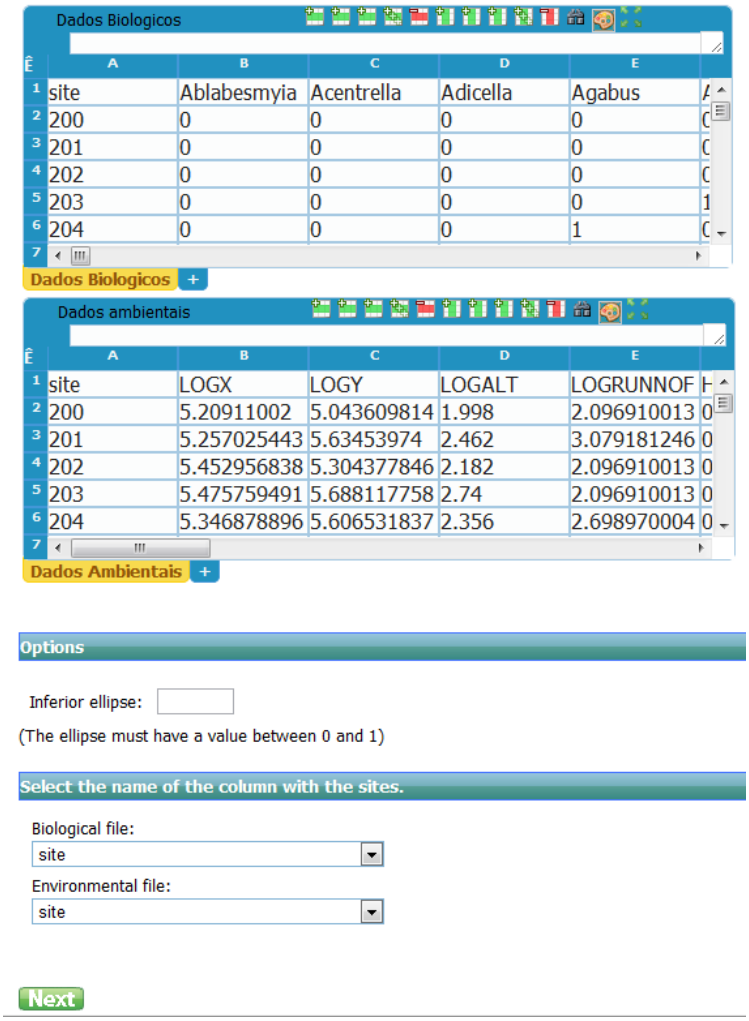
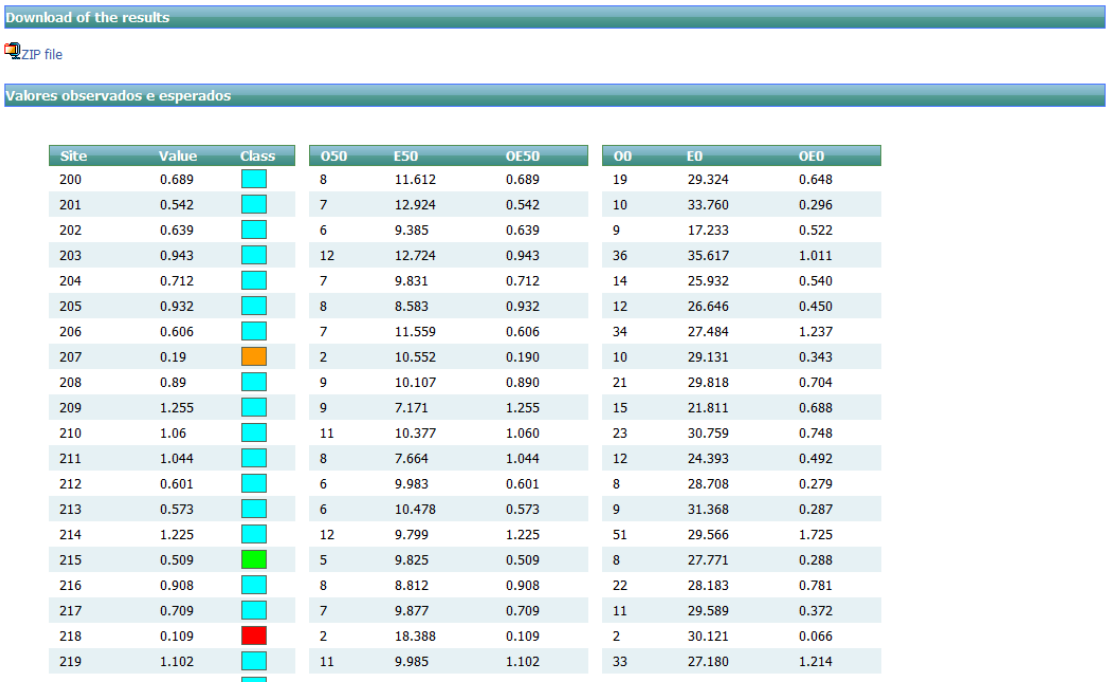


Figure 40 The table edition in model utilization

The final step of the analysis varies according to the selected type of model: if it is RIVPACS, the compressed file with all the results and a table with the quality classes and the OE50 and OE0 values are presented, as illustrated in Figure 41.



Download of the results

ZIP file

Valores observados e esperados

Site	Value	Class	O50	E50	OE50	O0	E0	OE0
200	0.689		8	11.612	0.689	19	29.324	0.648
201	0.542		7	12.924	0.542	10	33.760	0.296
202	0.639		6	9.385	0.639	9	17.233	0.522
203	0.943		12	12.724	0.943	36	35.617	1.011
204	0.712		7	9.831	0.712	14	25.932	0.540
205	0.932		8	8.583	0.932	12	26.646	0.450
206	0.606		7	11.559	0.606	34	27.484	1.237
207	0.19		2	10.552	0.190	10	29.131	0.343
208	0.89		9	10.107	0.890	21	29.818	0.704
209	1.255		9	7.171	1.255	15	21.811	0.688
210	1.06		11	10.377	1.060	23	30.759	0.748
211	1.044		8	7.664	1.044	12	24.393	0.492
212	0.601		6	9.983	0.601	8	28.708	0.279
213	0.573		6	10.478	0.573	9	31.368	0.287
214	1.225		12	9.799	1.225	51	29.566	1.725
215	0.509		5	9.825	0.509	8	27.771	0.288
216	0.908		8	8.812	0.908	22	28.183	0.781
217	0.709		7	9.877	0.709	11	29.589	0.372
218	0.109		2	18.388	0.109	2	30.121	0.066
219	1.102		11	9.985	1.102	33	27.180	1.214

Figure 41 The final step of model utilization for RIVPACS-type models

On the other hand, if the model type is BEAST, along with the compressed file containing the analysis results, several plots with the multi-dimensional scaling (MDS) of each test site with its group of reference sites[6] are presented. By selecting any test site in the left area of the page, the three combinations of the first three MDS axes, two by two are shown: the first and the second dimensions (Figure 42), the first and the third (Figure 43) or the second and the third (Figure 44).

[6] The green points represent reference sites, the red point represents the selected test site and the black point represents the centroid of the ellipses.

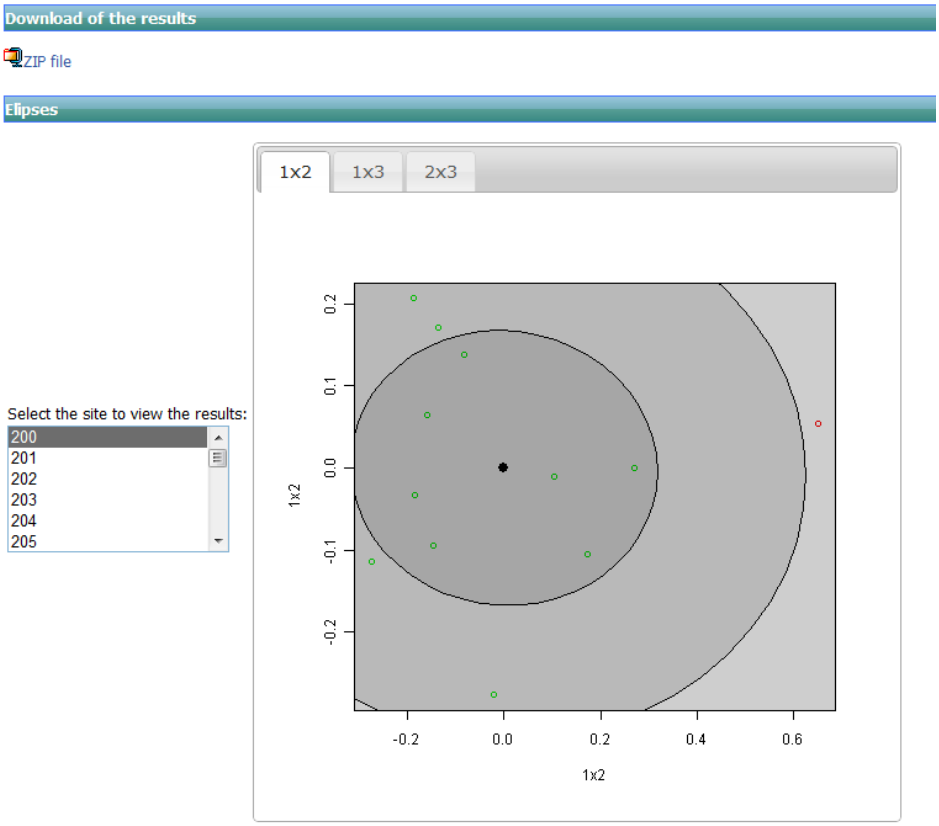


Figure 42 The final step of a model utilization of BEAST type

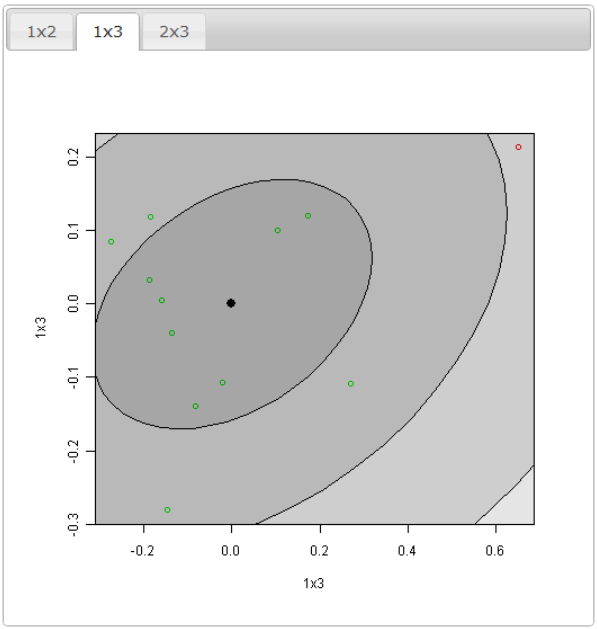


Figure 43 Illustration of the selection of the second tab of the ordination spaces of a site

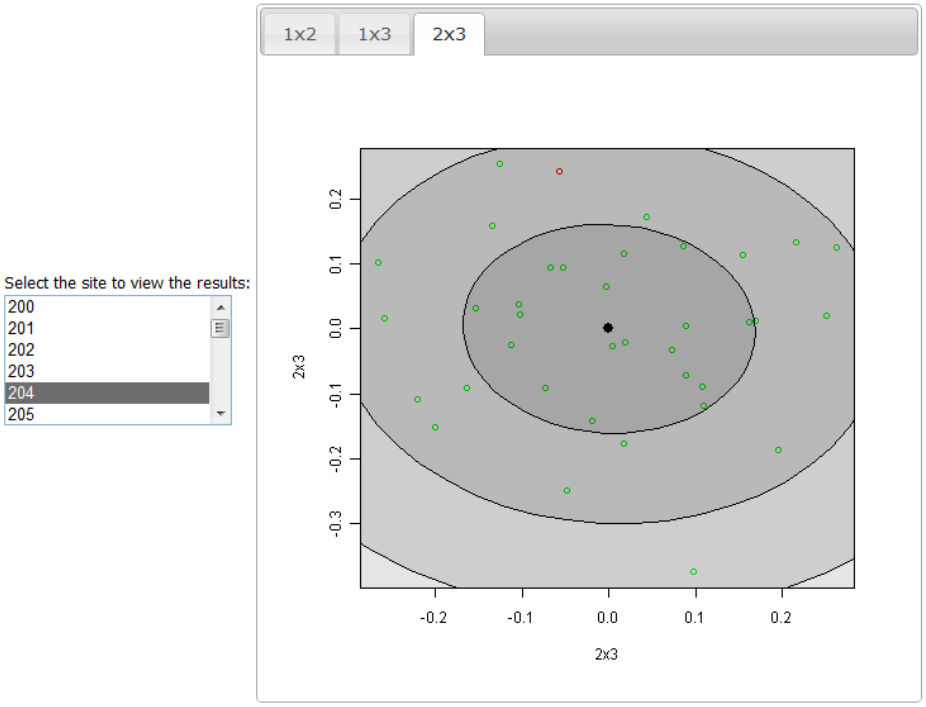


Figure 44 Demonstration of the selection of the third tab (y and z coordinates) of a site (site "204")

4.2.6. Analysis execution performance

Regarding the performance, the web site has instantaneous response to the pages not related to predictive modeling.

On the model creation pages, the responses typically take up to 3 seconds to execute, except on the following cases:

- The table edition page, since the loading of the page on client side after server response may take more than 30 seconds to complete, owing to tables with large amounts of data, which are all loaded at once. On this matter, the user has to rely on the performance of the web browser. The table edition is instantaneously responsive after loading the page.
- The model building step of the all-subsets regression, the 6th step. This is due to the time needed for the R script to perform all calculations, as it creates and processes many possible models. This script may take several minutes depending highly on the number of environmental variables provided by the user.

4.2.7. Model Outputs

As required in the initial objectives, this implementation of the predictive models provides more and better information to the user.

For model creation, the following information is provided in a compressed file:

- Histograms of each environmental variable with its original distribution and its normalized distribution, if available;
- The dendrogram of the distributed sites based on their biological distributions, without and with the selected grouping;
- The latitude and longitude of all the sites;
- The biological and environmental tables, both original (without any errors) and with transformations in CSV file type. The environmental table is provided without the location columns;
- The transformations performed on the environmental variables;
- A plot of the RMSE (Root Mean Squared Error)[7] from the best models of each order against model order, for both calibration and validation data;
- A plot of the classification accuracy of the 5 best models of each model order;
- The predictor importance (through the percentage of best models that include the predictor);
- The entire all-subsets processing;
- The limits of the quality classes;
- The selected discriminant variables;
- The selected number of groups;
- Also, specifically for BEAST:
    - An MDS plot of the reference sites;
    - The MDS processing for both the calibration data and all the data;
    - The Shepard distribution diagram;
    - The model stress.
- Also, specifically for RIVPACS:
    - Several files with O/E statistics such as the O50, O0, E50, E0, OE50 and OE0 for the null model, in CSV;
    - The OE50 and OE0 processing description and full information;
    - OE50 “O vs E” scatterplot;
    - A q-q plot[8] of the O vs. E;
    - A histogram of the OE values;
    - O vs. E linear regression plot;
    - A table with each site’s resulting class.

[7] The RMSE is a measure that measures the differences between estimated and observed values of a model, for purposes of accuracy.
[8] A q-q plot illustrates the quantile distributions of two variables against each other.

Regarding the model using, the following results are provided to the user in the final step of the analysis, in a compressed file:

- Specifically for BEAST:
    - Plots of the MDS results of each test site (as a red point) with its reference group (as green points), for each 3 combinations of two coordinates of the 3D points;
    - Number of iterations and stress of the calculation of the MDS of each test site;
    - The original correct biological and environmental test tables.
- Specifically for RIVPACS:
    - The classes limit and the resulting classes for each test site;
    - The histogram of the O vs. E values;
    - A q-q plot of the O vs. E;
    - The OE50 and OE0 processing description and full information;
    - The OE50 and OE0 tables;
    - The original correct biological and environmental test tables.

4.3.Global matrix

In order to maintain, in a single place, all the biological and environmental information collected from several different sites and years, the global matrix page was created. This page contains a table for biological data and another for environmental data. These tables are fully replaceable or editable. To be editable means that a user may edit the existing sites or add new sites to the tables manually or by uploading a new file of multiple sites. The new sites data table is concatenated at the end of the existent table. All the information is downloadable in three different formats (CSV, XLS and XLSX) at any time. These features are depicted in Figure 45 and Figure 46.

Since the global matrix also supports location columns (with manual point introduction, as described later in section 4.5.1), a web page was also created to show the location of all the points listed in a global map, similarly to the page of model listing.

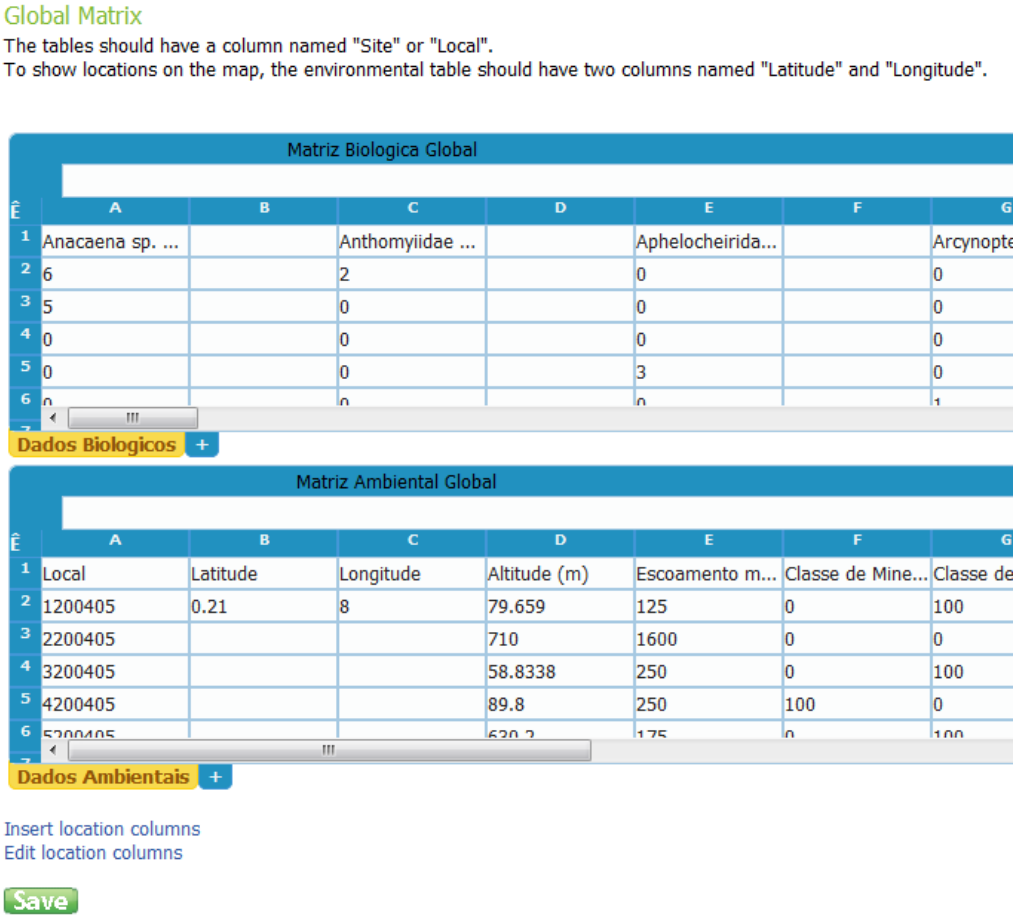


Figure 45 The first half of the edition page for the biological and environmental global matrices.

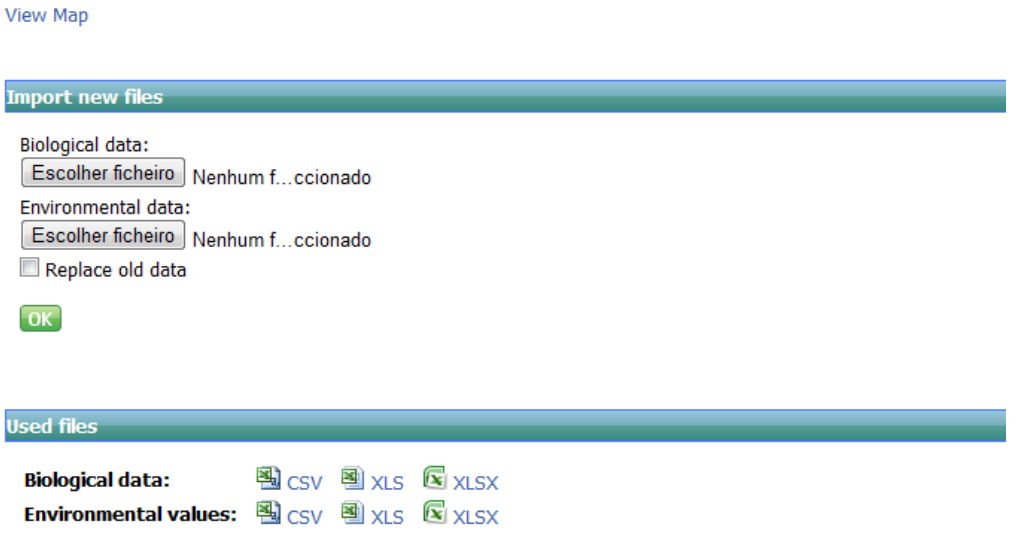


Figure 46 The second half of the global matrix edition page

4.4.Taxonomic keys

The identification of taxa is a complex and time consuming work but is the basis for all bioassessment methods. Taxonomic keys, a series of paired statements that describe the physical characteristics of different organisms, are used to find the correct name of the species, genus or family of a given individual. Based on dichotomic taxonomic keys, a software tool called Lucid Phoenix (Lucid.org, University of Queensland) was used by the Institute of Marine Research to develop a digital key for Portuguese aquatic freshwater macroinvertebrates.

This tool provides an embeddable interface for iteratively reducing the universe of possibilities with the help of visual content, through videos and photographic examples. Therefore, this project aimed the incorporation of the built key.

For that purpose, a new page for utilization of the taxonomic key identifier was implemented. This page incorporates the Lucid Phoenix Player embedded as a java applet. This applet is interactive: the user may navigate through the key builder by answering questions and visualizing images/videos to help support the decisions. An example of the second step of identification (the user already answered one question as "Present") is portrayed in Figure 47.

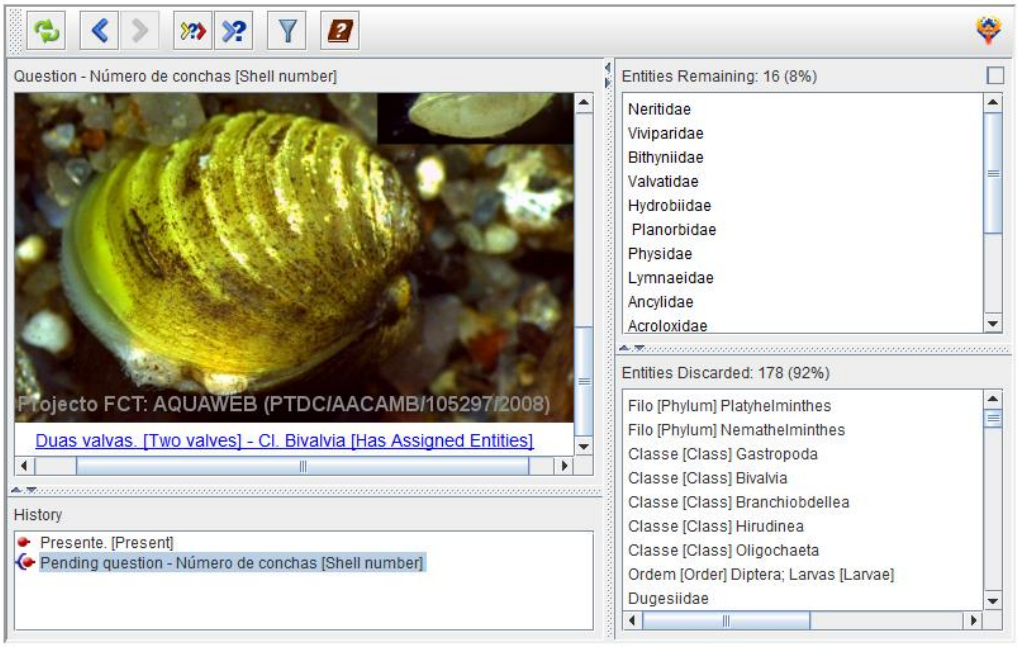


Figure 47 Taxonomic key identification tool (Lucid Phoenix Player)

Besides the image of species, another form of visual representation is supported by this feature to facilitate the identifications: video. The videos were firstly uploaded to *YouTube*[9] by the IMAR team and provided when creating the Lucid Phoenix taxonomic key project. Then, embedded in the web page of AQUAWEB, the video is presented (Figure 48).

[9] A video hosting web site in available in *www.youtube.com*.

Figure 48 Demonstration of a video for the Taxonomic key identifier

Moreover, an administrative page where the user may upload a new version of the identification tool was created, as depicted in Figure 49. The folder necessary for uploading must include, besides the content for the decisions, a Java based player for the interaction in the web site. In order to create this folder, the user must create all of the taxonomic keys in the Lucid Phoenix software, export the project (resulting in the folder with the contents described above) and place it in a compressed file. This compressed file may then be provided to the web site.

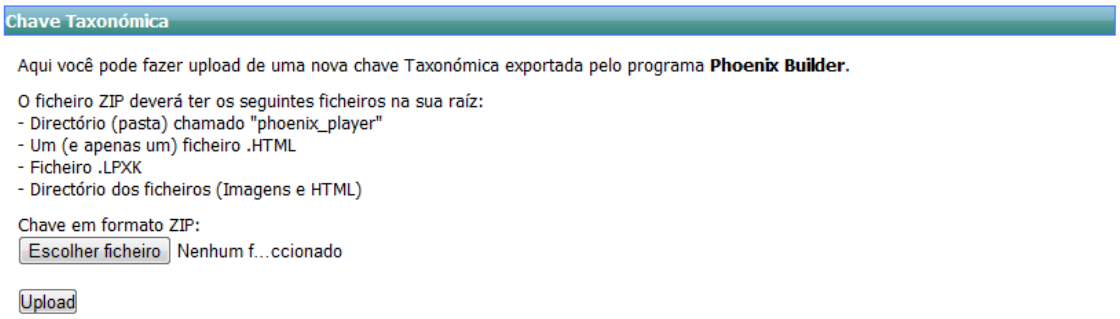


Figure 49 Uploading a taxonomic key identification version

4.5.Maps

This last section of the platform implementation indicates the features that support maps, namely the manual point introduction by pressing a point in a map that saves the selected coordinates and the representation of the classes of the sites included in a freshwater quality assessment through colored pins in the map.

4.5.1. Manual point introduction

Several features of the project deal with location attributes: the model creation and use, the global matrix, the listed models and the model edition in the administrative section. These location attributes are represented as columns in environmental tables.

When editing a table, the user may edit the longitude and latitude columns if they exist, or even add these new columns, which will be placed next to the last column of the environmental table. Upon this insertion or edition of the location columns, which must be identified as “longitude” and “latitude”, the user may manage the location of each site of the table by selecting a point in a map. The longitude and latitude of that point are automatically placed in the respective columns of the table.

Also, after indicating the location of a site in the map, if the user edits a new site location, it will be placed in the location of the previous site, for easier placing in close sites. After indicating the desired locations, the user must to click the “Save” button at the end of the list for all the changes to be saved. An example of the manual point introduction is illustrated in Figure 50.

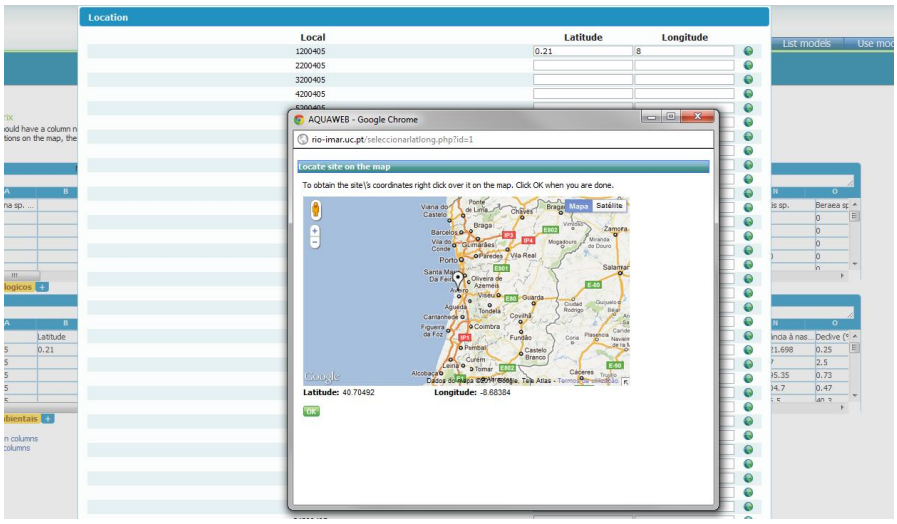


Figure 50 Illustrative example of the selection of the location of a site through the map

4.5.2. Analysis resulting classes

When a model is created, it may be listed upon administrative permission. A published model presentation provides the most important information about the model. This information includes the defined intervals of the classes (Figure 51) and a map marking the sites with the associated quality class assigned to each one of them (Figure 52).

Band	Interval	Colour
HIGH	1.8254 - 0.7948	
GOOD	0.7948 - 0.5961	
MODERATE	0.5961 - 0.3974	
POOR	0.3974 - 0.1987	
BAD	0.1987 - 0	

Figure 51 Definition of the classes

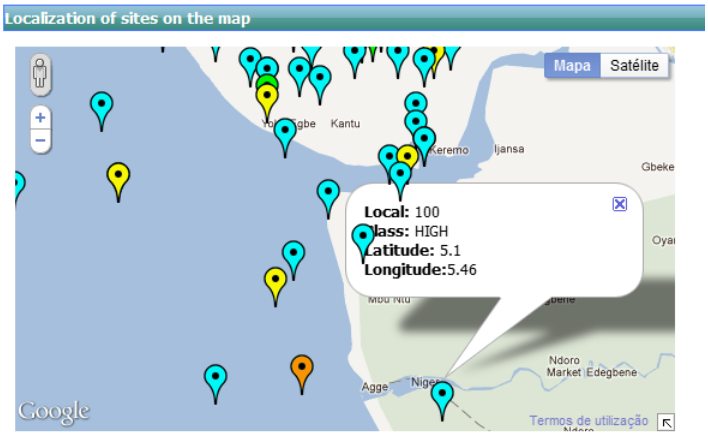


Figure 52 Visual demonstration of the sites in the analysis

5. Conclusions

The main objective in this dissertation, implemented in the scope of the project FCT (PTDC/AAC-AMB/105297/2008) and currently in development, was creating an integrated tool that would support water quality assessment through predictive modeling approaches, whilst knowing these approaches need heavy long calculations and information management that discourage their application.

To achieve these goals, we created a web application to allow the processing of all the necessary steps. The data that researchers gather does not need now any prior treatment to be provided to the platform, and may be provided in the most common file types. A visual and detailed edition of the data allows the users to only proceed to the following steps when the data is completely correct.

Moreover, the processing of the sequence involved with the assessments is now divided in more and smaller steps, so that the researcher may have a closer notion of intermediate results and have a more informed decision about the configurations provided for the following step. Finally, when an assessment is completed, all the results presented throughout the analysis may be downloaded for further considerations.

Also, a new version of the database of the platform was designed and implemented in order to improve scalability, performance and reliability for future developments on the project.

All the requirements listed in chapter 3 were accomplished by the platform. In addition, an exceptional goal was achieved, the improvement of the database design structure.

5.1. Result

The resulting application of this dissertation is a practical utility to be used by researchers in bioassessment of streams and rivers ecosystems through predictive modeling, to detect potential threats and prevent further deterioration of these ecosystems, as well as to report to national regulation committees and international organizations.

To conclude, we present a SWOT analysis of the system in Figure 53.

Strengths	Weaknesses
• Fully integrated processing of predictive modeling • Effortless and intuitive error detection in data tables • Scalability of the platform • Information detail on models	• Long processing time in certain model creation steps • Need of providing even better calculating features of the predictive models
• Need of a tool to support the entire processing in predictive modeling	• New developments in the area may require updating the techniques
Opportunities	**Threats**

Figure 53 SWOT Analysis of the application

5.2.Future work

The platform is been used in real environment and user feedback has been provided. However, some features can be improved and some work was left to be done in the future. Next, we will summarize what should be done to improve and expand the AQUAWEB solution:

- The presentation of the results might be improved, namely the display of intermediary results in each step. For instance, the results could be identified by colors and the order of table columns could be changed. This will help the researcher to better comprehend the values presented;
- The R scripts may also be worked in order to provide detailed logs of performed calculations. A more detailed log of results errors is requested, identifying the problematic data and the algorithm point associated to the execution error. Moreover, the R variables associated to each algorithm step could be also provided to the user for further analysis of the intermediate and final results;
- Still regarding the R scripts, a few improvements on the techniques might be achieved, such as allowing the taxa abundance matrix instead of a presence/absence matrix to be considered for site clustering; and reusing of the calibration data for the both the validation of the model and creation of the null model, instead of a separate set of validation data, thus using all the available data to process the model without disregarding any sites;
- Furthermore, the loading time of the biological and environmental data tables in the supported use cases could be improved with, for example, a library that would allow partial loading of the table, and providing further data as the user scrolls the table.

Appendix A

This appendix provides the visual presentation of the RIO and AQUAWEB databases.

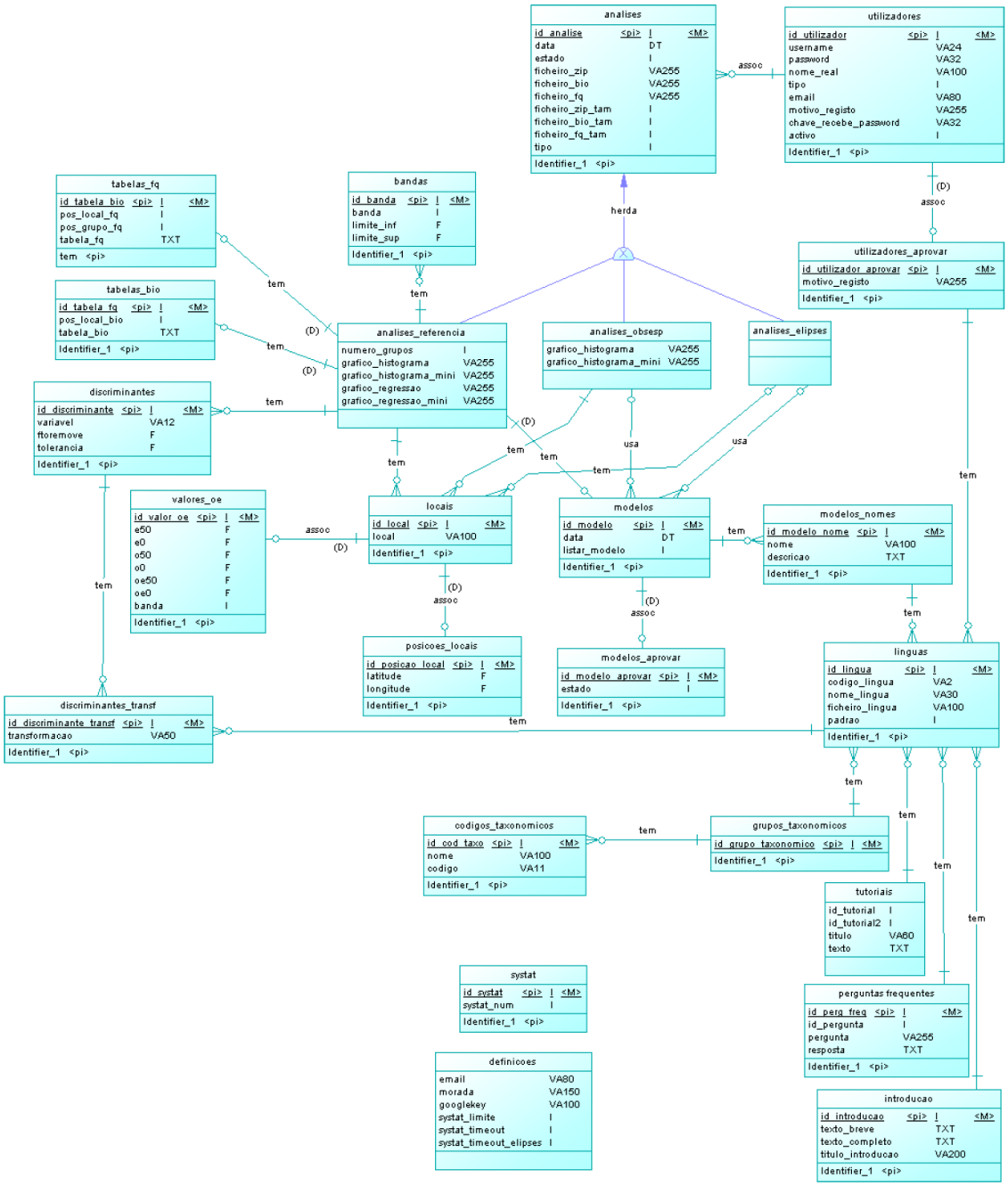


Figure 54 Database model from RIO

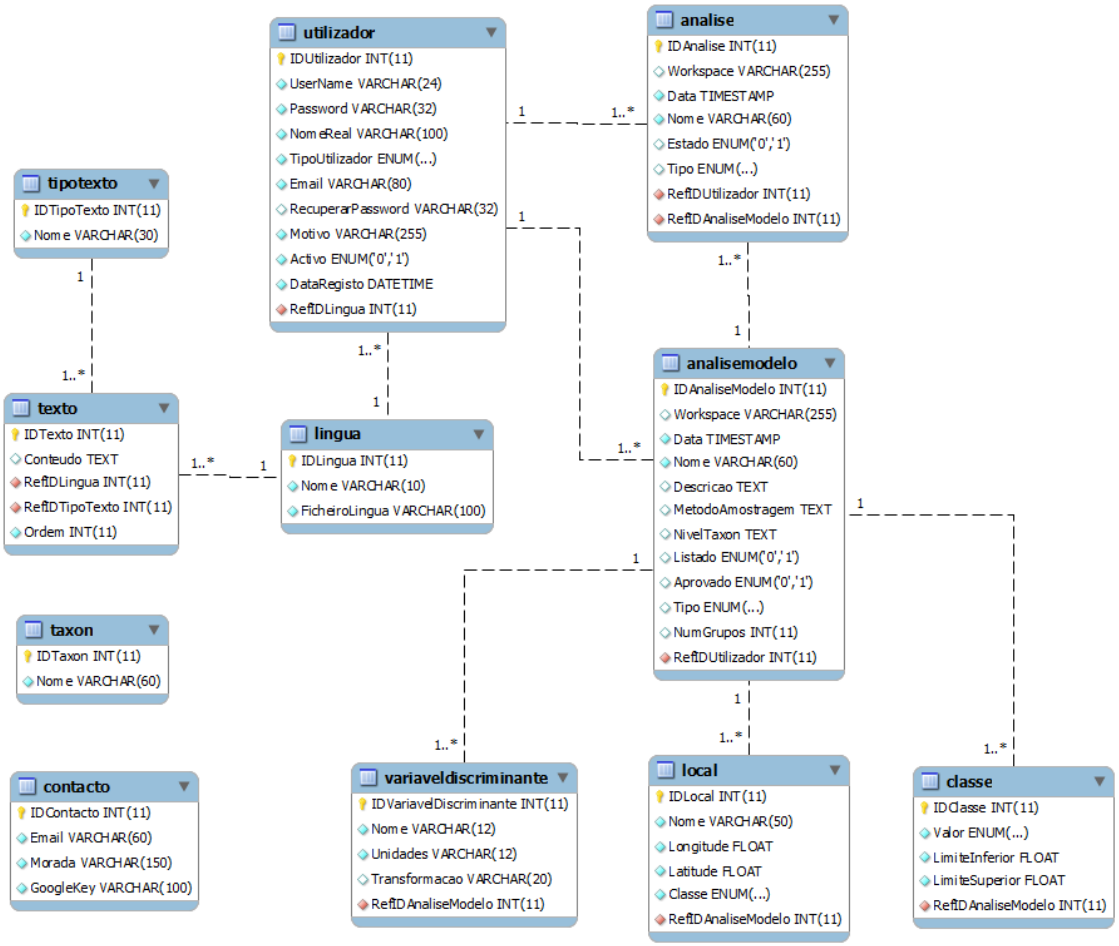


Figure 55 Database model from AQUAWEB

Appendix B

The creation of a predictive model follows several steps, as depicted in Figure 56. Firstly, the user may manage the biological and environmental tables. Then, he may transform environmental variables by normalizing their distributions. After this, the user must select the number of groups in the dissimilarity dendrogram of the sites. When the preferred grouping is chosen, the user may then select, from the list of the models suggested by the algorithm, a model. At the final step, all the final results are presented and the user may download the information provided in all steps of the model creation.

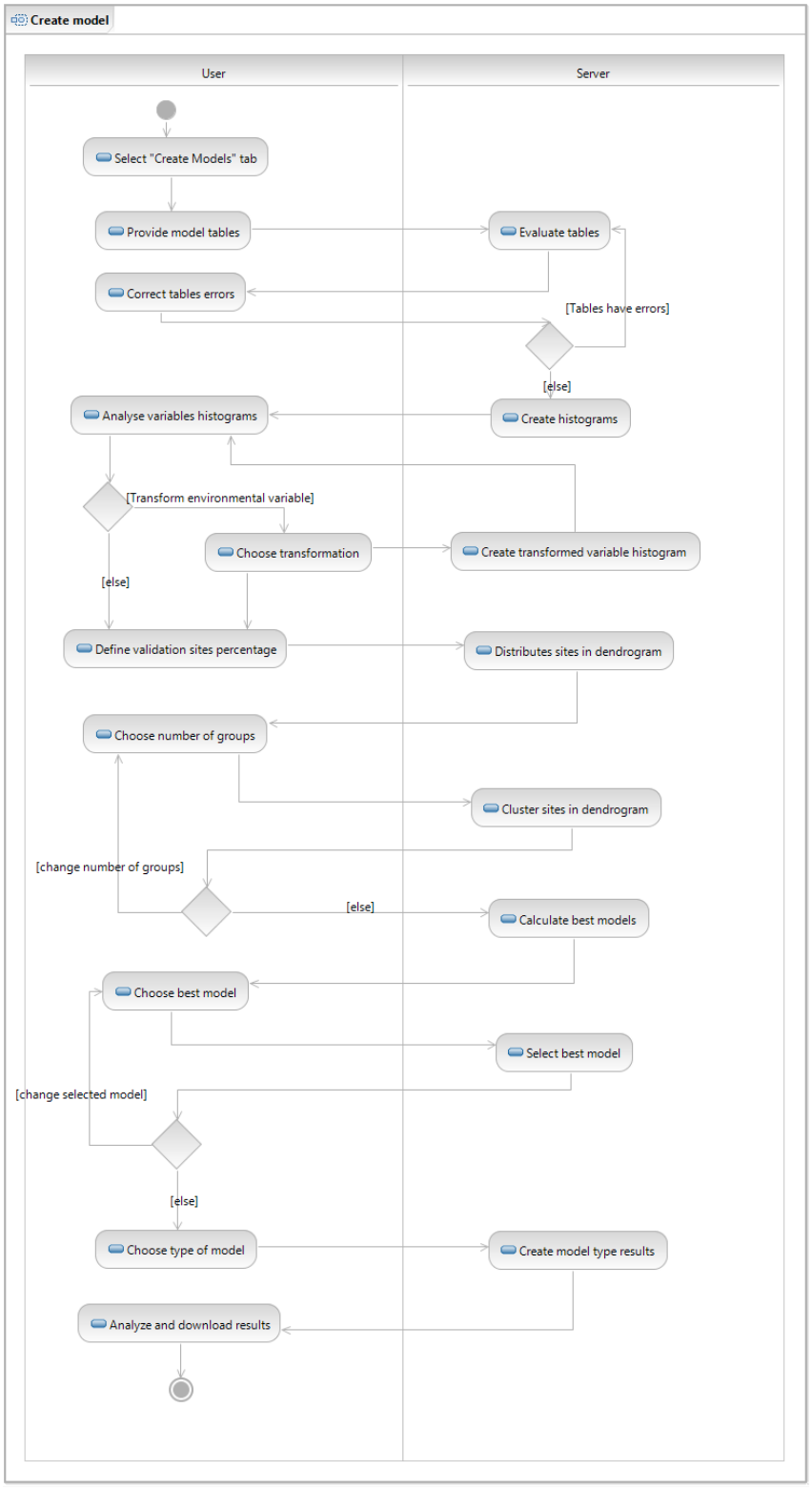


Figure 56 The necessary steps for creating a model

As for the model utilization, the necessary steps are, as shown in Figure 57, the selection of the desired existent predictive model, managing the biological and environmental test tables and analyzing and downloading the analysis results.

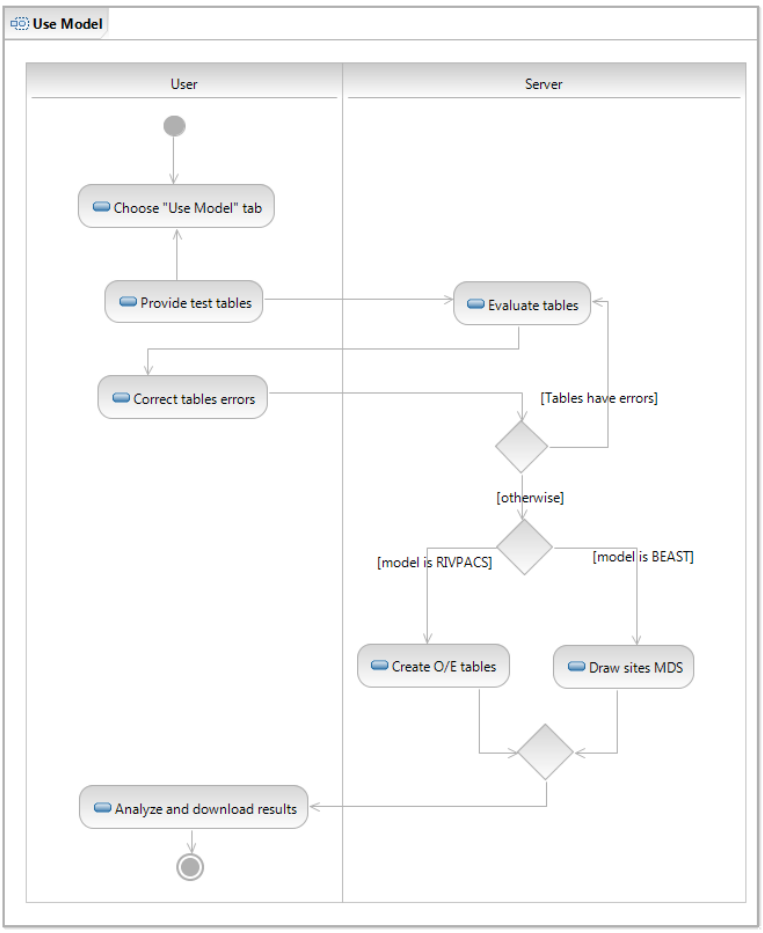


Figure 57 Activity diagram illustrating the utilization of a model

Appendix C

The server side processing tasks associated to all activities presented in the “Create model” and “Use model” Use Case Diagrams are executed with the help of R scripts, with the exception of the table editions. The first step involving R scripts is the normalization of the entire biological table at once, applying the indicated selection of the user in the first step. The available transformations are squared or fourth root, logarithm and presence/absence. This transformation is optional, since the user can choose not to apply any transformation. This process is illustrated in Figure 58.

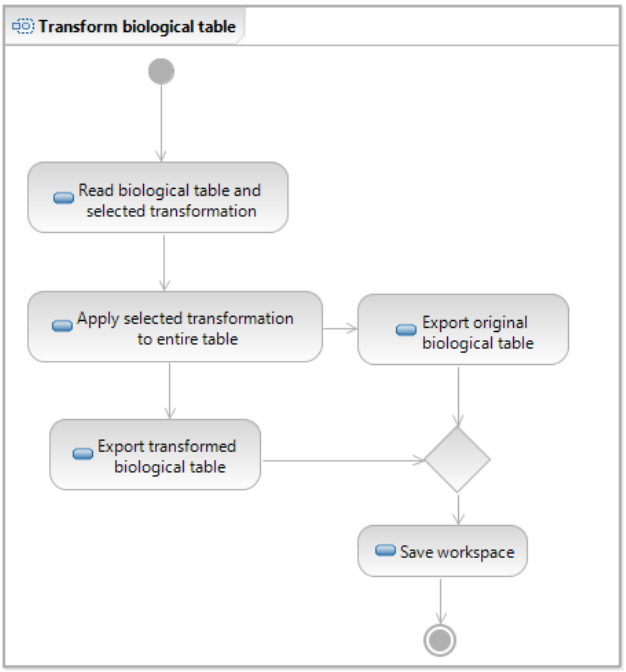


Figure 58 Applying a transformation to all the variables included in the biological table

Also in this same step, distribution histograms are drawn for each of the environmental variables, for the user to evaluate the need of also applying a transformation to any of them. When a transformation is selected, a histogram of the transformed variable is presented for further evaluation, as can be seen in Figure 59.

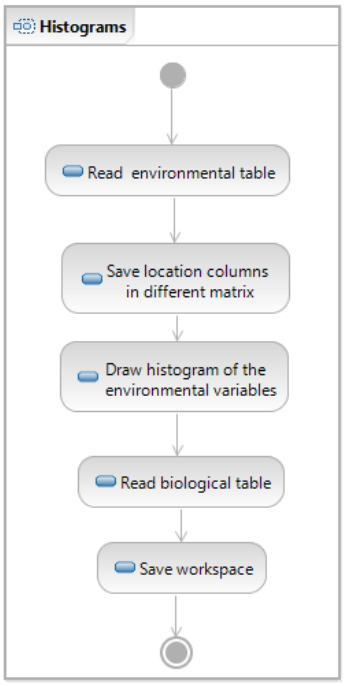


Figure 59 Activity diagram describing the steps for creating the histograms of the environmental variables

The third script used in the creation of predictive models generates the clustering of the sites based on dissimilarities of non-rare taxa, through a Bray-Curtis dissimilarity matrix, presented in the form of a dendrogram (Figure 60).

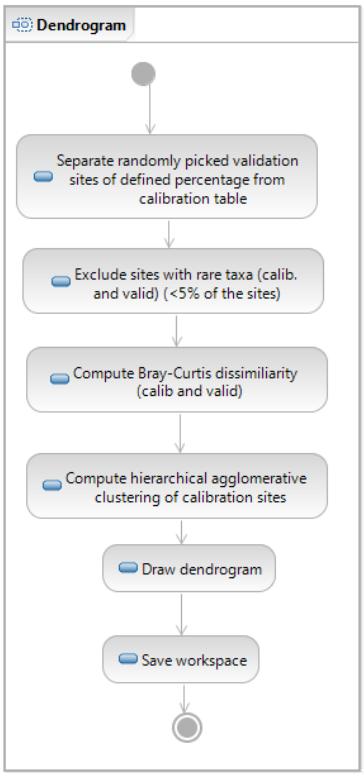


Figure 60 Activity diagram of the creation of the dendrogram of the non-rare sites

After the number of groups is selected, the all-subsets regression method is performed as depicted in Figure 61.

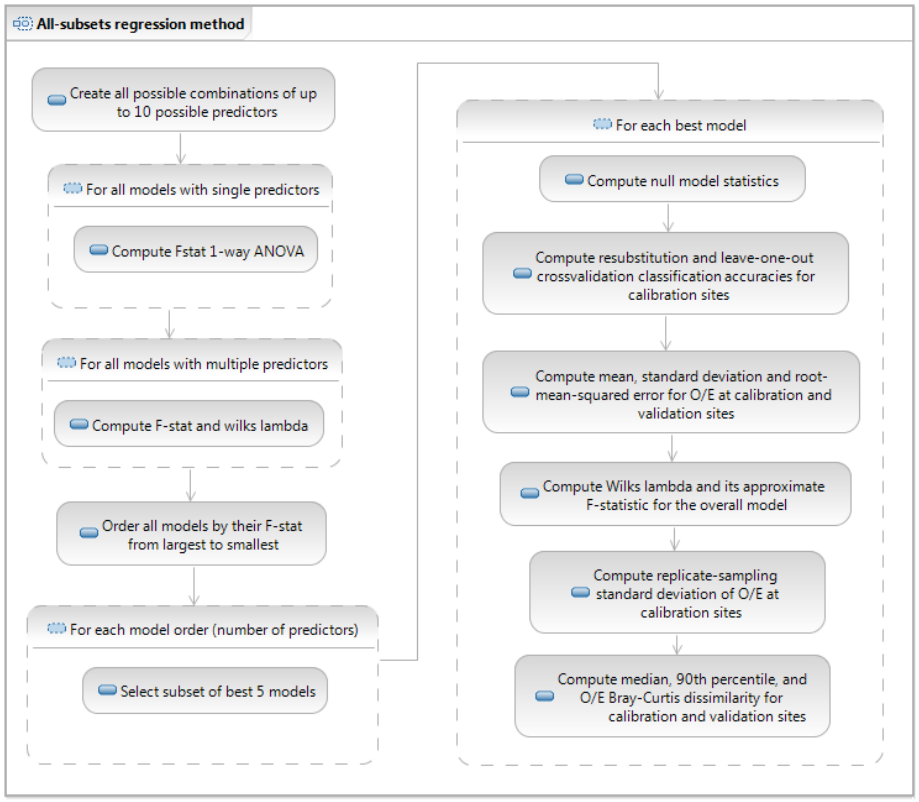


Figure 61 Activity diagram demonstrating the process of the all-subsets regression method

Figure 62 illustrates how the results of the previous script are used to create plots for more detailed information when choosing the desired model from the presented subset of best models.

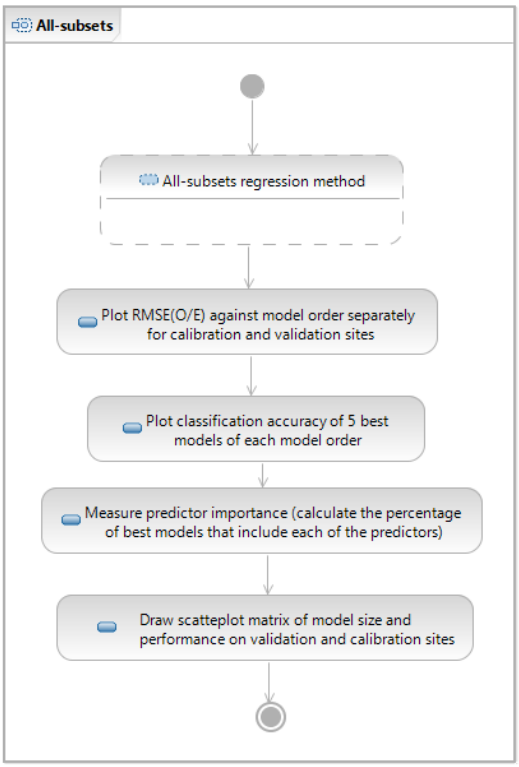


Figure 62 Processing of all the data provided by the all-subsets regression

The last step of the creation of models depends on the type of model the user selected to create: RIVPACS or BEAST. The results of both resulting scripts are the quality class definition, according to the respective model type, as well as a few other results, which are demonstrated later in section 4.2.5 (Step by step algorithm).

Regarding the “Use model” Use Case, there is a script for each of the model-types. Each of them calculate the O/E values and define the resulting quality classes as well as some other model-type specific results, which are demonstrated in section 4.2.5 (Step by step algorithm).

References

1. Clarke, R.T., J.F. Wright, and M.T. Furse, *RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers.* Ecological Modelling, 2003. **160**(3): p. 219-233.
2. Wright, J.F., D.W. Sutcliffe, and M.T. Furse, *Assessing the biological quality of fresh waters: RIVPACS and other techniques.* 2000.
3. Bailey, R.C., R.H. Norris, and T.B. Reynoldson, *Bioassessment of freshwater ecosystems: using the reference condition approach*. 2004: Springer Netherlands.
4. Aguiar, F.C., M.J. Feio, and M.T. Ferreira, *Choosing the best method for stream bioassessment using macrophyte communities: Indices and predictive models.* Ecological Indicators, 2011. **11**: p. 379-388.
5. Joy, M.K. and R.G. Death, *Development and application of a predictive model of riverine fish community assemblages in the Taranaki region of the North Island, New Zealand.* New Zealand Journal of Marine and Freshwater Research, 2000. **34**(2): p. 241-252.
6. Feio, M.J., et al., *Diatoms and macroinvertebrates provide consistent and complementary information on environmental quality.* Fundamental and Applied Limnology/Archiv für Hydrobiologie, 2007. **169**(3): p. 247-258.
7. Gordon, N.D., *Stream hydrology: an introduction for ecologists*. 2004: Wiley.
8. Chapman, D.V., et al., *Water quality assessments: a guide to the use of biota, sediments, and water in environmental monitoring*. 1996: E & FN Spon.
9. Wehr, J.D. and R.G. Sheath, *Freshwater algae of North America: ecology and classification*. 2003: Academic Press.
10. Naddeo, V., T. Zarra, and V. Belgiorno. *European procedures to river quality assessment*. 2005.
11. Liefferink, D., M. Wiering, and Y. Uitenboogaart, *The EU Water Framework Directive: A multi-dimensional analysis of implementation and domestic impact.* Land Use Policy, 2011. **28**(4): p. 712-722.
12. Reynoldson, T., et al., *The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates.* Journal of the North American Benthological Society, 1997: p. 833-852.
13. Stoddard, J.L., et al., *Setting expectations for the ecological condition of streams: the concept of reference condition.* Ecological Applications, 2006. **16**(4): p. 1267-1276.
14. Wright, J.F., P.D. Armitage, and M.T. Furse, *RIVPACS : a technique for evaluating the biological quality of rivers in the UK*. Vol. European Water Pollution Control ;. 1993.
15. Krzanowski, W.J., *Principles of multivariate analysis: a user's perspective*. 2000: Oxford University Press, USA.
16. Moss, D., et al., *comparison of alternative techniques for prediction of the fauna of running water sites in Great Britain.* Freshwater Biology, 1999. **41**(1): p. 167-181.
17. Moss, D., et al., *The prediction of the macro invertebrate fauna of unpolluted running water sites in Great Britain using environmental data.* Freshwater Biology, 1987. **17**(1): p. 41-52.
18. Norris, R.H. and K. Morris, *The need for biological assessment of water quality: Australian perspective.* Australian Journal of Ecology, 1995. **20**(1): p. 1-6.
19. Simpson, J., et al. *Biological assessment of river quality: development of AUSRIVAS models and outputs*. 2000: Freshwater Biological Association (FBA).

20. Belbin, L. and C. McDonald, *Comparing three classification strategies for use in ecology.* Journal of Vegetation Science, 1993: p. 341-348.
21. Faith, D.P., P.R. Minchin, and L. Belbin, *Compositional dissimilarity as a robust measure of ecological distance.* Plant Ecology, 1987. **69**(1): p. 57-68.
22. Johnson, R.K. and L. Sandin, *Development of a prediction and classification system for lake (littoral, SWEPAC LLI) and stream (riffle, SWEPAC SRI) macroinvertebrate communities.* Department of Environmental Assessment, Swedish University of Agricultural Sciences, Rapport, 2001. **23**: p. 1-66.
23. Kokeš, J., et al., *The PERLA system in the Czech Republic: a multivariate approach for assessing the ecological status of running waters.* Hydrobiologia, 2006. **566**(1): p. 343-354.
24. Feio, M., et al., *Water quality assessment of Portuguese streams: Regional or national predictive models?* Ecological Indicators, 2009. **9**(4): p. 791-806.
25. Poquet, J.M., et al., *The MEDiterranean Prediction and Classification System (MEDPACS): an implementation of the RIVPACS/AUSRIVAS predictive approach for assessing Mediterranean aquatic macroinvertebrate communities.* Hydrobiologia, 2009. **623**(1): p. 153-171.
26. Reynoldson, T.B., et al., *Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state.* Australian Journal of Ecology, 1995. **20**(1): p. 198-219.
27. Rosenberg, D., et al. *Establishing reference conditions in the Fraser River catchment, British Columbia, Canada, using the BEAST (BEnthic Assessment of SedimenT) predictive model*. 2000: Freshwater Biological Association (FBA).
28. Feio, M.J. and J.M. Poquet, *Predictive Models for Freshwater Biological Assessment: Statistical Approaches, Biological Elements and the Iberian Peninsula Experience: A Review.* International Review of Hydrobiology, 2011. **96**(4): p. 321-346.
29. Belbin, L., *Semi-strong hybrid scaling, a new ordination algorithm.* Journal of Vegetation Science, 1991: p. 491-496.
30. Reynoldson, T., et al. *The development of the BEAST: a predictive approach for assessing sediment quality in the North American Great Lakes*. 2000: Freshwater Biological Association (FBA).
31. Feio, M.J., T.B. Reynoldson, and M.A. Graça, *Effect of seasonal changes on predictive model assessments of streams water quality with macroinvertebrates.* International Review of Hydrobiology, 2006. **91**(6): p. 509-520.
32. Feio, M.J., et al., *A predictive model for freshwater bioassessment (Mondego River, Portugal).* Hydrobiologia, 2007. **589**(1): p. 55-68.
33. Moreno, P., et al., *Use of the BEAST model for biomonitoring water quality in a neotropical basin.* Hydrobiologia, 2009. **630**(1): p. 231-242.
34. Linke, S., et al., *ANNA: a new prediction method for bioassessment programs.* Freshwater Biology, 2005. **50**(1): p. 147-158.
35. Olden, J.D., M.K. Joy, and R.G. Death, *An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data.* Ecological Modelling, 2004. **178**(3-4): p. 389-397.
36. Nigrin, A., *Neural networks for pattern recognition*. 1993: MIT Press.
37. Hoang, T.H., et al., *Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam.* Ecological Informatics, 2010. **5**(2): p. 140-146.
38. De'ath, G. and K.E. Fabricius, *Classification and regression trees: a powerful yet simple technique for ecological data analysis.* Ecology, 2000. **81**(11): p. 3178-3192.
39. Adriaenssens, V., et al. *Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers*. 2004: Cambridge Univ Press.

40. Hocking, R., *A BIOMETRICS INVITED PAPER: THE ANALYSIS AND SELECTION OF VARIABLES IN LINEAR REGRESSION.* Biometrics, 1976. **32**: p. 1-49.
41. Cohen, J., *Applied multiple regression/correlation analysis for the behavioral sciences*. Vol. 1. 2003: Lawrence Erlbaum.
42. Rawlings, J.O., S.G. Pantula, and D.A. Dickey, *Applied regression analysis: a research tool*. 1998: Springer.
43. Rousseeuw, P.J. and A.M. Leroy, *Robust regression and outlier detection*. 2003: Wiley-Interscience.
44. Sickle, J.V., D.D. Huff, and C.P. Hawkins, *Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates.* Freshwater Biology, 2006. **51**(2): p. 359-372.
45. Corder, G.W. and D.I. Foreman, *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. 2011: John Wiley & Sons.
46. Stephenson, F.H., *Calculations for Molecular Biology and Biotechnology: A Guide to Mathematics in the Laboratory*. 2010: Academic Pr.
47. Berrar, D.P., W. Dubitzky, and M. Granzow, *A practical approach to microarray data analysis*. 2003: Kluwer Academic Publishers.
48. Özsu, M.T. and L. Liu, *Encyclopedia of Database Systems*. 2009: Springer.
49. Ihaka, R. and R. Gentleman, *R: a language for data analysis and graphics.* Journal of computational and graphical statistics, 1996: p. 299-314.
50. *PHPExcel*. Available from: http://phpexcel.codeplex.com/.
51. *JQuery stylesheet*. Available from: http://visop-dev.com/Project+jQuery.sheet.
52. Sickle, J.V. *Western Ecology Division | US EPA: R-language scripts for RIVPACS-type predictive modeling*. Available from: http://www.epa.gov/wed/pages/models/rivpacs/rivpacs.htm.