DeepData: Web application for Azores deep sea

Magda Resende
magda.resende@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2018

Abstract

Azores' deep sea is the habitat of several living beings. Knowing the relationships that these living beings establish with their habitats is essential for the development of appropriate measures for protecting marine ecosystems. The species distribution models (SDM), which associate environmental conditions with the occurrence of species, are an effective way to predict the relationships between living beings and the environment. This paper describes the web application DeepData, which integrates several data sources that store marine data, stores the data in a relational database and enables the application of predictive models of species distribution to the marine region of the Azores. In this paper, we also present the validation of the application DeepData. The validation results highlight the ease of use of the application DeepData and its usefulness for the calculation of SDM. However, some features need to be adjusted to allow for greater user interaction with statistical models.
Keywords: Data Integration, Azores Deep Sea, Species Distribution Models, Database

1. Introduction

The study of the marine region and the knowledge about its climatic and morphological characteristics is essential to understand the life of aquatic organisms. Through the analysis of the interaction of aquatic organisms with their habitat, it is possible to evaluate the impact of human interactions and climate change on these ecosystems. Understanding the environmental conditions favorable to the development of the species enables predicting the location of hitherto unknown habitats and taking the necessary preventive measures to protect them. To obtain knowledge about the habitats, marine researchers use a set of data sources for species occurrence and environmental conditions, on which they apply Species Distribution Models (SDM).

A group of researchers from IMAR, Instituto do MAR, with headquarters in the Department of Oceanography and Fisheries of the University of the Azores [1], who study the deep sea of the Azores, intend to automate the computation of SDM to this region. SDM provide knowledge about the relationship between species and their habitats, statistically relating species distributions with environmental data to obtain predictions of the geographic distribution of these species [5]. In the calculation of SDM, statistical models are used to fit the data. These statistical models are trained with environmental data and species occurrence data, extrapolating a prediction of where the species may exist, according to the environmental conditions similar to those where the species were sighted. Examples of statistical models used in the calculation of SDM are: Support Vector Machines (SVMs), Random Forest (RF), Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) [7].

Currently, to apply predictive models of species distribution, IMAR researchers access each source of species and environmental data separately, needing to perform data preparation manually, according to the type of data found in each source.

For the calculation of the SDM, researchers access three types of data separately:

- Environmental conditions (for example, from the World Ocean Atlas 2013 data source (WOA13[2])).
- Occurrence of species (for example, from the European Marine Observation and Data Network (EMODnet[3]) and *Ocean Biogeographic Information System* (OBIS[4]) data sources.
- Bathymetry (for example, from EMODnet data sources), which corresponds to the measure-

[1] http://www.horta.uac.pt/intradop/index.php/imar-topmenu-123

[2] https://www.nodc.noaa.gov/OC5/woa13/woa13data.html
[3] http://www.emodnet.eu/bathymetry
[4] http://www.iobis.org/

ment of the depth of the ocean floor and its relief.

Outside the scope of the SDM calculation, but to obtain more information about the species, they also access species taxonomy data (for example, from the World Register of Marine Species Data (WoRMS[5])).

Nowadays, for the calculation of a SDM, the researchers face three different challenges. The first is a data integration challenge. Since the data is distributed across several heterogeneous data sources, they require manual selections and different treatments, depending on their types. In this treatment it is necessary to select and treat the data manually, in order to prepare them for the application of the SDM. The second challenge is the automation of the data processing. Currently, the treatment process requires time to be done manually. Ideally, the processing of data from the various data sources would be automatic, taking as little time as possible.

The third challenge is the insertion of new data. As of now, every time the user wants to consider new data or new data sources for the calculation of a SDM, it is necessary to re-select and repeat the data preparation process. This insertion implies the repetition of all manual handling of the data.

In order to improve the process to compute a SDM, the objective is to design and build a software solution that automates the selection, preprocessing of the data, its integration and the application of SDM to data from several sources. In particular, this solution should allow the integration of data obtained from four data sources: EMODnet, OBIS, WOA13 and WoRMS. The solution must support insertion of new data entered by the users, which are stored in the application, in the format suitable for the SDM application.

This paper is organized as follows. Section 2 describes the research work made in order to understand the state of art. Section 3 describes the web application DeepData. Section 4 describes the validation of the web application DeepData. Section 5 concludes and presents future work.

2. Research work

The research work made for better understanding the theme and to build a solution is divided in two groups: (i) the study of the Azores sea and its characteristics and (ii) the study of existing integration systems.

For the study of the Azores sea and its characteristics two papers were particularly relevant: the paper (i) Seafloor Characteristics in the Azores Region (North Atlantic) [10] and the paper (ii) Predictive modeling of deep-sea fish distribution in the Azores [9].

The first paper characterizes the geomorphology and the type of marine soil of the Azores region, through the analysis of the environmental variables depth, slope, orientation, curvature, soil type and sediment thickness. It also presents an analysis of the correlation between the environmental variables temperature, salinity, chlorophyll concentration, carbon, oxygen (dissolved, saturated and used), phosphates, nitrates and silicates.

The second paper presents the results of a SDM that was used to predict the presence of deep-sea fish species with commercial value in the Azores. This paper was particularly valuable to understand the process for computation of a SDM used by the IMAR researchers nowadays.

The study of existing integration systems is divided in two groups: (i) polystore systems and (ii) materialized integration systems.

Polystore systems are a solution for the integration of Big Data. A polystore system follows a virtual data integration approach and is an integrated system that exposes multiple storage mechanisms, with queries that intercept multiple data models, through a single interface. The following polystore systems were studied: BigDAWG [4], Myria [12], Awesome [3], Tatooine [1] and ClouMdsQL [8].

These materialized integration systems were also studied: World Register of Marine Species (WoRMS[2]), a database for uniformity of species taxonomy; Yeastract [11], a database with information about the orgamism of the beer yeast; and RegulonDB [6], a database with information about the genes of the organism *Escherichia coli*.

3. DeepData Architecture and Implementation

The web application DeepData allows the execution of Species Distribution Models (SDMs) of species of the Exclusive Economic Zone (EEZ) of the Azores. The application integrates four data sources: *(i)* EMODnet, which provides species data of the Azores EEZ and bathymetry data of the oceans; *(ii)* OBIS providing species data from the Azores EEZ; *(iii)* World Ocean Atlas 2013 (WOA13), which provides data on oceanic environmental conditions of the oceans; and *(iv)* World Register of Marine Species (WoRMS) which provides species taxonomy data. The DeepData application was designed according to the requirements of a group of researchers from IMAR, Institute of MAR, with headquarters in the Department of Oceanography and Fisheries of the University of the Azores, who will be the end users of the application. The web application DeepData is composed of 5 modules: (i) Relational database, (ii) Data integration, (iii) Computation of SDM, (iv) In-

[5] http://www.marinespecies.org/

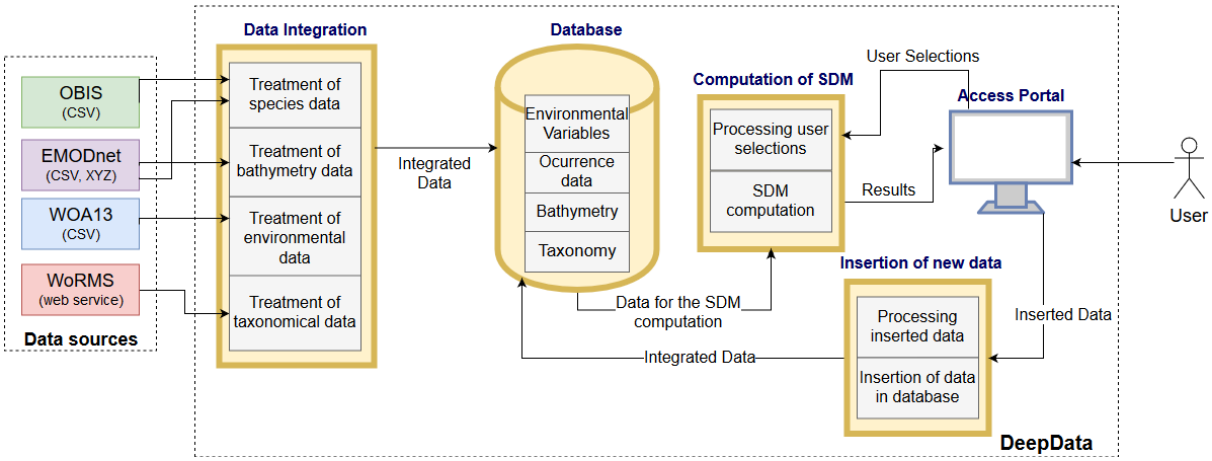

Figure 1: Architecture of web application DeepData.

sertion of new data and (v) Access Portal.

The Figure 1 presents the architecture of the web application DeepData.

3.1. Relational Database

The relational database stores four kinds of data: environmental data, species occurrence data, bathymetric data and taxonomic data. The database is composed of ten tables. The table of environmental data stores the environmental conditions temperature, salinity, silicate, nitrate, phosphate, density, conductivity, dissolved oxygen, oxygen saturation and oxygen utilization for each tuple of (latitude, longitude, depth, decade). It also stores the name of the data source from where the data came from. The table of species occurrence data stores the species names, their location, date of sight and the name of the data source. The table of bathymetric data stores the depth of each tuple (latitude, longitude) and the name of the data source. The taxonomic data is stored in a hierarchy of 7 tables. The relational model of the database is as follows. Primary keys are underlined and foreign keys are specified as FK.

Environmental_conditions (latitude, longitude, depth, decade, temperature, salinity, silicate, nitrate, phosphate, density, conductivity, dissolvedOxygen, oxygenSaturation, oxygenUtilization, source)

Occurrence (specie name, latitude, longitude, depth, decade, source, month, year)
latitude, longitude, depth, decade: FK (Environmental_conditions)

Bathymetry (latitude, longitude, depth, source)

Taxonomy_Kingdom (kingdom)

Taxonomy_Phylum (phylum, kingdom)
kingdom: FK (Taxonomy_Kingdom)

Taxonomy_Class (class, phylum)
phylum: FK (Taxonomy_Phylum)

Taxonomy_Order (order, class)
class: FK (Taxonomy_Class)

Taxonomy_Family (family, order)
order: FK (Taxonomy_Order)

Taxonomy_Genus (genus, family)
family: FK (Taxonomy_Family)

Taxonomy_Specie (specie, genus)
genus: FK (Taxonomy_Genus)

3.2. Data integration

The data integration module performs the data transformation and integration process. The data integration is carried out in a materialized way because the data does not suffer major changes in its sources. Therefore, in addition to the initial loading, later uploads due to updates of the data sources will not be very frequent. In the process of data integration four types of data were transformed and integrated: (i) species occurrence data, (ii) environmental data, (iii) bathymetric data and (iv) taxonomic data.

Species occurrence data The species occurrence data obtained from OBIS and EMODnet is available in two CSV files. The process of integrating and processing this data is carried out through the following 8 steps:

(i) Remove from the files the columns that have irrelevant content for the computation of a SDM. The remaining columns after this step are scientificname, yearcollected, monthcollected, longitude, latitude, minimumdepth e maximumdepth.

(ii) Remove from the files the lines that have missing fields. The remaining lines have all the fields needed for SDM computation.

(iii) Calculate average depth for each species occurrence. This calculation is necessary because, when the species data are crossed with the environmental data, there must be only one depth value per species occurrence, so that it is possible to obtain the value of the environmental variable at that depth.

(iv) Standardize species names, eliminating special characters.

(v) Add column with decade, considering the year of occurrence. This is necessary since the data used for calculating the SDM will be split by decades.

(vi) Standardize latitude, longitude and depth values. In this step, the depth, latitude and longitude values of species are standardized to be in agreement with the values of depth, latitude and longitude available for the values of the environmental variables. The depth values range from 0 meters up to 5500 meters, ranging from 5 meters up to 100 meters, from 25 meters up to 500 meters, from 50 meters up to 2000 meters, and from 100 meters up to and 5500 meters. Latitude values range from 32,125 to 43,875, with intervals of 0.25. Longitude values range from -34.125 to -21.875, with intervals of 0.25. These values of latitude and longitude correspond to the Exclusive Economic Zone of the Azores.

(vii) Remove duplicate species names, in order to insert them in the *Taxonomy_Specie* table of the database.

(viii) Insert species occurrence data in the table *Occurrence* of the database and species names in the table *Taxonomy_Specie* of the database.

Environmental data The environmental data obtained from WOA13 is available in 26 CSV files. These files have data about 10 environmental conditions (temperature, salinity, silicate, nitrate, phosphate, density, conductivity, dissolved Oxygen, oxygen Saturation, oxygen Utilization) between the years 1955 and 2012, separated by decades. The process of integrating and processing this data is carried out through the following 3 steps:

(i) Remove from each file the lines that do not belong to the EEZ of Azores. The remaining lines have latitude values ranged from 32,125 to 43,875 ans longitude values ranged from -34.125 to -21.875.

(ii) Enter default value (-9999) in incomplete fields (fields where a value for the environmental variable is not available for a given combination [latitude, longitude, depth]). The value '-9999' is used because, later on, when creating ASCII files for representing all points in the environmental region, it indicates that there is no value for the environmental condition at that combination (latitude, longitude, depth). After this step, the environmental data for the EEZ of the Azores is complete, having values (default or not) for all combinations (latitude, longitude, depth).

(iii) Insert environmental data in the table *Environmental_conditions* of the database.

Bathymetric data The bathymetric data obtained from EMODnet is available in 2 XYZ files. The process of integrating and processing this data is carried out through the following 4 steps:

(i) Remove from each file the lines that do not belong to the EEZ of Azores. The remaining lines have latitude values ranged from 32,125 to 43,875 ans longitude values ranged from -34.125 to -21.875.

(ii) Standardize latitude and longitude. Since the initial latitude and longitude values have 8 decimal places (and require only 3 to be crossed with the remaining environmental variable data and occurrence of species) rounding is done to three decimal places.

(iii) Enter default value (-9999) in incomplete fields, that is, in fields where a depth value for a given set of (latitude, longitude) is not available. This is done for the same reason as in the treatment of environmental data.

(iv) Insert bathymetric data in the table *Bathymetry* of the database.

Taxonomic data The taxonomic data from WoRMS is obtained through a web service. The process of integrating and processing this data is carried out through the following 3 steps:

(i) Invoke the web service with the names of the species to obtain their taxonomy. Initially, the web service is invoked for all species stored in the database. This gives the kingdom, phylum, class, order, family and genus for each species.

(ii) Standardize taxonomy, eliminating special characters from names.

(iii) Insert the taxonomic data in the database, according to their category. The tables for taxonomic data are: *Taxonomy_Kingdom*, *Taxonomy_Phylum, Taxonomy_Class, Taxonomy_Order, Taxonomy_Family, Taxonomy_Genus, Taxonomy_Specie*.

3.3. Computation of SDM

The module for computation of SDM receives the user's choices for the SDM computation and computes it. It consists of two components: the first component queries the database and processes the data for the SDM computation. The second component computes the SDM. This module works as follows:

Get average depth of occurrence A database query is made to the *Taxonomy_Species* table to search the average occurrence depth of the specie chosen in the access portal. This depth is standardized according to the depth values of environmental conditions.

Get environmental data A query is made to the database, to the *Environmental_ conditions* table, to search the values of the environmental conditions in the decade chosen in the access portal and in the depth obtained in the previous step.

Create ASCII grid files The environmental conditions data need to be transformed into an ASCII grid file to be interpreted by the statistical models. These files are used to represent the environmental conditions of the region for which the SDM is to be calculated, in this case the EEZ. In an ASCII grid file the first six lines indicate the properties of the zone they represent, such as the total number of latitude values, the total number of longitude values, the endpoints, the size of each cell, and the value for the "no data" existence (whose default value is '-9999'). The ASCII grid files created represent values of environmental variables between latitudes 32.125 and 43.875, with intervals of 0.25 and between longitudes -34.875 and -21.125, with intervals of 0.25. The created ASCII files thus represent the environmental conditions in the Azores EEZ.

Get species data A query is made to the database, to the table *Occurrence*, to search the occurrence data of the specie indicated in the access portal. With this data, a CSV file is created. It has as columns the name of the species, the latitude and the longitude at which it was sighted. This file represents the occurrences of the species.

Compute SDM In this step, the SDM is calculated by applying the statistical models selected by the user in the access portal. The SDM is calculated using the SSDM software package of the R language, a package for SDM calculation. The package performs the data preparation and computation functions of the models to the environmental and species data that were processed in the previous steps. As a result of the models computation, it produces a plot representing the probability of species occurrence at all the given point of coordinates (latitude, longitude), taking into account the environmental conditions that exist in that place, compared to the environmental conditions that exist in the places where the species was sighted. The statistical models for computing SDM available in the SSDM package are Generalized linear model (GLM), Generalized additive model (GAM), Multivariate adaptive regression splines (MARS), Generalized boosted regressions model (GBM), Classification tree analysis (CTA), Random forest (RF), Maximum entropy (MAXENT), Artificial neural network (ANN) and Support vector machines (SVM). However, given the small volume of data, some models do not perform correctly as is the case of CTA, GBM and MAXENT and, therefore, are not available for selection in the access portal.

3.4. Insertion of new data

Through the access portal, the user can insert new data and new data sources of environmental conditions and species in the application. The data insertion module receives data from environmental conditions and species occurrence entered by the user, through the access portal, and inserts them into the database. The environmental conditions data inserted must be contained in a CSV file with the fields: latitude, longitude, depth, decade, environmental variable value and data source name. The species occurrence data inserted must be contained in a CSV file with the fields: species name, latitude, longitude, month collected, year collected, depth and data source name.

3.5. Access Portal

Through the access portal the users choose the parameters for the computation of the SDM, and visualize the plot resulting from the SDM computation. The parameters available for the computation of the SDM include: the data sources, the specie, the environmental variables, the decade and the statistical models (Figure 2). These selections are sent to the computation of SDM module where they are processed and the statistical models are calculated.

After the calculation of the models, the plot generated can be accessed through the portal (Figure 3). The plot corresponds to the result of the application of the chosen models to the species and environmental conditions chosen. This plot represents the probability of species occurrence in the EEZ of the Azores, taking into account the environmental conditions existing in the places where

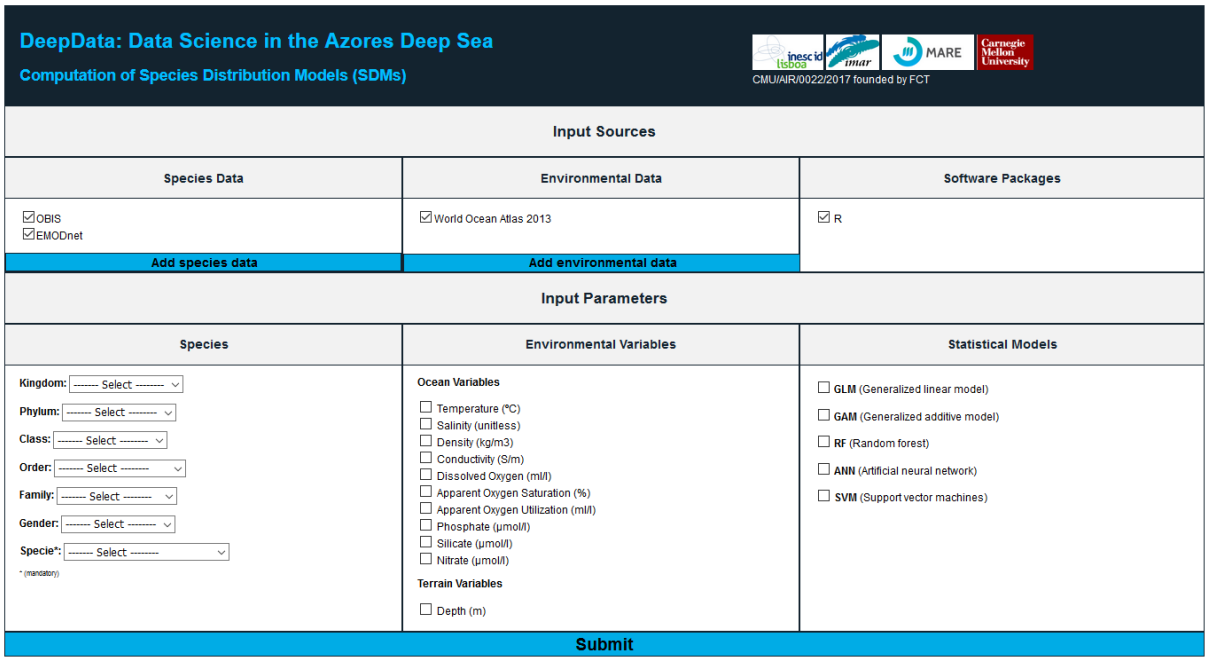

Figure 2: Access portal from web application DeepData.

the specie was sighted. Together with the result plot, two tables are presented: on the left, a table with the parameters chosen for the SDM calculation (specie taxonomy, statistical model and environmental conditions) and on the right, a table with the result of the statistical model evaluation. This evaluation presents the performance of execution of the statistical model and allows the verification of whether the suitability of the chosen model is suitable for the available data. From the performance of the model it is possible to infer the reliability of the SDM result plot. The metrics presented for model evaluation are: (i) AUC Area under the receiving operating characteristic (ROC) curve, which indicates the accuracy of an estimate. If AUC has a value of 1 it means that it is a perfect estimate, while values below 0.5 indicate that m is the estimate; (Ii) Kappa, which compares the accuracy obtained with the estimated accuracy. If kappa shows the value 1, then the accuracy is in total agreement. If there is no agreement between the accuracy of both metrics, the kappa value is 0. If kappa presents a negative value, it means that there is no effective relation between the two metrics; (iii) Sensitivity, which measures the true positives; (iv) Specificity, which measures the true negatives; And (v) the proportion of occurrences correctly predicted. If the model has not been calculated (due to insufficient environmental data or species occurrence), instead of the plot and the result of the model evaluation, a message is presented informing the user of what happened.

Through the access portal it is also possible to insert new species occurrence and environmental data.

3.6. Implementation

The data integration module is implemented in Python using the CSV library, to handle CSV files and python-mysql.connector, for insertion of data and queries to the MySQL database.

The module for computation of SDM was written using two programming languages: Python, to prepare the data to calculate the SDM, and R to calculate the SDM. The Python language is used to query the database (which returns the environmental conditions and the species occurrence), to create the ASCII grid file of the environmental conditions and the CSV file of the species occurrence. After the data is prepared, an R script is invoked. It calculates the SDM using the SSDM software package.

The implementation of the access portal is done using PHP, HTML/CSS and Javascript.

4. Validation and Discussion

The experimental validation of the web application DeepData validates the access portal and the module for computation of SDM. The validation involves the group of IMAR researchers as they will be the end users of the web application DeepData. The validation process consists of two phases: (i) execution of tasks in the DeepData application and (ii) completion of a usability questionnaire. The execution of tasks in the DeepData application allows the validation of the access portal. The usability questionnaire allows the validation of the computation of SDM module and the usability of the access portal.

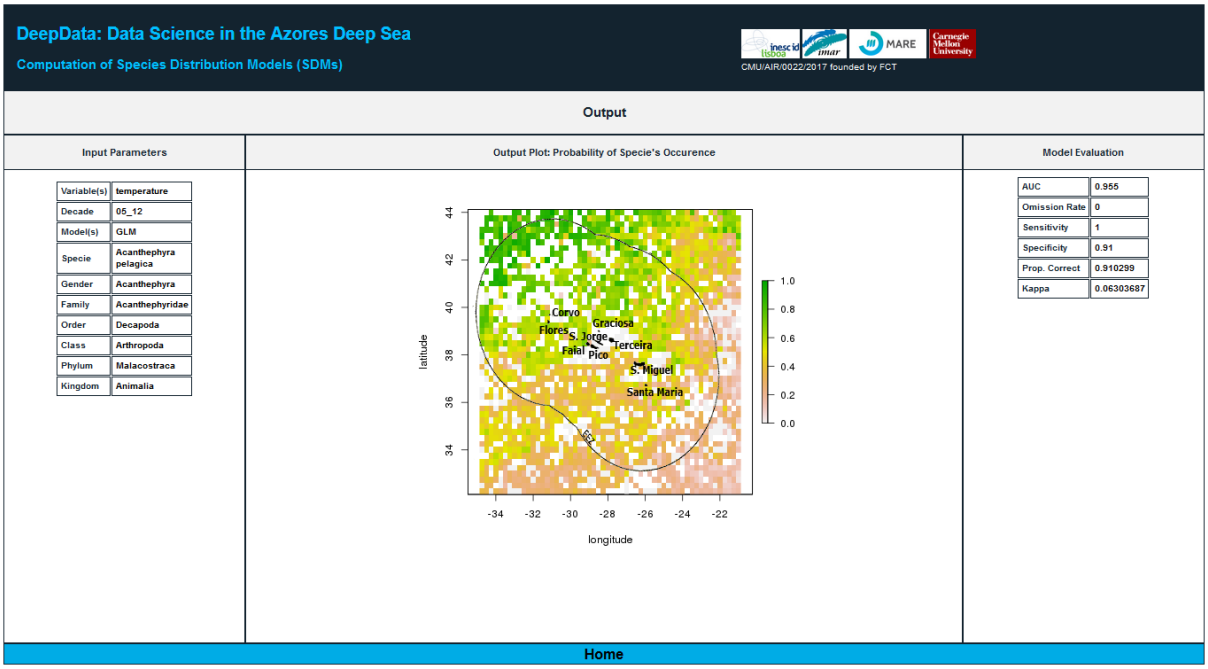

Figure 3: Access portal with result of SDM computation.

The validation of the DeepData application involved 10 IMAR researchers, 20% aged between 26 and 35 years, 60% aged between 36 and 45 years and 20% aged between of 46 and 55. 60% of the participants are male and 70% have a PhD degree, while the remaining have a master's degree. Their occupations are divided between biologists (20%), marine researchers (60%) and PhD students (20%). 70% of validation participants are familiar with the SDM calculation.

4.1. Access Portal Validation

The access portal was validated by measuring the execution time and the number of errors made while performing the tasks in the DeepData application. The tasks consist of: (i) running an SDM using one statistical model; (ii) execution of an SDM, using two statistical models; (iii) insertion of new species occurrence data. The metrics used allow the evaluation of the difficulty that users feel when computing an SDM. The usability of the access portal is evaluated through the usability questionnaire, namely through the questions that address the presentation of the interface, such as the color scheme, for example.

4.2. Computation of SDM Validation

The module of computation of SDM is validated through the usability questionnaire, namely through the questions that ask about the results of the SDM, the evaluation of the SDM, the parameters available for input and the utility of the system as a tool for calculating SDM.

4.3. Results

In the table 1 are the measured times and the count of the errors made during the execution of the tasks in the application DeepData. Measuring the time that users took to perform tasks in the DeepData application demonstrates that on average the calculation of an SDM (which corresponds to tasks 1 and 2) takes between 1 minute and 04 seconds and 1 minute and 18 seconds. During execution of the first task, no errors were performed. In the execution of the second task 3/10 errors were made, all of them because they did not select the data sources intended in the task.

As for entering new data (corresponding to task 3), users take an average of 2 minutes and 12 seconds, never exceeding 2 minutes and 50 seconds. In the execution of the third task 4/10 errors were executed, which corresponded to the poor formation of the inserted CSV file.

The result of the questionnaires is presented in the Table 2. The first two questions, which refer to the usability of the interface, present an average of positive responses, emphasizing the ease of interaction with the DeepData application. The remaining questions aim to validate the module of predictive models. Although all the questions present an average of positive answers, there are some answers with negative values, indicating that there may be improvements related to statistical models, environmental variables and the way of exposing the results of SDM. 80% of users of the DeepData application claim that they would use the SDM application but only 30% would use instead of their current method. Users who prefer to continue us-

Table 1: Time and number of errors in the execution of the validation tasks in the DeepData application.

	Task 1		Task 2		Task 3	
	Time (minutes)	Errors	Time (minutes)	Errors	Time (minutes)	Errors
	1.07	0	1.05	0	1.36	0
	0.5	0	1.23	0	2.5	1
	1.1	0	1.01	1	2.02	1
	1.24	0	1.14	0	2.14	0
	2.03	0	1.55	1	2.05	0
	1.31	0	0.56	0	1.55	0
	1.05	0	0.57	0	2.01	0
	1.4	0	1.07	0	1.33	1
	1.12	0	1.11	0	1.58	0
	1.11	0	1.19	0	1.5	0
	1.02	0	0.57	1	1.2	1
Average time (minutes):	1.18		1.04		2.12	
Minimum time (minutes):	0.5		0.56		1.2	
Maximum time (minutes):	2.03		1.55		2.5	

ing their current method, say that they would use the DeepData application as a tool for comparison and validation and that, with the insertion of new features such as the adjustment of the parameters of the models, verification of the collinearity between environmental conditions and the export of the SDM results, would also use the DeepData application with more confidence in its results.

4.4. Discussion

The time measurement results and the result of the number of errors during the execution of tasks are acceptable and demonstrate that users quickly understand how to interact with the DeepData application for SDM calculation and insertion of new data without any interaction or previous knowledge about the application.

Usability surveys demonstrate users' interest in the DeepData application and the ease they feel in interacting with the application. The results of the usability questionnaires are positive but there are new features that need to be incorporated into the DeepData application to improve the SDM results. Users say that adding new features that allow greater interaction in the calculation of statistical models, and later validation, would make the application more robust and reliable. Thus, features to be added in the application will allow to choose the parameters for the calculation of statistical models, insertion of new terrain variables, identification of colinearities between environmental conditions and computation of SDM only on the surface or ocean floor.

5. Conclusions

This paper presents the web application DeepData, an application for the SDM calculation of the Azores EEZ. The application DeepData aims to satisfy the need that a group of researchers of the IMAR manifested, guaranteeing an platform of integrated data of occurrence of species and environmental conditions, that allows the calculation species distribution models. The application allows the users to insert new data and data sources of species occurrence and environmental conditions. The DeepData application integrates four data sources: OBIS, WOA13, EMODnet and WoRMS. These data sources provide data on species, environmental conditions, bathymetry and taxonomy of the species of the Azores EEZ. The integrated data are stored in the relational database and then SDM are applied, according to the selections the user made in the access portal. The results of the application of the SDM are presented, in the access portal, in the form of a plot and a table containing the SDM evaluation. This paper also presents the validation of the web application DeepData, carried out by the IMAR researchers.

Table 2: Usability questionnaire responses. Responses with value 1 mean "completely disagree", responses with value 5 mean "completely agree".

	Responses					
Question	1	2	3	4	5	Average
The color scheme is adequate.	0	0	1	8	1	4
It is easy to understand what I have to select to calculate a SDM.	0	0	0	4	6	4.6
The environmental variables are adequate to calculate a SDM.	0	1	3	5	1	3.6
The models are adequate to calculate a SDM.	0	1	2	5	2	3.8
The result of the SDM is shown in a adequate way.	1	0	3	5	1	3.5
The model evaluation is useful.	1	0	2	6	1	3.6
The application has all the functionalities needed to calculate a SDM.	0	2	4	3	1	3.3

5.1. Future Work

The main current limitation of the web application DeepData is the lack of interaction, on the part of the user, with the parameters of the statistical models. Ideally, the user could adjust the parameters of the statistical models, according to the data used for the SDM calculation. In its current state, the DeepData application does not allow this type of interaction with the statistical models because the software package used to calculate SDM (SSDM package of the R language) does not allow much interaction with the parameters of the statistical models, working as an abstraction.

Based on feedback from the IMAR researchers about the DeepData application, it was possible to highlight the features that need to be developed in the near future. This features include:

- Allow the selection of environmental conditions only from the ocean floor.
- Allow the selection of environmental conditions only on the oceanic surface.
- Automatic refresh of species taxonomy data after insertion of new species data.
- Allow the choice of terrain variables such as slope.
- Allow the SDM results to be downloaded.
- In the case of calculating an SDM with several statistical models (ensemble), present not only the result of the calculation of the statistical models together but also the result of the SDM obtained by calculating each statistical model separately.
- Allow the selection of the parameters used in the statistical models.
- Allow colinearity to be verified between environmental conditions.

References

[1] R. Bonaque, T. D. Cao, B. Cautis, F. Goasdoué, J. Letelier, I. Manolescu, O. Mendoza, S. Ribeiro, and X. Tannier. Mixed-instance querying: a lightweight integration architecture for data journalism. *Proceedings of the VLDB Endowment*, 9(13), 2016.

[2] M. J. Costello, P. Bouchet, G. Boxshall, K. Fauchald, D. Gordon, B. W. Hoeksema, G. C. Poore, R. W. van Soest, S. Stöhr, T. C. Walter, et al. Global coordination and standardisation in marine biodiversity through the world register of marine species (worms) and related databases. *PloS one*, 8(1), 2013.

[3] S. Dasgupta, K. Coakley, and A. Gupta. Analytics-driven data ingestion and derivation in the awesome polystore. In *IEEE International Conference on Big Data*, (2016).

[4] J. Duggan, A. J. Elmore, M. Stonebraker, M. Balazinska, B. Howe, J. Kepner, S. Madden, D. Maier, T. Mattson, and S. Zdonik. The bigdawg polystore system. *Association for Computing Machinery (ACM) Sigmod Record*, 44(2), 2015.

[5] J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 2009.

[6] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muniz-Rascado, J. S. Garcia-Sotelo, K. Alquicira-Hernandez, I. Martinez-Flores, L. Pannier, and J. A. Castro-Mondragon. Regulondb version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic acids research*, 44, 2015.

[7] A. Guisan, T. C. Edwards, and T. Hastie. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157, 2002.

[8] B. Kolev, C. Bondiombouy, P. Valduriez, R. Jiménez-Peris, R. Pau, and J. Pereira. The cloudmdsql multistore system. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD)*. Association for Computing Machinery (ACM), 2016.

[9] H. E. Parra, C. K. Pham, G. M. Menezes, A. Rosa, F. Tempera, and T. Morato. Predictive modeling of deep-sea fish distribution in the azores. *Deep Sea Research Part II: Topical Studies in Oceanography*, 2016.

[10] A. D. Peran, C. K. Pham, P. Amorim, F. Cardigos, F. Tempera, and T. Morato. Seafloor characteristics in the azores region (north atlantic). *Frontiers in Marine Science*, 3, 2016.

[11] M. C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A. R. Fernandes, N. P. Mira, M. Alenquer, A. T. Freitas, A. L. Oliveira, and I. Sá-Correia. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic acids research*, 34, 2006.

[12] J. Wang, T. Baker, M. Balazinska, D. Halperin, B. Haynes, B. Howe, D. Hutchison, S. Jain, R. Maas, P. Mehta, et al. The myria big data management and analytics system and cloud services. In *8th Biennial Conference on Innovative Data Systems Research (CIDR), Chaminade, CA, USA*, 2017.