Automatic Procedures to Assist in Manual Review of Marine Species Distribution Maps

Gianpaolo Coro[1], Pasquale Pagano[1], and Anton Ellenbroek[2]

[1] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR Pisa, Italy,
{gianpaolo.coro,pasquale.pagano}@isti.cnr.it,
[2] FAO - Food and Agriculture Organization,
anton.ellenbroek@fao.org

Abstract. Ecological Niche Modeling (ENM) is a branch of biology that uses algorithms to predict the distribution of species in a geographic area on the basis of a numerical representation of their preferred habitat and environment. Algorithmic maps can be produced for suitable or native habitats and require a review by human experts. During the review operation biologists use their knowledge about a species to modify the maps. They usually take algorithmic maps as starting point in the review. In this paper we provide a methodology for biologists to use the automatic maps as references also during and after the review process. Our approach is based on a comparison between the reviewed map and two systems: an expert system and a Feed Forward Neural Network. Furthermore we suggest an evaluation procedure of the quality of the environmental features used as training set, for assessing the models reliability.

Keywords: Ecological Niche Modeling, Feed Forward Neural Networks, AquaMaps, Maps Review

1 Introduction

The term Ecological Niche Modeling (ENM) refers to a set of methods that use algorithms to predict the distribution of species in a geographic area. ENM techniques focus on a numerical representation of a species habitat, which should be complete and made up of independent variables. Several approaches have been used for this task, ranging from physiological to mathematical models [16]. Results are usually produced in the form of GIS heat maps, where a color gradient represents a range of probabilities and hotter colors correspond with higher probabilities. These maps can be generated by automatic procedures using different kinds of approaches [15]. Automatically generated maps are usually reviewed by a biologist in order to correct inconsistencies. The biologist uses his/her knowledge about the species to make little corrections or to completely redesign the distribution map. Such corrections can consist of adjusting models parameters or in manually editing distribution values.

In this article, we propose a methodology that can help biologists in discovering gaps in their knowledge when reviewing a map. They can investigate

inconsistencies and revise some assumptions made during a review. Furthermore, the method proposes statistical analysis techniques for validating the quality of the map. The method has been applied to marine species. It is based on a comparison between the reviewed map and two systems: an automatically generated distribution based on Feed Forward Neural Networks, and an expert system (Aquamaps [12]) for marine species distribution prediction. A features processing technique, based on Principal Component Analysis, is then used for assessing the reliability of the models and consequently of the reviewed map.

In Section 2 we report an overview of common approaches for producing species maps. In Section 3 we describe the methodology and propose an example of application. In Section 4 we draw the conclusions.

2 Overview

Ecological Niche Modeling is usually a complex and iterative process including [8]: (*i*) identification of *relevant data*, (*ii*) *modeling*, (*iii*) *projection* of predictions onto a geographic space. The first step is crucial and usually involves the identification of environmental features related to species preferences. In this paper we don't focus on features selection as our method is applied after the projection step. For what regards modeling, techniques are usually based on occurrence records (*presence points*), i.e. places where the species has been observed in its habitat. Some approaches need even to use *absence points*, i.e. locations where the environment is considered unsuitable for a species [9]. Models need representative occurrence data and independent and complete environmental parameters. These possibly give robustness and reliability to the models [11,8]. The choice of a suitable modeling technique for a specific scenario is not trivial. There is not a general pattern to follow when designing an ENM experiment: each species can be specific in terms of habitat and presence/absence information. In [8] the authors report several applications and eventually indicate possible directions for producing robust models. Such directions involve (*i*) improvement of methods for modeling presence-only data, (*ii*) accounting for biotic interactions and (*iii*) assessing model uncertainty. Similar advices come from the BAM diagram for Biotic Interactions described in [17]. Other issues involve the choice of the kind of modeling technique to apply: a model could try to explicitly catch the behavior of a species and its physiological limits and tolerances [4] (*mechanistic* approaches). Otherwise it could automatically extract the correlations between the environmental features vectors and the species presence (*correlative* approaches) [16]. In our methodology a scientist is suggested to use a particular algorithm for simulating the opinion of another biologist on the reviewed map. This algorithm uses a partial mechanistic approach that involves expert knowledge about the species. On the other side we use a correlative technique for generating maps according to different inputs sets.

2.1 ENM Algorithms: Aquamaps and Feed Forward Neural Networks

Several tools allow scientists to produce maps by applying Niche Modeling algorithms [19,5]. In [16] a big collection of techniques is presented along with the kind of scenarios these should be applied to. Artificial Neural Networks (ANNs), in particular, have demonstrated to gain high performances when absence and presence points are available for the species to model. ANNs implement a correlative approach as they try to automatically simulate the probability for a species given certain environmental conditions. Also other models, like SVMs, have gained very good results in ENM [7], but we chose to rely on the advices in [16] and to use ANNs in our method at first stage.

In [18] the authors report a comparison among the most used correlative approaches applied to some marine species of commercial and scientific interest. Among these, the Aquamaps algorithm [13,12] is a presence-only species model, that allows the incorporation of expert knowledge about the species habitat. The Aquamaps distributions are generated using information about species preferences on environmental properties like depth, salinity, temperature, primary production, distance from land and sea ice concentration. Maps are produced at 0.5 degrees resolution. The expert knowledge is used in modeling the habitat parameters and the species preferences. The environmental features values are manually edited before applying the model, then a trapezoidal function is traced for each of them. This function represents the species 'preferred' values for that parameter and can be automatically produced by processing the values ranges associated to the presence points. In particular, the trapezoid is traced on 4 values called *minimum*, *preferred minimum*, *preferred maximum* and *maximum*. These values are calculated, for each parameter, by a rule-based procedure [12] using percentiles of the values observed at the presence points. In some cases the trapezoid can be manually traced by a biologist. The probabilities are produced by multiplying the values of the functions at a certain 0.5 degrees cell in the oceans. Aquamaps presents mechanistic assumptions combined with automatic estimation of parameters values. After the model projection, a scientist can review a map by manually changing the trapezoidal curves or by modifying the values in the produced distribution table. Aquamaps is a reference algorithm for marine species distribution modeling, as it gains high performances if compared to other purely automatic procedures [6]. For such reason we used it for simulating the point of view of another biologist in our method.

2.2 Features Analysis: PCA and HRS

Features analysis is crucial in ENM. A preliminary processing of the features vectors constituting the training set could highlight useless features or could evaluate the potential robustness of the models to produce. One of the most used techniques is the Principal Component Analysis (PCA) [10] a mathematical procedure that aims to reduce the number of dimensions of the features

space. PCA uses an orthogonal transformation in the features space for producing independent variables called principal components. This transformation can be useful for investigating the correlations among the environmental features used in ENM. Adding more dependent variables, in fact, usually does not result in better models. PCA is not specific to biological applications, but other topic-oriented transformations can rely on it. The Habitat Representativeness Score (HRS) [14] is an algorithm based on PCA that applies to marine species environmental features. It measures how much representative sampled habitats are for a certain area of study. HRS has been used for assessing the minimum number of surveys on a study area that are needed to cover a good heterogeneity of species habitat variables. HRS can be applied to two datasets of environmental features, one representing a sampled area and the other a geographical region of interest. A score is produced for each feature, ranging from 0 to 2, with 2 representing completely non-overlapping distributions of values. The lower the HRS the more similar data obtained from a survey to the study area. In this paper we show how HRS can be useful for investigating the robustness of trained models. We applied HRS for assessing how much the features associated to species occurrence points represent a projection area.

3 Methodology

The proposed methodology aims to assist biologists in the evaluation of their manually created maps. The basic assumption is that biologists use their expert knowledge about the species spatial distribution and environmental preferences. This knowledge could be general or specific about some place in which the species lives. In the case of local knowledge it could be that the training set of a model contains indications about a disjoint location. In this case, the biologist and the automatic system start from different assumptions, referring to completely dissimilar environments. This can be generalized by considering the expert knowledge and the automatic system's training set as two sets of multi-dimensional points, where each dimension refers to a separate environmental feature. This leads to one of the following: (*i*) the biologist's knowledge includes the training set, (*ii*) the biologist's knowledge and the training set are disjoint, (*iii*) the biologist's knowledge is completely included in the training set.

The first part of our methodology aims to suggest to biologists in which of these cases the reviewed map could fall. For such aim, they could follow these 6 steps:

1. produce an Aquamaps distribution for the species;
2. review the distribution and produce a *reviewed map*;
3. variate the training set of a Feed Forward Neural Network in order to simulate a spot knowledge or a wide knowledge about the species. Take the best performing network topology in each case;
4. perform a numeric comparison between the reviewed map and the Neural Network maps;
5. use a statistical quality analysis for evaluating similarities among the maps;

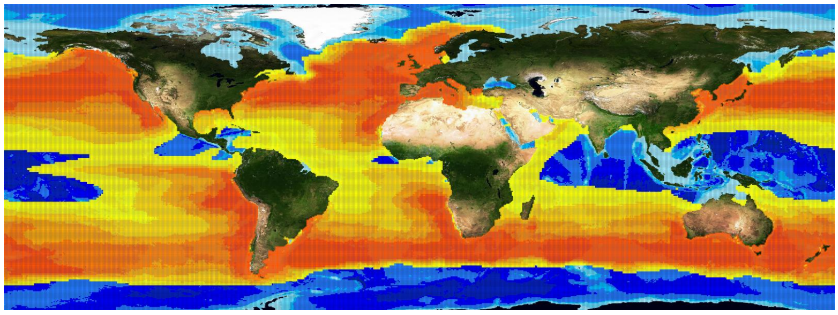

Fig. 1. Manually reviewed Aquamaps distribution for the Basking Shark taken from the AquaMaps website [2].

6. compare the maps with the Aquamaps distribution as this was the opinion of another biologist.

The second part of the proposed methodology then assesses the reliability of the automatic maps. In order to produce a robust model, the training set has to cover the projection area in terms of the variations in the environmental features values. For example if a Neural Network was trained on a partial set of knowledge (e.g. some points in the Mediterranean sea) but must be projected on a totally different environment (e.g. the Arctic Ocean), then its performances are likely to be unreliable. If a biologist finds the reviewed map to be similar to a possibly unreliable map then he/she might consider to revise it. The final steps in our method can be resumed as follows:

1. apply the HRS to the training sets used for the Neural Networks respect to the projection area;
2. evaluate, numerically, which is the most representative training set;
3. consider the reliability of the reviewed map by looking at the nearest correlative map, the representativeness of its training set and the simulated opinion of another expert.

In the next subsection we show a use case in which the described methodology has been applied to a manually created map for the Basking Shark.

3.1 Use Case

We performed an experiment on the Basking Shark (*Cetorhinus maximus*) marine species starting from the manually created map reported in figure 1. The AquaMaps website [2] provides (*i*) 449 occurrence points, (*ii*) a manually reviewed map and (*iii*) a map generated by the Aquamaps Suitable algorithm [12]. We used the Aquamaps Suitable distribution as a reference to simulate the opinion of another biologist. Figure 2 depicts the presence data distribution (*Presences* set), while figure 3 depicts the Aquamaps Suitable distribution. The maps illustrate that the Aquamaps and the reviewed map are very different.

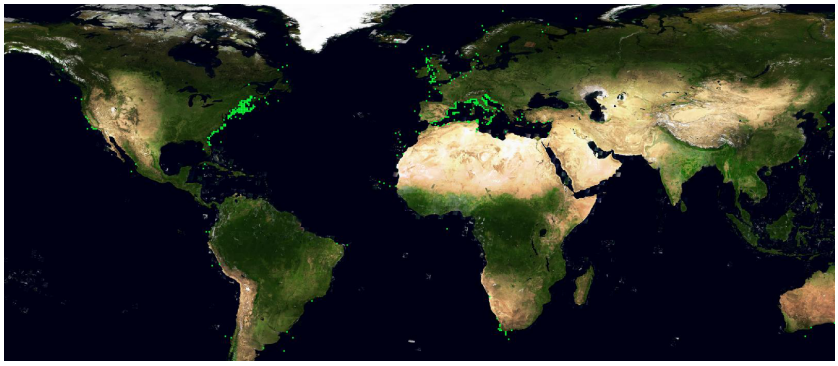

Fig. 2. Presence points for the Basking Shark, provided by Fish-Base [3] and AquaMaps [2].

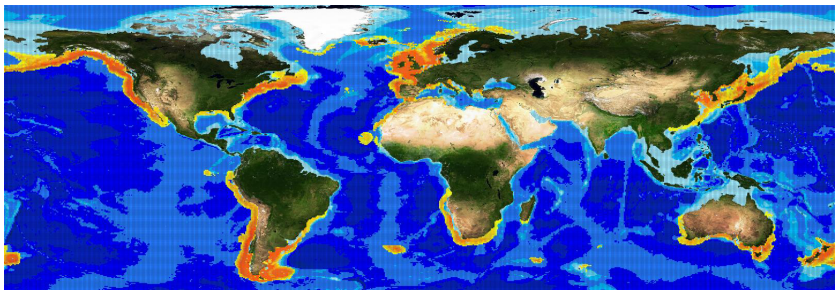

Fig. 3. Aquamaps Suitable range for the Basking Shark.

	Accuracy	Sensitivity	Specificity
NN Dense	64%	96%	33%
NN Sparse	72%	49%	95%

Table 1. Performances of the Neural Networks trained on Dense and Sparse Absences.

Following the approach described in Section 3, we trained a Feed Forward Neural Network with the same environmental information used by the Aquamaps Suitable algorithm. The inputs were vectors of 10 real numbers reporting (for 0.5 degrees oceans cells): the minimum, maximum and mean depth, the mean annual water values for salinity, bottom salinity, surface temperature, bottom temperature, primary production, distance from land and sea ice concentration. Thus, the network had 10 input neurons and 1 output neuron, returning real numbers ranging from 0 to 1. Hidden layers were necessary because the function to be simulated by the network was not linear.

We decided to simulate two variations in the knowledge about the species, by selecting two sets of 449 absence points: (*i*) a *Dense Absences* set, where absence points were close together, (*ii*) a *Sparse Absences* set, where absence points were

Fig. 4. Dense and Sparse Absences sets for the Basking Shark.

widely distributed in the oceans. Figure 4 shows the Dense Absences set (left) and the Sparse Absences set (right). Absence points were necessary because in our case Neural Networks required both positive and negative cases. We chose to variate the absence points sets because presence points are usually reliable information. Reliable absences are much more difficult to find, and it is likely that biologists rely on undocumented knowledge about them. Absence points were not available for the Basking Shark on [2], thus we used the reviewed map for generating them. We took locations with probability lower than 0.1 and higher than 0 to simulate places where the Basking Shark has very low probability of occurrence. We trained a Feed Forward Neural Network on the Presences and Dense Absences sets (*NN-Dense*), and another one on the Presences and Sparse Absences sets (*NN-Sparse*). In the training sessions we changed the number of hidden layers and of neurons in each layer. We adopted a *growing* approach, in which we added neurons and layers as far as the performances on the test set increased. Eventually we took the best performing topologies. We wanted the evaluation to be independent on other maps, therefore we used the 80% of the points for training and 20% for testing. The sets were not overlapping and the best performing model was chosen on the basis of the accuracy on the test set. The best NN-Dense model had 2 hidden layers with 100 neurons in the first layer and 2 neurons in the second, while the best NN-Sparse model had 1 hidden layer with 300 neurons. Table 1 reports the performances of the two models. The accuracy indicates the correct classification rate, the sensitivity the true positives fraction and the specificity the true negatives fraction. The accuracies are comparable, but while the NN-Dense model fits better to true positive classifications, the NN-Sparse prefers true negative classifications.

Figure 5 depicts the NN-Dense model projected on the world oceans, while figure 6 shows the map by the NN-Sparse model. The NN-Dense projection is widespread while the NN-Sparse one seems more similar to the Aquamaps Suitable distribution. We evaluated the similarities among the maps by considering as *true positives* the points in which the reviewed map reported at least a 0.8 probability and as *true negatives* the less than 0.1 probability points. The accuracy of NN-Dense respect to the reviewed map was 96.08% while NN-Sparse gained 66.75%. Taking the Aquamaps Suitable map as reference, NN-Dense gained 82.41% while NN-Sparse gained 93.71%. This evaluation can indicate that the NN-Dense model is more similar to the reviewed map while NN-Sparse is more similar to the Aquamaps Suitable map. For biologists this can be crucial information. The method indicates that the reviewed map is similar to another

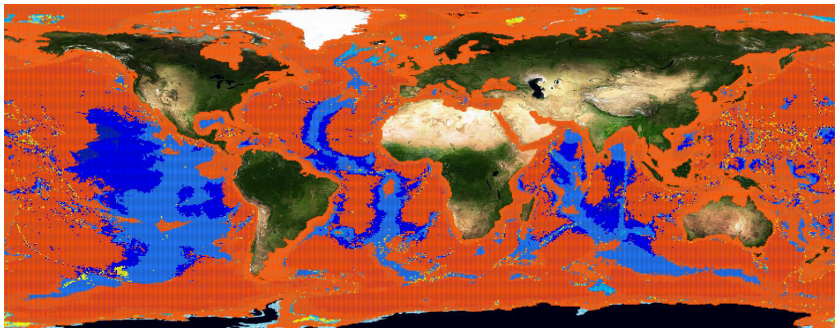

Fig. 5. Neural Network Dense model projected on the world oceans.

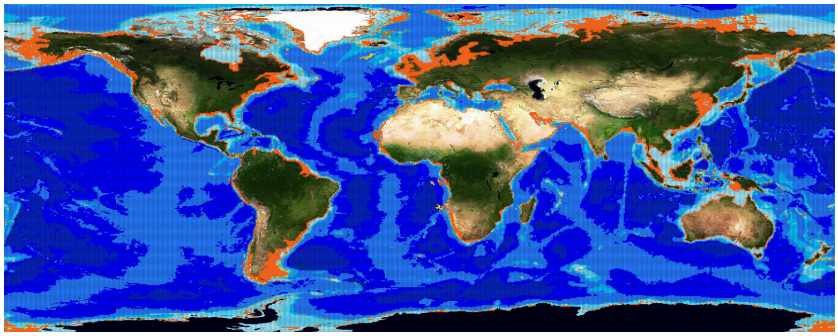

Fig. 6. Neural Network Sparse model projected on the world oceans.

	HRS on world oceans
Presences + Dense Absences	4.49
Presences + Sparse Absences	3.39

Table 2. Overall Habitat Representativeness Score between the projecting area and the training sets.

one generated by an automatic model that uses only limited and local knowledge. A wider knowledge is instead in agreement with an expert system (the Aquamaps Suitable distribution). At this point the biologists choose if the reviewed map has to be revised or if both the expert and the automatic systems are wrong.

As second step, we applied the Habitat Representativeness Score (HRS) technique between the projected area and the presences and absences sets. An overall score was calculated by summing the HRS scores of the involved features. This is different respect to the original procedure. In [14] the author suggests to weigh each score by the inverse of the eigenvalues of the PCA transformation. We avoided this inverse weighting because in our case (i) the eigenvalues depended on the ordering of the vectors used for calculating the PCA, (ii) we kept all the

principal components, because we considered all the features as independent and equally important, (*iii*) the HRS scores of the parameters were commensurable and an inverse weighting would have given too much importance to less variable dimensions. Table 2 reports the overall HRS of the projection area respect to two combinations of points. As we used 10 environmental features, the maximum (worst) HRS was 20, and the minimum (best) was 0. The combination of Presences and Sparse Absences offer the best performances but according to [14] the score is still high. The maps generated by the Neural Networks might not be robust because the training set does not represent the projection area. On the other side this does not mean they are performing badly, as the scores on the test sets are not low (table 1). At this point the technique offers a more complete scenario about the review: the manual map is similar to a map trained on a local knowledge which is not robust. Furthermore a simulated expert opinion agrees on using a wider knowledge. It is then on the biologists' side the decision to revise or to confirm the reviewed map.

4 Conclusions

We have described a method for using automatic techniques to evaluate a manually reviewed map for a species distribution. The method aims at scientists who want additional tools to revise their maps. We used Feed Forward Neural Networks to simulate different scenarios and calculated the similarity between the processed and the manual maps. By highlighting the similarities a biologist can quickly notice inconsistencies in the manually reviewed map. Furthermore, we used an expert system to simulate the point of view of another scientist that could be taken into account. Finally, by using a PCA based technique we evaluated the reliability of the automatic models by indicating the completeness of the training set. In the Basking Shark experimental case, we demonstrated that the reviewed map was similar to a map trained on a local knowledge and generated by an incomplete training set. Furthermore, by using a more distributed training set with the same dimension, the resulting map was more similar to an expert system's one. This methodology has been adopted in the i-Marine project [1] and is currently used on its stored data sources. Our future activity will concentrate on analyzing the application of our methodology by real users in order to collect more examples and to extend it. Feed Forward Neural Networks may be substituted by better performing models [7] or by *presence points-based* models [18]. Moreover, the current approach can be combined with additional analytical techniques that can reveal further qualitative information about the feature set, and thus improve the reliability of the maps.

Acknowledgments The reported work has been partially supported by the D4Science-II project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019) and by the i-Marine project (FP7 of the European Commission, INFRASTRUCTURES-2011-2, Contract No. 283644).

References

1. The i-Marine European Project (2011), http://www.i-marine.eu
2. Aquamaps Website (2012), http://www.aquamaps.org
3. Fish Base: Searchable global database (2012), http://www.fishbase.org
4. Chuine, I., Beaubien, E.: Phenology is a major determinant of tree species range. Ecology Letters 4(5), 500–510 (2008)
5. Coro, G.: Ecological modelling library for gcube vre. Software (2011), [Sofware] Release 1.0.0 , 18 May 2011.
6. Corsi, F., de Leeuw, J., Skidmore, A.: Modeling species distribution with gis. Research Techniques in Animal Ecology. Columbia University Press, New York pp. 389–434 (2000)
7. Drake, J.M., Randin, C.: Modelling ecological niches with support vector machines. Journal of Applied Ecology 43(3), 424 – 432 (2006)
8. Elith, J., Leathwick, J.: Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40, 677–697 (2009)
9. Guisan, A., Zimmermann, N.E.: Predictive habitat distribution models in ecology. Ecological Modelling 135(2-3), 147 – 186 (2000)
10. Jolliffe, I.: Principal component analysis. Wiley Online Library (2005)
11. Kamino, L., Stehmann, J., Amaral, S., De Marco, P., Rangel, T., de Siqueira, M., De Giovanni, R., Hortal, J.: Challenges and perspectives for species distribution modelling in the neotropics. Biology letters 8(3), 324–326 (2012)
12. Kaschner, K., Ready, J.S., Agbayani, E., Rius, J., Kesner-Reyes, K., Eastwood, P.D., South, A.B., Kullander, S.O., Rees, T., Close, C.H., Watson, R., Pauly, D., Froese, R.: AquaMaps: Predicted range maps for aquatic species. `http://www.aquamaps.org/` (2008)
13. Kaschner, K., Watson, R., Trites, A.W., Pauly, D.: Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. Marine Ecology Progress Series 316, 285–310 (July 2006)
14. MacLeod, C.: Habitat representativeness score (hrs): a novel concept for objectively assessing the suitability of survey coverage for modelling the distribution of marine species. Journal of the Marine Biological Association of the United Kingdom 90(07), 1269–1277 (2010)
15. Owens, H.L., Bentley, A.C., Peterson, A.T.: Predicting suitable environments and potential occurrences for coelacanths (latimeria spp.). Biodiversity and Conservation 21, 577 – 587 (2012)
16. Pearson, R.G.: Species distribution modeling for conservation educators and practitioners. (2012), synthesis. American Museum of Natural History. Available at http://ncep.amnh.org.
17. Peterson, A., Soberon, J., Pearson, R., Anderson, R., Martinez-Meyer, E., Nakamura, M., Araujo, M.: Ecological Niches and Geographic Distributions (MPB-49), vol. 49. Princeton University Press (2011)
18. Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E., Kullander, S., Froese, R.: Predicting the distributions of marine organisms at the global scale. Ecological Modelling 221(3), 467 – 478 (2010), `http://www.sciencedirect.com/science/article/pii/S030438000900711X`
19. de Souza Muñoz, M.E., De Giovanni, R., de Siqueira, M.F., Sutton, T., Brewer, P., Pereira, R.S., Canhos, D.A.L., Canhos, V.P.: openmodeller: a generic approach to species' potential distribution modelling. GeoInformatica 15(1), 111–135 (2011)