# EO BIG DATA ANALYTICS FOR THE DISCOVERY OF NEW TRENDS OF MARINE SPECIES HABITATS IN A CHANGING GLOBAL CLIMATE

*Sabeur, Z. A*[1]*; Correndo, G.* [1]*; Veres, G.* [1]*; Arbab-Zavar, B.* [1]*; Neumann, G*[1]*. Ivall, T*[1]*; Castel, F.* [2]*; Zigna, J-M.*[3]*; Lorenzo, J.*[4]

[1]University of Southampton IT Innovation Centre, UK.
[2]Atos France, Toulouse, France.
[3] Collecte Localisation Satellite, Toulouse, France.
[4]Atos Spain, Madrid, Spain.

## ABSTRACT

Climate change has been observed using multiple methods of Earth Observation (EO) including in situ, air-borne and space-borne sensing methods. These use multi-modal observation platforms, with various geospatial coverages, spatio-temporal resolutions and accuracies. The resulting EO Big Data from heterogeneous sources constitute valuable sources for scientists to investigate on the manifested responses of natural species behaviour to climate change. In the EO4wildlife[1] research project, we have access to Copernicus and Argos EO Big Data for conducting studies on the changes of habitats for a variety of marine species. The challenge is to discover causality of Metocean environmental observations and their relationship with the changing habitats of species. Nevertheless, there is a need to deploy Big Data technologies for connecting, ingesting, processing of EO data, as well as implementing specialised open data analytics services in this study. The particular services shall be made accessible to the scientific community for setting up modelling scenarios concerning the potential discovery of new trends of marine species habitats due to climate change. Three marine species are being studied in the EO4wildlife project. They include the Bluefin Tuna in the Atlantic-Mediterranean migratory regions, the black-footed albatross seabirds across the sub-tropical Atlantic Ocean and Loggerhead sea turtles along the North West coast of the African continent and Cape Verde. Large data representing geospatial migratory tracks and settlements of these respective marine species have been acquired in the project over period of times together with Metocean EO data from Copernicus and Argos satellites. These are currently analysed and modelled with a set of features obtained by searching in a large space of possible measured and derived Metocean parameters. A two-step search was used involving significance measurement and an iterative breadth first search based wrapper type feature selection algorithm. Furthermore, the analysis is useful for improving the performance of our habitat prediction models across the three marine species in the study. The discovery of new habitats geospatial and temporal trends which may be associated to the changing climate under these analyses will be achieved through the deployment of web-enabled data mining and analytics open services. A dedicated Big Data platform supported by generic data management services in the cloud is therefore deployed for assuring the scalability of the data processing and analytics services.

***Index Terms—*** Big Data, Earth Observation, Copernicus satellite, Climate change, habitat modelling

## 1. INTRODUCTION

EO4wildlife brings large number of multidisciplinary scientists such as marine biologists, ecologists and ornithologists around the world to collaborate closely together while using European Sentinel Copernicus Earth Observations more efficiently [1]. In order to reach such important capability, an open service oriented platform with an interoperable toolbox, that is compliant with OGC standards and supported by scalable cloud infrastructure is being implemented. The EO4wildlife platform offers dedicated open services that enable scientists to connect to marine species tracks databases and Big EO data in order to run habitat modelling simulations under a scalable processing environment. In particular, the platform enables the full integration of Copernicus sentinel data, ARGOS archive databases and animal track databases which can be effectively mined and fused for advanced big data analytics concerning the discovery of new trends of animal behaviour in the marine environments.

## 2. OPEN SERVICE ARCHITECTURE FOR BIG DATA MANAGEMENT

The EO4wildlife platform is composed of various functional components: 1- An internal data catalogue for aggregating geo-referenced products from external heterogeneous sources; 2- An ingestion module that allows the retrieval of data for exploitation by the platform services and; 4- A service Manager with which developers and/or data scientists manage the life cycle and execution of deployed services. Finally, the platform has built-in visualization features for the

[1] http://eo4wildlife.eu/

resulting geographic data from the processing services. The service management mechanism in the Big Data infrastructure is built on the containerization concept (i.e. Docker) which allows to encapsulate each service into an independent component that can be easily deployed on the cloud. An orchestration technology (i.e. Kubernetes) is used to manage container life cycle so that the underlying infrastructure becomes totally transparent [2].

## 3. BIG EO DATA ANALYTICS

In order to provide proofs of concept of the EO4wildlife platform and its dedicated Big EO data analytics services, a number of scenarios on habitat modelling for marine species behaviour are being developed. These required a pre-processing and analysis of the acquired big data for the discovery of strengths and relationships between data features prior to achieving efficiently performing models.

### 3.1. Big Data Features Selection

Features selection is the process of selecting the most dominant and connected variables or features for modelling environmental processes. Although initially there is only a small set of features (e.g. 8 features in the case of the pelagic fish use case) derived features such as gradients over time, averages over time and gradient over horizontal and vertical space are important to consider as they are related to the physical dispersion of nutrients and other hydrodynamic transport processes that take place within the marine environment [3]. In this case a genetic algorithm is used to search in the space of potential feature subsets. For each subset of selected features, ecological envelopes based on percentiles that the algorithm chooses and combines it into trees using "AND" and "OR" logical assertions are discovered. This process is performed stochastically and repeatedly so that a good number of possible subsets (therefore models) can be explored and trialled on the training set. Prior to this step a systematic search to find the best granularity for each derived feature (e.g. establishing whether temporal gradient for a given feature should be on a 10 or 30 day scale) is also conducted. This big data features selection process aims at optimizing the feature set to be used for best niche modelling the relationship between EO data and processes with trends on observed animal presence in space and time.

### 3.2. Habitat Modelling

Habitat niche modelling is a method for discovering and modelling the link between where the animal has been found (presence) and the environmental conditions at those points. These methods give an indication of the conditions which are favourable for the animal. Similarly, where the animals have not been found (absences) give an indication on the conditions that are not suitable for the animals. Given a model of climatic changes that forecasts metocean environmental conditions, a habitat model for given species can be used to predict how the boundaries of its habitat do change due to such environmental conditions. One of the most concerning results of climate change is the vulnerability of habitats of certain species. Other problems may include rapid shifts in the spatial positioning of these habitats which can have severe consequences for less mobile species. In order to visualise the animal tracks as they evolve in time, and compare the distribution of metocean observations where the animals have been detected, a working demonstrator is being developed (see Figure 1). The demonstrator allows users to integrate and explore different types of data under a single user interface.

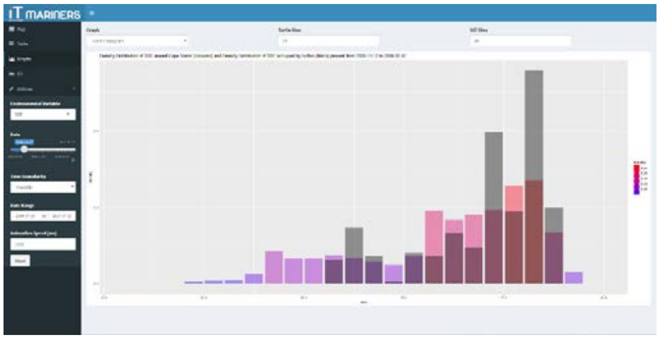


**Figure 1** *Spatial density distribution of sea turtles versus sea surface temperature environmental Observations*

#### *3.2.1. Habitat Modelling for Atlantic Bluefin Tuna (ABFT)*

An Ecological Niche Modelling (ENM) framework which uses using observed animal presence data (animal tracks) has been developed for predicting probabilities of *Potential Habitats*. Specifically, monthly ENMs on *Potential Habitat* predictions of ABFT in the Mediterranean Sea were developed. The most relevant Earth Observation (EO) variables which influence habitat preferences were also identified [4]. These include Bathymetry, Sea Surface Temperature (SST), Chlorophyll (CHL), $CO_2$ Net Primary Production (NPP), Sea Level Anomalies (SLA) and Eddy Kinetic Energy (EKE). Environmental Envelopes (EE) were calculated during the model training stages for each variable through using pre-defined bounds. During the testing stage, geospatial areas of interest in the Mediterranean Sea were analysed with [0.1 x 0.1] degrees grid resolution. Each grid cell was set up to unity (*Potential Habitat = 1)*, if for example, the sampled EO variables at the grid cell satisfies some specific environmental conditions, such as:

$$CHL\_min \leq CHL_{(i)} \leq CHL\_max,\ SST\_min \leq SST_{(i)} \leq SST\_max$$

The model for predicting *Potential Habitat(0/1)* is simply defined as follows:

$$Bathy_{range}\ (0/1)*SST_{range}(0/1)*CHL_{range}\ (0/1)*NPP_{range}\ (0/1)*EKE_{range}\ (0/1)$$

As a result, 99% percentiles for EE bounds were obtained. (See Table 1). *Proportion of Sea* notes the fraction of the spatial region that was classified as Habitat. *Number found* is the number of observed relocations that are considered as *Potential Habitat*. *Out of* is the number of all observed relocations. *%* in the last column is the percentage of correct predicted relocations in potential habitat.

| **Description** | **Proportion of Sea** | **Number found** | **Out of** | **%** |
|---|---|---|---|---|
| ABFT habitat | 0.679 | 80 | 85 | **94.12** |

**Table 1.** Potential Habitat Modelling for ABFT

### *3.2.2. Habitat Modelling for Black-Browed Albatross (BBA)*

Though for the BBA species, only presence data are available, it is common practice to generate animal pseudo-absences techniques [5]. The generated pseudo-absences should be well separated from presences both in spatial and environmental (or ecological) space. The pseudo-absences are selected using a two-step approach. First, Correlated random Walk is used to generate 10 pseudo-absences for each presence relocation, where a constraint function is used to implement a spatial separation of presences from pseudo-absences. Second, EE and ENM is used to select the number of pseudo-absences which are well separated in environmental space. Though [6] performed a number of experiments and gave some recommendations on a number of pseudo-absences for different habitat modelling techniques, the experiments showed that equal number of presences and pseudo-absences lead to more robust performance for our Big data. Therefore we selected as many pseudo-absences as presences in the second step of pseudo-absence selection. This led us to a two-class problem for each geographic grid cells. Basically classified as either as *Potential Habitat (=1) or no Potential Habitat (=0)*. Two regression techniques were used to predict *Potential Habitat* for the BBA. These include: A Generalised Additive Model (GAM) and Boosted Regression Trees (BRT). The EO data which influence *Potential Habitat* selections were in this case: *Bathymetry, SST, SLA and EKE*. The *Potential Habitat* modelling was done for each animal breeding stage (or monthly for non-breeding stage). The comparison of GAM and BRT for incubation stage both on training and testing set are given in Table 2, where Correct Classification Rates (CCRs) are shown for each class. The threshold for selecting habitat/no habitat was set to 0.5. Table 2 also shows that BRT produces better results both on training and testing modes.

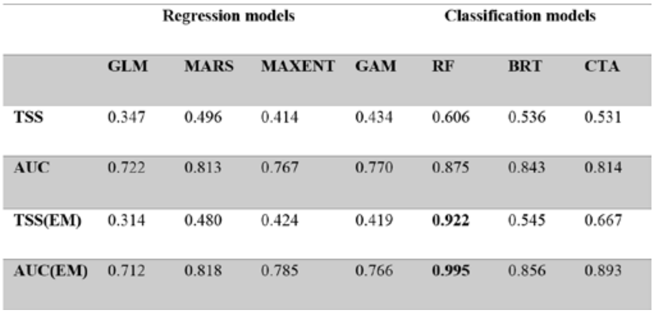
| **Classifiers** | **Training** | | **Testing** |
|---|---|---|---|
| | Habitat | No Habitat | Habitat |
| GAM | 77.1% | 76.3% | 68.45% |
| BRT | ***93.39%*** | ***99.51%*** | ***91.65%*** |

**Table 2.** Correct Classification Rates(CCRs) for BBA

### *3.2.3. Habitat Modelling for Loggerhead sea turtles*

Twenty one tracks of data on adult loggerhead sea turtles capturing their post-nesting movements during the years of 2004-2009 were also used for habitat modelling in this work. Two different foraging behaviours were observed with this animal population. These have been manually identified, and each animal was labelled as either an oceanic or a neritic forager. The overall modelling, pre-processing and pseudo-absence selection methods in this case were based on the works by Pikesley et al. [7], [8]. Three classification methods have been added and compared to the regression methods which were investigated in these works. Different spatial extents and numbers have also been examined for pseudo-absences. Data pre-processing stages include discarding relocations with unlikely speeds and turning angels. Best non-interpolated daily locations were then extracted for each of the tracks. Pseudo-absences were then generated within the convex hull of the presences via a random spatial-temporal sampling technique. Similar number of pseudo-absences as available presences were also generated (prevalence≅1).

The post-nesting habitat for oceanic adult loggerhead sea turtles was modelled using different classification and regression models. These experiments on EO data were performed eight times (8 replications) using different random sets of pseudo-absences [9]. In each replication, the data is split with a 75%/25% ratio for training and validation purposes. This random data splitting to training and validation sets is independently repeated four times in each replication. Table 3 shows the modelling evaluation results using TSS (*True Skill Statistic*), which is the most widely used stat alongside kappa for evaluating the accuracy of the species distribution models [10], and AUC (*Area Under Curve*) as the only non-threshold based evaluation method. The reported results are the mean of all the performances in all the runs and replications. It can be seen that overall classification methods provide better models while they can be further improved by building ensemble models with the four runs in each replication.



| | Regression models | | | | Classification models | | |
|---|---|---|---|---|---|---|---|
| | **GLM** | **MARS** | **MAXENT** | **GAM** | **RF** | **BRT** | **CTA** |
| **TSS** | 0.347 | 0.496 | 0.414 | 0.434 | 0.606 | 0.536 | 0.531 |
| **AUC** | 0.722 | 0.813 | 0.767 | 0.770 | 0.875 | 0.843 | 0.814 |
| **TSS(EM)** | 0.314 | 0.480 | 0.424 | 0.419 | **0.922** | 0.545 | 0.667 |
| **AUC(EM)** | 0.712 | 0.818 | 0.785 | 0.766 | **0.995** | 0.856 | 0.893 |

**Table 3**. Habitat modelling results for Loggerhead sea turtles (GLM (Generalized Linear Model), MARS (Multiple Adaptive Regression Splines), MaxEnt (Maximum Entropy), GAM (Generalized Additive Model), RF (Random Forest), BRT (Boosted Regression Trees), CTA (Classification Tree Analysis). *(TSS (EM) and AUC (EM) are ensemble models based on the four runs in one replication.)*

## 4. SUMMARY

The above research work used Copernicus and Argos Big data resources while adopting Big data infrastructure for wrapping a new generation of big data analytics services for predicting marine species habitats. The main focus of the EO4wildlife project was to establish performing mining and data analytics methods which automatically extract new knowledge from the newly available Copernicus Big EO data combined with those from Argos. The extracted knowledge, specifically concerns the confirmation of existing causalities between new emerging ecological conditions, due to climate change, and the response of selected vulnerable animal species at various oceanic regions. The next activity in the project will be on validating the elasticity and scalability of the big data analytics services which are being implemented on the EO4wildlife platform in collaboration with our project partners.

## 5. ACKNOWLEDGEMENT

This work is partly funded by the European Commission under H2020 Grant Agreement number: 687275. Also, the Authors would specifically like to thank the European Commission for providing access to COPERNICUS data, under the EO4life project, without which this study could not have been conducted.

## 6. REFERENCES

[1] Z. Sabeur, G. Correndo, G. Veres, B. Arbab-Zavar, J. Lorenzo, T. Habib, A. Haugommard, F. Martin, J-M. Zigna, and G. Weller. (2017) *EO Big Data Connectors and Analytics for Understanding the Effects of Climate Change on Migratory Trends of Marine Wildlife*. Proceedings of the 12th International Symposium of Environmental Information Systems. Zadar, Croatia, May 10th – 12th 2017. Springer International Publishing.

[2] G. Correndo, J-M Zigna, A. Haugommard. (2016). D3.1: Knowledge Base Service architecture Specification v1. Deliverable of EO4wildlife project. (see http://www.eo4wildlife.eu/deliverables - WP3 Advanced Analytics and Knowledge Base

[3] Druon, Jean-Noël, et al.(2011). *Potential feeding and spawning habitats of Atlantic bluefin tuna in the Mediterranean Sea*. Marine Ecology Progress Series 439: 223-240.

[4] J.-N. Druon and et al. (2016) Habitat suitability of the Atlantic bluefin tuna by class size:An Ecological niche approach. Progress in Oceanography, vol. 142, pp. 30-46.

[5] S.D. Senay, S.P. Worner, T. Ikeda (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. PLoS ONE, p. e71218

[6] Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? Methods in Ecology and Evolution 3: 327–338.

[7] S. K. Pikesley, S. M. Maxwell, K. Pendoley, D. P. Costa, M. S. Coyne, A. Formia and S. Ngouessono. (2013) "On the front line: integrated habitat mapping for olive ridley sea turtles in the southeast Atlantic. "Diversity and Distribution". Vol.19, no 12, pp.1518-1530.

[8] S. K. Pikesley, A. C. Broderick, D. Cejudo, M. S. Coyne, M. H. Godfrey, B. J. Godley and M. J. Witt. (2015). "Modelling the niche for a marine vertebrate: a case study incorporating behavioural plasticity, proximate threats and climate change," *Ecography*, vol. 38, no. 8, pp. 803-812.

[9] M. Barbet-Massin, F. Jiguet, C. H. Albert and W. Thuiller. (2012). "Selecting pseudo-absences for species distribution models: how, where and how many?," *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 327-338.

[10] O. Allouche, A. Tsoar and R. Kadmon. (2006)."Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)," *Journal of applied ecology*, vol. 43, no. 6, pp. 1223-1232.