Machine learning as a successful approach for predicting complex spatio–temporal patterns in animal species abundance

B. Martín, J. González–Arias, J. A. Vicente–Vírseda

Martín, B., González–Arias, J., Vicente–Vírseda, J. A., 2021. Machine learning as a successful approach for predicting complex spatio–temporal patterns in animal species abundance. *Animal Biodiversity and Conservation*, 44.2: 289–301, Doi: https://doi.org/10.32800/abc.2021.44.0289

Abstract

Machine learning as a successful approach for predicting complex spatio–temporal patterns in animal species abundance. Our aim was to identify an optimal analytical approach for accurately predicting complex spatio–temporal patterns in animal species distribution. We compared the performance of eight modelling techniques (generalized additive models, regression trees, bagged CART, k–nearest neighbors, stochastic gradient boosting, support vector machines, neural network, and random forest –enhanced form of bootstrap. We also performed extreme gradient boosting –an enhanced form of radiant boosting– to predict spatial patterns in abundance of migrating Balearic shearwaters based on data gathered within eBird. Derived from open–source datasets, proxies of frontal systems and ocean productivity domains that have been previously used to characterize the oceanographic habitats of seabirds were quantified, and then used as predictors in the models. The random forest model showed the best performance according to the parameters assessed (RMSE value and R^2). The correlation between observed and predicted abundance with this model was also considerably high. This study shows that the combination of machine learning techniques and massive data provided by open data sources is a useful approach for identifying the long–term spatial–temporal distribution of species at regional spatial scales.

Key words: Balearic shearwater, Machine learning, Random forest, Chlorophyll, NAO index

Resumen

Aprendizaje automático: una buen método para predecir patrones espacio–temporales complejos en la abundancia de especies animales. Nuestro objetivo fue determinar un método analítico óptimo para predecir de manera precisa patrones espacio–temporales complejos en la distribución de especies animales. En concreto, utilizando los datos recopilados en el proyecto eBird sobre la pardela balear, que es un ave migratoria, se compararon ocho técnicas diferentes de elaboración de modelos (modelos aditivos generalizados, árboles de regresión, *bagged* CART, *k–nearest neighbors, stochastic gradient boosting,* máquinas de vectores de soporte, redes neuronales, así como el bosque aleatorio –una forma mejorada de *bootstrap*– y *extreme gradient boosting* –una forma mejorada de *gradient boosting*) con objeto de predecir los patrones espaciales observados en la abundancia de esta especie. Utilizando conjuntos de datos de código abierto, se han cuantificado los indicadores de los sistemas frontales y la productividad de los océanos que ya se habían empleado con anterioridad para caracterizar los hábitats oceánicos de las aves marinas y, posteriormente, se han utilizado como predictores en los modelos. El bosque aleatorio resultó ser el modelo que ofreció el mejor rendimiento de acuerdo con los parámetros evaluados (RMSE y R^2). La correlación obtenida con este modelo entre la abundancia observada y la predicha también fue considerablemente alta. En este trabajo, mostramos la utilidad de combinar técnicas de aprendizaje automático y los datos masivos proporcionados por diferentes fuentes de código abierto para determinar la distribución espacio–temporal a largo plazo de las especies a escalas espaciales regionales.

Palabras clave: Pardela balear, Aprendizaje automático, Bosque aleatorio, Clorofila, Índice NAO

Received: 27 V 21; Conditional acceptance: 30 VI 21; Final acceptance: 24 VIII 21

Beatriz Martín, Randbee Consultants, c/ Carretería 67, Málaga, Spain.– Julio González–Arias, Juan A. Vicente–Vírseda, Departamento de Economía de la Empresa y Contabilidad, Paseo de la Senda del Rey 11, 28040 Madrid

Corresponding author: B. Martín. E–mail: beatriz.martin@randbee.com

ORCID ID: 0000-0001-6893-2187

ISSN: 1578–665 X
eISSN: 2014–928 X

Introduction

The continuous technical development of electronic devices such as geolocators, GPSs, and PTT devices for tracking animal movements have promoted the understanding of the distribution of many species (Katzner and Arlettaz, 2020). However, for migratory species, even with good sample sizes, tracking data cannot give a fully representative sample from all the possible migratory route alternatives. The reason for this is that data derived from these studies are temporally constrained to a few years of monitoring of a few individuals, but migratory species usually exhibit a great capacity to alter migratory behavior in response to environmental variability and specific individual traits (Martín et al., 2016). Consequently, even though they can provide highly detailed information on an individual's movements, electronic devices have a limited ability to improve our understanding of the adaptation of migration strategies to deal with a changing environment at both population and species levels (Martín et al., 2019).

In contrast, although direct observation methods such as census are unable to detect birds when they use areas out of human sight, they can provide a valuable overall picture despite the missing information if we apply predictive models to fill the spatial and temporal gaps in the original datasets (Gouraguine et al., 2019). In this sense, active public involvement in scientific research (citizen science) has become an excellent source of data for scientists and policymakers (Strasser et al., 2019). Citizen science projects provide observations of millions of species each year (Kelling et al., 2013; Chandler et al., 2017). Compared with the detailed data gathered from electronic devices, the massive datasets from citizen science projects have the advantage of offering low–cost information on long–term temporal and large spatial extents from many individuals belonging to several populations.

Prominent among citizen science projects devoted to birds and used by a large number of scientists is eBird (eBird, 2017), a biodiversity–related citizen science project gathering records of birds provided by professional and amateur ornithologists around the world. Caveats on the use of datasets from citizen science projects are related to errors in species identification, biases in count estimates and uneven sampling effort, both in terms of temporal and spatial coverage, among others, which make volunteer data highly variable in terms of precision. Recent advances in Big Data analysis and, more specifically, in Data Mining techniques, thanks to the application of artificial intelligence (AI) and, more specifically, machine learning (ML) techniques, allow to obtain robust predictions from these 'noisy' and incomplete datasets (Schain, 2015). For this reason, these techniques have become an extremely useful tool to extract relevant information in many disciplines of knowledge.

As model species, we focused on the Balearic shearwater (*Puffinus mauretanicus*), a critically endangered seabird species (BirdLife International, 2018) which spends about one quarter of the year on migration (Guilford et al., 2012). Previous research founded on boat–based survey counts showed Balearic shearwater abundance to be extremely difficult to predict, and it failed to provide reliable predictions on the spatial distribution of shearwater numbers (Oppel et al., 2012).

Our aim was to identify an optimal analytical approach for accurately predicting spatio–temporal patterns in the abundance of this species during migration from observations collected by professional and amateur ornithologists in citizen science projects, namely eBird database. To this end, we built several predictive models applying traditional statistical analysis (generalized additive models) and the latest machine learning techniques (neural networks, gradient boosting, random forest, support vector machines, among others). With these models we predicted the abundance of the Balearic shearwater along its migratory route over the Mediterranean and Atlantic coasts. These models were founded on the assumption that there is a relationship, direct or indirect, between environmental variables and the shearwater distribution. In this way, from a selected sect of environmental predictors derived from large–scale open datasets (NCEP/NCAR, AIS, NOAA, among others), we modelled the abundance of shearwaters over time and across space. The various models were then compared in order to determine the best one in terms of accuracy and predictive ability. Our final aim was to describe a successful analytical approach that can be extended to other animal species for which citizen science projects are collecting abundance distribution data.

Material and methods

Study species

The Balearic shearwater is included as 'Critically Endangered' on the IUCN Red List (BirdLife International, 2018) and is also considered as threatened bird for the European Union (rare or vulnerable bird species as listed in annex I of the E.U. Bird Directive). Most of the population of Balearic shearwater leaves the Mediterranean each year after breeding (Guilford et al., 2012) and stays mainly in the Atlantic during the non-breeding season (Arcos, 2011). At sea, it usually occurs in productive shelf areas related to oceanographic frontal systems (Louzao et al., 2006; Oppel et al., 2012; Pérez–Roda et al., 2017). During migration, Balearic shearwaters tend to fly very close to the shoreline (Arroyo et al., 2014), with an average off-shore distance of about 1,190 m at the Strait of Gibraltar (Mateos and Arroyo, 2011). Fluctuations in Balearic shearwater migration, both seasonally and inter-annually, seem to be related to changes in food resources (Wynn et al., 2007; Jones et al., 2014). The diet of the Balearic shearwater includes small pelagic and also demersal fish, frequently obtained from trawling discards. The species can also feed on plankton and macrozooplankton, specifically krill (Arcos and Oro, 2002; Louzao et al., 2015).

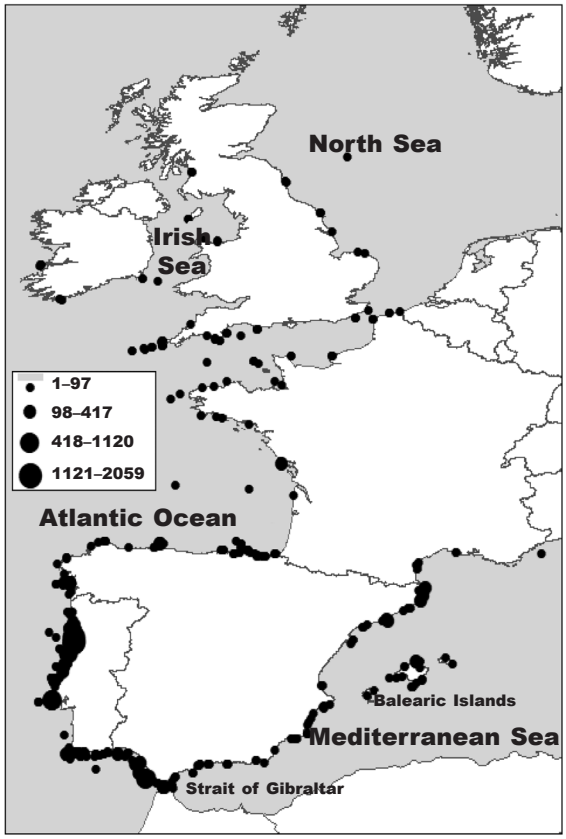


Fig. 1. Spatial distribution of the shearwater abundance data during the post–breeding migration period (May–August). Years 2005–2017; n = 1,881.

Fig. 1. Distribución espacial de los datos sobre la abundancia de la pardela correspondientes al periodo de migración post–nupcial (mayo–agosto). Años 2005–2017; n = 1.881.

The extent of the study area (fig. 1) covers the whole distribution range of the Balearic shearwater throughout the year (BirdLife International, 2018).

Variables in the models

Response variable

Our response variable was Balearic shearwater abundance (abundance considering only presence, thus, absence of abundance –i.e., zeros– was not considered in the analysis) (Pearce and Boyce, 2006), expressed as the number of birds sighted on a given date at a given latitude/longitude, during migration. Daily abundance of the species was recorded from 1964 to 2018, both as opportunistic sighting records and within systematic effort–based surveys (with a standardized duration of the sampling effort) obtained from the online database eBird (eBird, 2017). Due to the opportunistic nature of some of these data, we needed powerful modelling techniques to obtain robust results (see below). Although it was possible to differentiate opportunistic and systematic surveys within the dataset, we opted to keep all records for our analysis in order to test the robustness of the modelling approach in case our methods will be extended to other species for which this information is not available.

Specifically, we considered abundance during the 'post–breeding migration' period, defined as the northward migration of birds leaving the Mediterranean and molt, between May 1st–August 30th (Mourino et al., 2003; Yésou, 2003).

Environmental predictors

Shearwater abundance was modelled using 15 environmental predictors (table 1) that have been previously described to be related with the spatial distribution at sea of this and other seabird species (Yésou, 2003; Dias et al., 2012; Oppel et al., 2012; Jones et al., 2014). As migrating birds need to replenish energy reserves during stopover periods at key locations where they maximize their refueling

opportunities (Wynn et al., 2007; Benoit–Bird et al., 2013), fluctuations in bird abundance during migration are closely related to changes in food resources (Wynn et al., 2007). Therefore, most of the variation in seabirds appears to be mediated by changes in prey abundance (Frederiksen et al., 2006; Benoit–Bird et al., 2013). We used chlorophyll concentration (Chla, measured in mg/m^3) as a proxy of marine productivity (Wakefield et al., 2009). Specifically, we downloaded satellite–based monthly products at 4 km spatial resolution for the 2003–2017 period (JRC Data Catalogue; http://gmis.jrc.ec.europa.eu/satellite/4km/). Discard availability may influence the at–sea distribution of shearwaters (Cortés et al., 2018; Genovart et al., 2018). As a proxy of food availability (both discards and fishes) we used information on fisheries. This information was inferred from the distribution of fishing vessels (Natale et al., 2015) between 2014 and 2015, sourced by the JRC Data Catalogue at 1 km resolution (http://gmis.jrc.ec.europa.eu/dataset/jrc–fad–ais1415). This product identifies the areas where fishing is most frequent. Vessel tracking data were derived from the Automatic Identification System (AIS), an open source data system that allows analysis of the relation between fishing communities and fishing areas at high spatial resolution across Europe. Specifically, data used to build the map consists of 150 million positions from European fishing vessels above 15 m in length. In spite of its limited temporal coverage, the main strength of this dataset is its fine spatial resolution. This proxy on food availability in spatial terms, however, is complemented in our analysis with the high temporal resolution in marine productivity provided by chlorophyll concentration. Together with chlorophyll concentration, sea surface temperature (SST) is also a proxy of water mass distributions, frontal systems, and ocean productivity (Ramos et al., 2012; Robinson et al., 2013; Afán et al., 2014). In shearwaters, which perform dynamic soaring flight, oceanic winds are also of major importance in modulating migratory behavior (González–Solís et al., 2009). Wind speed and direction usually affect both migratory behavior (Catry et al., 2011; Dias et al., 2012; but see Dell'Ariccia et al., 2018) and seabird detectability by the observers (Martín et al., 2019). Data on SST and wind were obtained through RNCEP package (Kemp et al., 2012) in R (R Core Team, 2018) which allowed to request data from the NCEP/NCAR Reanalysis dataset (NOAA ESRL Physical Sciences Division; https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis2.html) for a specified range of space and time based on the observations of shearwaters. Specifically, we requested daily values at midday at the surface level, since Balearic shearwaters tend to migrate at a very low height (Martín et al., 2019). Variables requested were 'air.995' (air temperature), 'uwind.sig995' (u–wind component) and 'vwind.995' (v–wind component). RNCEP interpolates the nearest data value (in terms of location and time) in the NCEP/NCAR dataset to provide the requested information. We also considered the standard deviation (associated standard deviation of the points used to perform the interpolation) of the meteorological values (u–wind, v–wind and temperature) used for this interpolation as additional predictors in our models (table 1). General flight activity of seabirds increases during moonlit nights, and moon phase has been shown to affect shearwater migration behaviour (Dias et al., 2012). As changes in the flying patterns during the night may also affect daytime flights, we considered the daily fraction of the moon illuminated at midnight as an additional predictor, sourced by the U.S. Naval Observatory and Astronomical Applications Department (http://aa.usno.navy.mil/data/).

Together with the usual inter–annual variability in food resources, long–term climate change may also affect the at–sea distribution of this species (Wynn et al., 2007; Votier et al., 2008; Luczak et al., 2011), likely through effects on fish stocks (Tsikliras et al., 2019). The possible impact of climate change on the distribution of shearwaters at a regional level was taken into account through the North Atlantic Oscillation Index (NAO; Visbeck et al., 2001) obtained from the Climate Prediction Center (U.S. National Weather Service, NOAA), as monthly data from 1950 to 2018 (http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml). As SST and wind predictors already provided a daily resolution of the weather conditions at specific locations, monthly values rather than daily NAO index were preferred, as a surrogate of regional climate conditions affecting shearwater migration at a more general and longer time frame.

Water depth has been shown to influence seabird distributions (Yen et al., 2004). Specifically, Balearic shearwaters occur in relatively shallow, coastal waters along the shoreline during migration (Martín et al., 2020). Bathymetry was used as a surrogate of coastal–pelagic areas (Afán et al., 2019). Bathymetric data were sourced by European Marine Observation and Data Network (EMODnet; http://www.emodnet-bathymetry.eu/data-products) at ~200 m resolution although they were later aggregated through a moving–average window to 10 km resolution in order to obtain a more general bathymetry pattern regarding the shearwater observation location. This cell size was selected as a compromise between this general pattern in bathymetry and the sufficient resolution for conservation purposes.

Fluctuations in seabird abundance during migration are closely related to changes in food resources (Frederiksen et al., 2006; Benoit–Bird et al., 2013). However, photoperiodic cues and/or endogenous rhythms may also modulate seabird breeding and migration periods (Marshall and Serventy, 1959). Therefore, apart from food availability, migration decisions in Balearic shearwaters might be partially dictated by day length and/or by an internal rhythm that make the bird move instinctively into the north–west, while the post–breeding season progresses, and then into the south–east during the pre–breeding period. Date, longitude and latitude variables allow us to include the endogenous rhythm of the bird in the models. In addition, longitude and latitude can be indirect proxies of the effects that variable wintering sites, length of the route and en–route environmental conditions may pose to migrant shearwaters. Finally, due to potential

Table 1. Descriptive summary of the predictors in the models. Sources: EMODnet, European Marine Observation and Data Network; AIS, Automatic Identification System (JRC Data Catalogue); NCEP/NCAR, Reanalysis dataset (NOAA ESRL Physical Sciences Division); MERIS, JRC Data Catalogue; CPC, Climate Prediction Center (US National Weather Service, NOAA); USNO & AAD, U.S. Naval Obervatory and Astronomical Applications Department.

Tabla 1. Resumen descriptivo de las variables utilizadas en los modelos. (Para las abreviaturas de las fuentes, véase arriba).

Name	Description	Unit	Source	Period
batim	Bathymetry	meters	EMODnet	2012
fish	Fishing intensity	number of vessels	AIS	2014–2015
tmmean	Mean temperature	ºK	NCEP/NCAR	1964–2018
tmstd	Atandard deviation of mean temperature	ºK	NCEP/NCAR	1964–2018
uwmean	Mean wind speed (u–wind: east–west direction)	m/s	NCEP/NCAR	1964–2018
uwstd	Standard deviation of mean wind speed (u–wind: east–west direction)	m/s	NCEP/NCAR	1964–2018
vwmean	Mean wind speed (v–wind: north–south direction)	m/s	NCEP/NCAR	1964–2018
vwstd	Standard deviation of mean wind speed (v–wind: north–south direction)	m/s	NCEP/NCAR	1964–2018
clorof	Chlorophyll concentration	mg/m^3	MERIS	2004–2017
NAO	NAO index	–	CPC	1950–2018
moon	Daily fraction of the moon illuminated at midnight	%	USNO & AAD	2005–2018

differences in the rates of change of the environmental predictors across space, interactions between predictors and 'latitude' and longitude, and between 'year' and 'latitude', may allow to quantify both the spatial and temporal heterogeneity in the migratory responses.

Statistical analysis

As usual in count data, abundance of shearwaters followed a Poisson distribution (Hilbe, 2014). Although there is a lack of assumptions in the distribution of the data in machine learning models, when we use approaches where data partitioning is applied, we can obtain better results if we model a dependent variable homogenously distributed, because the model dispersion increases as long as the variable increases. Supporting the previous statement, we observed that this log transformation increased the variance explained by all the models (up to 1.4 times in the case of the RF model). Therefore, before building our models we log–transformed (natural log) the abundance data. The estimates of abundance obtained from the models were then scaled back to a linear scale before assessing the model performance. In addition, before modelling, predictor variables were pre–processed: centered and scaled (subtracting the mean of the predictor's data from the predictor values and then dividing by the standard deviation).

To evaluate the different performance, up to eight different modelling techniques were carried out on the post–nuptial migration dataset. The set of models constitutes a representative sample of the various machine learning techniques available for predictive analyses in which the dependent variable is quantitative (regression approaches). Specifically, regression trees (CART), bootstrap aggregation (bagged CART), extreme gradient boosting, stochastic gradient boosting, K–nearest neighbours (KNN), support vector machine (SVM) and multilayer perceptron neural network (MLP). In addition to these machine learning approaches, we include Generalized Additive Models (GAM). GAMs are simple, transparent, and flexible models which do not assume a linear relationship between independent and dependent variables.

This approach is suitable to model count data such as animal abundance (Zuur et al., 2009). Apart from flexibility, in contrast with other multivariate models, such as multivariate adaptive regression splines (MARS) and machine learning techniques, it provides an interpretable solution, thus providing a good balance between the interpretable linear model and the extremely flexible, but 'black box' iterative/learning algorithms. To enhance model performance, according to the nature of the response variable (see above), we applied a GAM with Poisson distribution and log–link function. Prior to modelling we assessed collinearity from the total set of environmental predictors through the correlation between pairs of variables (Pearson correlation coefficient). Although machine learning techniques are not as highly subjected to the effects of collinearity as traditional statistical approaches, computation timing and model results can benefit from the removal of any redundant features from the training dataset, irrespectively of the model's algorithm.

Comparisons between specific modelling techniques and particular models were based on the Root Mean Square Error (RMSE; the average difference between the observed known values of the outcome and the predicted value by the model; Gareth et al., 2014), and on R^2, measured in the caret package in R (R Core Team, 2018) as the squared correlation between observed and predicted values (Kuhn, 2008). The validation procedure of the models was based on a random split of the abundance data into training and test data (setting aside 20 % of the data for testing the models; Araújo et al., 2006). Performance of the built models was estimated using 5–fold cross–validation on the training data. Models were calibrated on the training data and then evaluated on the test data to determine the model's ability to generalize to other datasets. Except for the GAM analysis, for model building and assessment we used the caret package (Kuhn, 2008) in R (R Core Team, 2018). Compared to a Gaussian identity–link GAM approach, a Poisson log–link GAM on the same training dataset increased both RMSE (from 124.13 to 139.55) and R^2 (from R^2 = 0.19 to R^2 = 0.36). As Poisson log–link GAMs are not supported by caret, predictions (and R^2) for these models were derived using the mgcv package (Wood, 2011) also in R. However, to ensure minimum RMSE values that were comparable between modelling techniques, RMSE for a Gaussian GAM model was quantified using caret.

All the models were built using the total set of predictors described in table 1. Most of the models assessed have parameters that must be tuned to obtain an optimal fitting. To determine the parameter values offering the best fit, we specified a set of tuning values to be tested during the calibration of the models. Generally, we applied a grid search method, thus we evaluated the model over different combinations of parameters included in the grid (table 2). Specifically for the GBM models, the range for the parameter tuning was based on recommendations derived from previous research (Friedman, 1999a, 1999b; Friedman et al., 2001; Ridgeway, 2005). To identify the model with the optimal parameter combination (providing the best fit) we compared the RMSE values (see supplementary material) of the models (Gareth et al., 2014).

To detect significant differences in the performance between modeling techniques, all pair–wise differences in RMSE and R^2 over the model resamples were quantified and tested to assess whether the difference was equal to zero. The resulting confidence level was adjusted using Bonferroni correction (Kuhn, 2008).

We also assessed the relative importance of the variables used for predicting shearwater abundance in the different models using the varImp function in caret package (Kuhn, 2007).

Results

Assessment of the available eBird data showed that records before 2005 (from 1964 to 2004) were scarce and did not properly cover the migration periods. For this reason, to ensure a minimum sample size for estimating predictions, models were built only with data from after 2005. Similarly, we did not consider data for 2018 in the analysis as they were seasonally incomplete.

After data filtering, a total of 1,881 post–breeding records (ranging from 1 to 2,059 birds; median = 5) on 58,901 individual observations of shearwaters remained for the years 2005–2017. According to the migration areas which are known to be used by the study species (BirdLife International, 2018), the spatial representation of the remaining observations was good (fig. 1).

The assessment of collinearity, measured as Pearson correlation coefficients between pairs of predictors, showed non–significant results at the 0.05 *p*–level for any of the pairs of variables, thus all predictors included in the initial set were considered for modelling.

According to the cross–correlation resamples (fig. 2), the best modelling technique was Random Forest (RF), showing the lowest RMSE mean value among resamplings (mean RMSE = 1.11) and the largest R^2 (mean R^2 = 0.47), closely followed by the results from XGBM (RMSE = 1.13; R^2 = 0.44). Differences in RMSE between models were only significant in the case of MLP against the best modelling techniques (RF, XGBM and GBM, in that order). The worst techniques both considering RMSE and R^2 were MLP followed by GAM, although the GAM model with Poisson distribution and log–link function almost doubled the R^2 value (mean R^2 = 0.36; min = 0.33; max = 0.41). Differences in RMSE and also in R^2 of KNN against RF, XGBM and SVM were almost significant at the Bonferroni *p*–level (*p*–value < 0.1). CART was also significantly better than MLP in terms of R^2 (fig. 3). The resamplings of all the modelling techniques assessed, including RF, showed high variability in terms of RMSE and R^2 values, indicating that both errors and the predictive ability of the models are considerably variable depending on the data subset used in each resampling. Mainly for this reason, the difference in terms of performance

Table 2. Parameter values used for model calibration.

Tabla 2. Valores de los parámetros utilizados en la calibración de los modelos.

Model Parameters	Tuning sequence	Tuning by	Method
Bagging			
# baggings	10, 20, 30, 35, 40		Manually assessed
CART			
cp	0–0.07	0.001	Grid search
Extreme gradient boosting			
# rounds	200–1000	50	Grid search
eta	0.01, 0.015, 0.025, 0.05, 0.1, 0.3		Grid search
maximum depth	3, 5, 7, 9, 12, 15, 17, 25	1	Grid search
gamma	0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0		
colsample_bytree	0.6, 0.7, 0.8, 0.9, 1.0		
min_child_weight	1, 3, 5, 7		
subsample	0.6, 0.7, 0.8, 0.9, 1.0		
Gradient boosting			
interaction depth	4–10	2	Grid search
# trees	0–2,500	500	Grid search
shrinkage	0.001, 0.01, 0.1		Grid search
minimum number of observations in tree's terminal nodes (n.minobsinnode)	5, 10, 20		Grid search
sampling fraction (bagg.fraction)	0.2, 0.3, 0.5		Manually assessed through iteratively repeating the grid search (see suppl. material)
KNN			
k	1–25	1	Grid search
Neural network			
# layers	1–25	1	Grid search
Random forest			
mtry	1–25	1	Grid search
# trees	1000, 1500, 1600, 1625, 1650, 1700, 1725, 1750, 2000	500	Manually assessed through iteratively repeating the grid search (see suppl. material)
Support vector machines			
sigma	60–180/400; 80:100/400; 100:120/400; 120–140/400; 140:180/400	1	Grid search that was iteratively refined based on RMSE results (see suppl. material)
C	1:10/10; 10:20/10; 20:30/10; 30/10	1	Grid search that was iteratively refined based on RMSE results (see suppl. material)

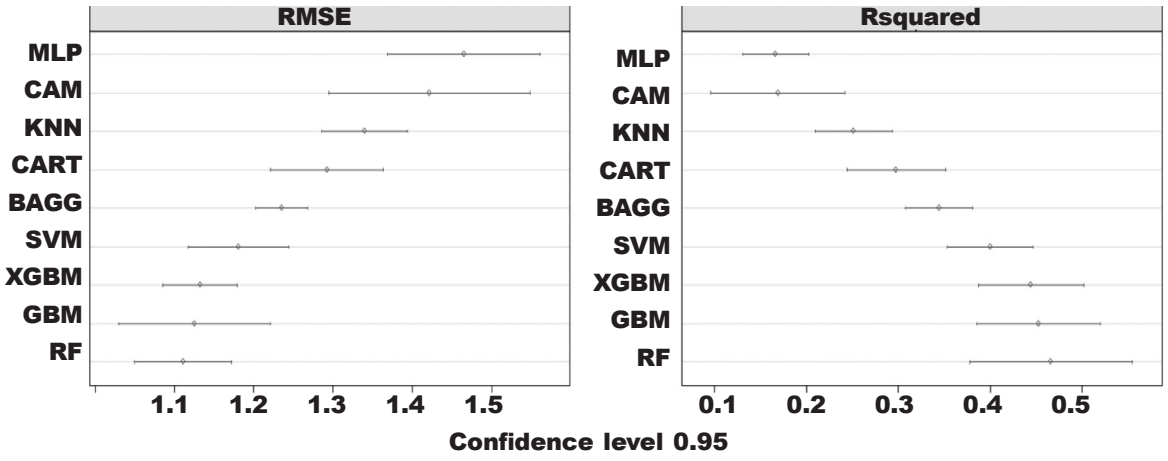


Fig. 2. Results from the resampling (5–fold cross–validation) analysis of the various models assessed. Range (minimum–maximum) and mean values; 95 % confidence intervals for the medians. Models based on data for 2005–2017: * GAM, Gaussian distribution and identity–link function.

*Fig. 2. Resultados de los remuestreos (validación cruzada sobre cinco conjuntos) de los diferentes modelos evaluados. Inter (valores mínimo y máximo) y valores medio; intervalos de confianza del 95 % para las medianas. Modelos basados en la muestra de los años 2005–2017: * GAM, distribución normal y función de enlace de identidad.*

(assessed through RMSE and R^2) between models was mostly non–statistically significant.

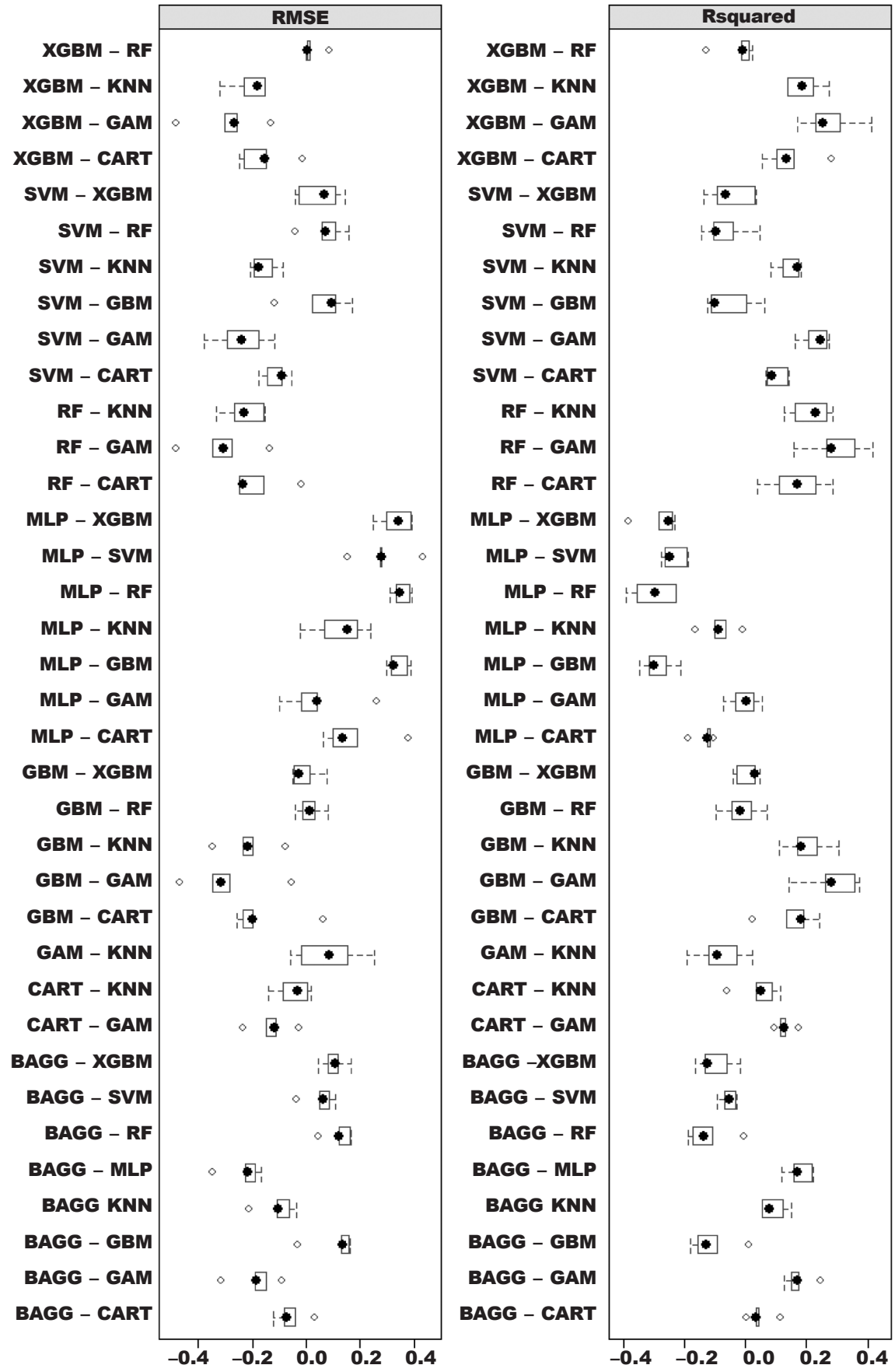
The relative importance of the variables was variable between models. However, standard deviations of eastward and westward wind speed (uwstd and vwstd, respectively) were important variables in most of the models assessed. Year (year) of the observation, chlorophyll concentration (clorof) and bathymetry (batim) were also variables of high importance in predicting shearwater abundance (fig. 4).

Discussion

This study showed that the combination of Random Forest techniques and massive data provided by various open data sources, including data gathered in citizen science initiatives, is a useful approach to identify the long–term spatial–temporal distribution of species at regional spatial scales. This machine learning technique was highly successful in capturing both spatial and temporal patterns in shearwater abundance at sea. The results from Random Forest techniques clearly outweighed those obtained with other traditional statistical modelling techniques such as GAM. In observational studies in Ecology and Conservation Biology, where the environmental variability is considerably high and it is not easy to take into consideration all the factors affecting the research subject, the ability of statistical models to predict the existing variability in the dependent variable is usually poor (<10 %), both because of the inherent 'noise' in the data as well the randomness of the natural phenomena that is being modelled (Møller and Jennions, 2002) but randomness and noise may reduce this amount considerably in biological studies. In contrast to traditional statistical techniques, we showed that Random Forest can largely increase the variability explained (up to 53 %) and the predictive ability (70 %) of the models calibrated with noisy data in complex scenarios where both temporal and spatial variation is present. The variance explained by our random forest model is not only larger than usual in Ecology and Conservation Biology studies (about 10 % according to Møller and Jennions, 2002) but it is also larger than the explanatory ability found

Fig. 3. Differences in the performance (in terms of RMSE and R^2) of the assessed models. Range and median; 95 % confidence intervals for the means: * GAM, Gaussian distribution and identity–link function.

*Fig. 3. Diferencias en los resultados obtenidos (en cuanto RMSE y R^2) con los modelos evaluados. Intervalo y mediana. Intervalos de confianza del 95 % para las medias: * GAM, distribución normal y función de enlace de identidad.*


RMSE
Rsquared
XGBM – RF
XGBM – KNN
XGBM – GAM
XGBM – CART
SVM – XGBM
SVM – RF
SVM – KNN
SVM – GBM
SVM – GAM
SVM – CART
RF – KNN
RF – GAM
RF – CART
MLP – XGBM
MLP – SVM
MLP – RF
MLP – KNN
MLP – GBM
MLP – GAM
MLP – CART
GBM – XGBM
GBM – RF
GBM – KNN
GBM – GAM
GBM – CART
GAM – KNN
CART – KNN
CART – GAM
BAGG – XGBM
BAGG – SVM
BAGG – RF
BAGG – MLP
BAGG KNN
BAGG – GBM
BAGG – GAM
BAGG – CART
–0.4
–0.2
0.0
0.2
0.4

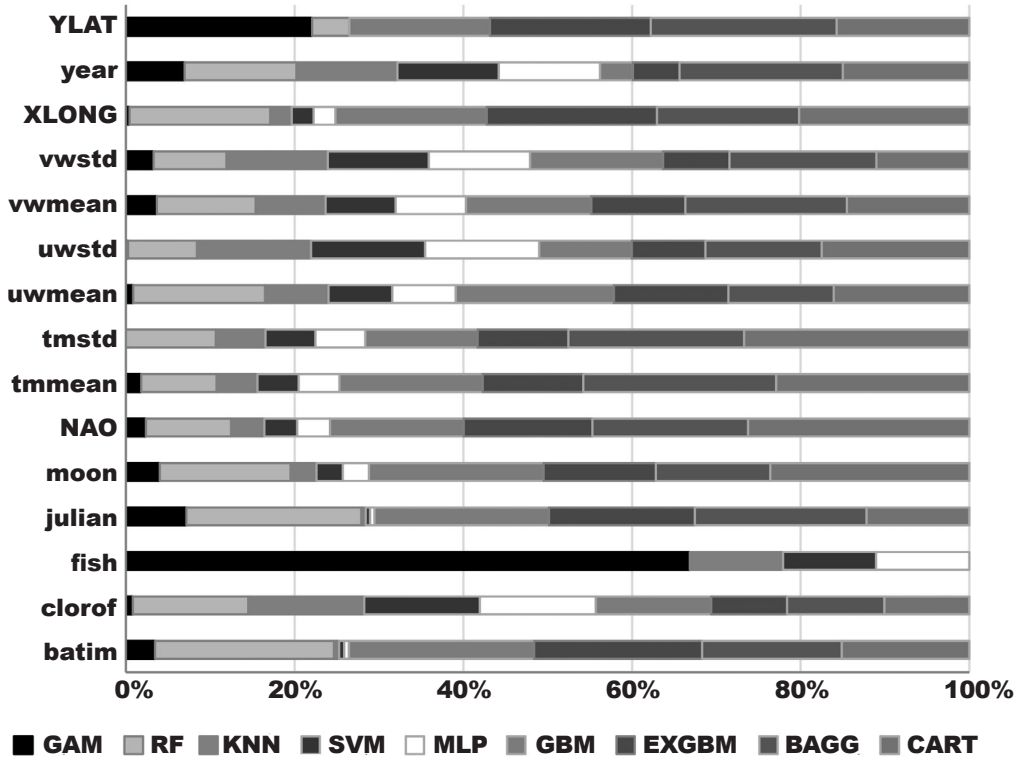


Fig. 4. Relative variable importance, as percentage, in the models (from varImp function in caret package; (Kuhn, 2007).

Fig. 4. Importancia relativa (en porcentaje) de las variables en los diferentes modelos (a partir de la función varImp en el paquete caret (Kuhn, 2007).

in similar studies (15 %) applying machine learning approaches to traditional survey (less 'noisy') data (i.e, boat–based surveys; Oppel et al., 2012).

Machine learning techniques are increasingly applied in order to obtain valid and accurate information from massive data sets that, due to their volume, noise and variety, until not too long ago it was not possible to analyze. However, according to our results, machine learning techniques in general are not a panacea, since other models also assessed in this study, particularly MLP, did not show good results in predicting shearwaters during post–breeding. Our results also highlighted that the predictive ability of the random forest model was highly conditioned on the data subset used to calibrate the model. For this reason, although this technique can deal with datasets with small sample size thanks to the bagging procedure implemented, its application requires suitable data pre–processing to ensure that the data for the analysis is fully representative of the phenomena to be modelled.

Conclusions

This study shows that the combination of machine learning techniques and massive data provided by open data sources is a useful approach to identify the long–term spatial–temporal distribution of species at regional spatial scales. Our research showed that machine learning techniques, and specifically random forest approaches, can be a good choice for the analysis of highly noisy massive datasets such as those collected by volunteers in citizen science projects. According to our results, these models allow the successful capture of the spatio–temporal variation in continuous biological variables such as bird migratory abundance recorded over long–time frames, thereby enabling long–term spatial monitoring of mobile species.

Acknowledgements

This study is part of the research project ('Environmental factors determining the interannual variation in the migration of Balearic and Scopoli's shearwaters in the Mediterranean') which formed part of the Annual Programme of Grants of the Instituto de Estudios Ceutíes (IEC, Autonomous City of Ceuta, Spain), 2018–2019. Data sourced by eBird can be obtained after registration and request at https://ebird.org/.

References

Afán, I., Navarro, J., Cardador, L., Ramirez, F., Kato, A., Rodríguez, B., Ropert–Coudert, Y., Forero, M., 2014. Foraging movements and habitat niche of two closely related seabirds breeding in sympatry. *Marine Biology*, 161: 657–668, Doi: 10.1007/s00227-013-2368-4

Afán, I., Navarro, J., Grémillet, D., Coll, M., Forero, M., 2019. Maiden voyage into death: are fisheries affecting seabird juvenile survival during the first days at–sea? *Royal Society Open Science*, 6: 181151, Doi: 10.1098/rsos.181151

Araújo, M. B., Thuiller, W., Pearson, R. G., 2006. Climate warming and the decline of amphibians and reptiles in Europe. *Journal of Biogeography*, 33: 1712–1728, Doi: 0.1111/j.1365-2699.2006.01482.x

Arcos, J., 2011. *International species action plan for the Balearic shearwater*, *Puffinus mauretanicus*. SEO/BirdLife & BirdLife International.

Arcos, J., Oro, D., 2002. Significance of fisheries discards for a threatened Mediterranean seabird, the Balearic shearwater *Puffinus mauretanicus. Marine Ecological Progress Series*, 239: 209–220, www.jstor.org/stable/24866058 [Accessed on 8 Sept. 2021].

Arroyo, G. M., Mateos–Rodríguez, M., Muñoz, A. R., De la Cruz, A., Cuenca, D., Onrubia, A., 2014. New population estimates of a critically endangered species, the Balearic Shearwater *Puffinus mauretanicus*, based on coastal migration counts. *Bird Conservation International*, 26: 87–99, Doi: 10.1017/S095927091400032X

Benoit–Bird, K. J., Battaile, B. C., Heppell, S.A., Hoover, B., Irons, D., Jones, N., Kuletz, K. J., Nordstrom, C. A., Paredes, R., Suryan, R. M., Waluk, C. M., Trites, A. W., 2013. Prey Patch Patterns Predict Habitat Use by Top Marine Predators with Diverse Foraging Strategies. *Plos One*, 8: e53348, Doi: 10.1371/journal.pone.0053348

BirdLife International, 2018. *Puffinus mauretanicus*. The IUCN Red List of Threatened Species 2018: e.T22728432A132658315, Doi: 10.2305/IUCN.UK.2018-2.RLTS.T22728432A132658315.en [Accessed on 15 March 2019].

Catry, P., Dias, M.P., Phillips, R.A., Granadeiro, J.P., 2011. Different Means to the Same End: Long–Distance Migrant Seabirds from Two Colonies Differ in Behaviour, Despite Common Wintering Grounds. *Plos One*, 6: e26079, Doi: 10.1371/journal.pone.0026079

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller–Rushing, A. J., Newman, G., Rosemartin, A., Turak, E., 2017. Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, 213: 280–294, Doi: 10.1016/j.biocon.2016.09.004

Cortés, V., García–Barcelona, S., González–Solís, J., 2018. Sex– and age–biased mortality of three shearwater species in longline fisheries of the Mediterranean. *Marine Ecological Progress Series*, 588: 229–241, Doi: 10.3354/MEPS12427

Dell'Ariccia, G., Benhamou, S., Dias, M. P., Granadeiro, J. P., Sudre, J., Catry, P., Bonadonna, F., 2018. Flexible migratory choices of Cory's shearwaters are not driven by shifts in prevailing air currents. *Scientific Reports*, 8: 3376, Doi: 10.1038/s41598-018-21608-2

Dias, M. P., Granadeiro, J. P., Catry, P., 2012. Do seabirds differ from other migrants in their travel arrangements? On route strategies of Cory's shearwater during its trans–equatorial journey. *Plos One*, 7(11): e49376, Doi: 10.1371/journal.pone.0049376

eBird, 2017. *eBird Basic Dataset*. Version: EB_reINov–2017. Cornell Lab of Ornithology, Ithaca, New York.

Frederiksen, M., Edwards, M., Richardson, A. J., Halliday, N. C., Wanless, S., 2006. From plankton to top predators: bottom–up control of a marine food web across four trophic levels. *Journal of Animal Ecology*, 75: 1259–1268, Doi: 10.1111/j.1365-2656.2006.01148.x

Friedman, J. H., 1999a. Greedy function approximation: a gradient boosting machine (Technical report.). Department of Statistics, Stanford University.

– 1999b. Stochastic Gradient Boosting. *Computational Statistics and Data Analysis,* 38: 367–378.

Friedman, J. H., Hastie, T., Tibshirani, R., 2001. *The Elements of Statistical Learning*. Springer Series in Statistics New York, NY, USA.

Gareth, J., Witten, D., Hastie, T., Tibshirani, R., 2014. *An Introduction to Statistical Learning*. Springer, New York.

Genovart, M., Bécares, J., Igual, J.–M., Martínez–Abraín, A., Escandell, R., Sánchez, A., Rodríguez, B., Arcos, J. M., Oro, D., 2018. Differential adult survival at close seabird colonies: The importance of spatial foraging segregation and bycatch risk during the breeding season. *Global Change Biology*, 24: 1279–1290, Doi: 10.1111/gcb.13997

González–Solís, J., Felicisimo, A., Fox, J., Afanasyev, V., Kolbeinsson, Y., Muñoz, J., 2009. Influence of sea surface winds on shearwater migration detours. *Marine Ecological Progress Series*, 391: 221–230, Doi: 10.3354/meps08128

Gouraguine, A., Moranta, J., Ruiz–Frau, A., Hinz, H., Reñones, O., Ferse, S., Jompa, J., Smith, D., 2019. Citizen science in data and resource–limited areas: A tool to detect long–term ecosystem changes. *Plos One*, 14: e0210007, Doi: 10.1371/journal.pone.0210007

Guilford, T., Wynn, R., McMinn, M., Rodríguez, A., Fayet, A., Maurice, L., Jones, A., Meier, R., 2012. Geolocators reveal migration and pre–breeding behaviour of the critically endangered balearic shearwater *Puffinus mauretanicus*. *Plos One*, 7: 1–8, Doi: 10.1371/journal.pone.0033753

Hilbe, J. M., 2014. *Modeling Count Data*. Cambridge University Press, Cambridge, Doi: 10.1017/CBO9781139236065

Jones, A., Wynn, R. B., Yésou, P., Thébault, L., Collins, P., Suberg, L., Lewis, K. M., Brereton, T. M., 2014. Using integrated land– and boat–based surveys to inform conservation of the Critically Endan-

gered Balearic shearwater *Puffinus mauretanicus* in northeast Atlantic waters. *Endangered Species Research*, 25: 1–18, Doi: 10.3354/esr00611

Katzner, T. E., Arlettaz, R., 2020. Evaluating Contributions of Recent Tracking–Based Animal Movement Ecology to Conservation Management. *Frontiers in Ecology and Evolution*, 7: 519, Doi: 0.3389/fevo.2019.00519

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W., Yu, J., Damoulas, T., Gomes, C., 2013. A Human/Computer Learning Network to Improve Biodiversity Conservation and Research. *AI Magazine*, 34: 10–20.

Kemp, M. U., Emiel van Loon, E., Shamoun–Baranes, J., Bouten, W., 2012. RNCEP: global weather and climate data at your fingertips. *Methods in Ecology and Evolution,* 3: 65–70, Doi: 10.1111/j.2041-210X.2011.00138.x

Kuhn, M., 2007. *Variable importance using the caret package*, https://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/caret/caretVarImp.pdf

– 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28: 1–26, Doi: 10.18637/jss.v028.i05

Louzao, M., García, D., Rodríguez, B., Abelló, P., 2015. Evidence of krill in diet of Balearic Shearwaters Puffinus mauretanicus. *Marine Ornithology* 43: 49–51.

Louzao, M., Hyrenbach, K. D., Arcos, J. M., Abelló, P., de Sola, L. G., Oro, D., 2006. Oceanographic habitat of an endangered Mediterranean Procellariiform: implications for marine protected areas. *Ecological Applications*, 16: 1683–1695, Doi: 10.1890/1051-0761(2006)016[1683:OHOAEM]2.0.CO;2

Luczak, C., Beaugrand, G., Jaffré, M., Lenoir, S., 2011. Climate change impact on Balearic shearwater through a trophic cascade. *Biology Letters*, 7: 702–705, Doi: 10.1098/rsbl.2011.0225

Marshall, A. J., Serventy, D. L., 1959. Experimental Demonstration of an Internal Rhythm of Reproduction in a Trans–Equatorial Migrant (The Short–Tailed Shearwater *Puffinus tenuirostris*). *Nature*, 184: 1704–1705, Doi: 10.1038/1841704a0

Martín, B., Onrubia, A., Ferrer, M., 2016. Migration timing responses to climate change differ between adult and juvenile white storks across Western Europe. *Climate Research*, 69: 9–23, Doi: 10.3354/cr01390

– 2019. Endemic shearwaters are increasing in the Mediterranean in relation to factors that are closely related to human activities. *Global Ecology and Conservation*: e00740, Doi: 10.1016/j.gecco.2019.e00740

Martín, B., Onrubia, A., González–Arias, J., Vicente–Vírseda, J. A., 2020. Citizen science for predicting spatio–temporal patterns in seabird abundance during migration. *Plos One*, 15: e0236631, Doi: 10.1371/journal.pone.0236631

Mateos, M., Arroyo, G. M., 2011. Ocean surface winds drive local–scale movements within long–distance migrations of seabirds. *Marine Biology*, 158: 329–339, Doi: 10.1007/s00227-010-1561-y

Møller, A., Jennions, M. D., 2002. How much variance can be explained by ecologists and evolutionary biologists? *Oecologia*, 132: 492–500, Doi: 10.1007/s00442-002-0952-2

Mourino, J., Arcos, F., Salvadores, R., Sandoval, A., Vidal, C., 2003. Status of the Balearic shearwater (*Puffinus mauretanicus*) on the Galician coast (NW Iberian Peninsula). *Scientia Marina*, 67: 135–142.

Natale, F., Gibin, M., Alessandrini, A., Vespe, M., Paulrud, A., 2015. Mapping Fishing Effort through AIS Data. *Plos One*, 10: e0130746, Doi: 10.1371/journal.pone.0130746

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A. F., Miller, P. I., Louzao, M., 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation,* 156: 94–104, Doi: 10.1016/j.biocon.2011.11.013

Pearce, J. L., Boyce, M. S., 2006. Modelling distribution and abundance with presence–only data. *Journal of Applied Ecology*, 43: 405–412, Doi: 10.1111/j.1365-2664.2005.01112.x

Pérez–Roda, A., Delord, K., Boué, A., Arcos, J. M., García, D., Micol, T., Weimerskirch, H., Pinaud, D., Louzao, M., 2017. Identifying Important Atlantic Areas for the conservation of Balearic shearwaters: Spatial overlap with conservation areas. Deep Sea Research Part II: *Topical Studies in Oceanography*, 141: 285–293, Doi. 10.1016/j.dsr2.2016.11.011

R Core Team, 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Available online at https://www.R-project.org/.

Ramos, R., Granadeiro, J. P., Nevoux, M., Mougin, J.–L., Dias, M. P., Catry, P., 2012. Combined Spatio–Temporal Impacts of Climate and Longline Fisheries on the Survival of a Trans–Equatorial Marine Migrant. *Plos One,* 7: e40822, Doi: 10.1371/journal.pone.0040822

Ridgeway, G., 2005. Generalized Boosted Models: A Guide to the GBM Package. *Compute*, 1: 1–12.

Robinson, J., Dornelas, M., Ojanguren, A. F., 2013. Interspecific synchrony of seabird population growth rate and breeding success. *Ecology and Evolution,* 3: 2013–9, Doi: 10.1002/ece3.592

Schain, M., 2015. *Machine Learning Algorithms and Robustness*. Tel–Aviv University, Tel–Aviv.

Strasser, B., Baudry, J., Mahr, D., Sanchez, G., Tancoigne, E., 2019. "Citizen Science"? rethinking science and public participation. *Science and Technology Studie*s, 32: 52–76.

Tsikliras, A. C., Licandro, P., Pardalou, A., McQuinn, I. H., Gröger, J. P., Alheit, J., 2019. Synchronization of Mediterranean pelagic fish populations with the North Atlantic climate variability. *Deep Sea Research Part II: Topical Studies in Oceanograph*y, 159: 143–151, Doi: 10.1016/j.dsr2.2018.07.005

Visbeck, M. H., Hurrell, J. W., Polvani, L., Cullen, H. M., 2001. The North Atlantic Oscillation: Past, present, and future. *Proceedings of the National Academy of Sciences USA*, 98: 12876, Doi: 10.1073/pnas.231391598

Votier, S. C., Bearhop, S., Attrill, M. J., Oro, D., 2008. Is climate change the most likely driver of range expansion for a critically endangered top predator

in northeast Atlantic waters? *Biology Letters*, 4: 204–205, Doi: 10.1098/rsbl.2007.0558

Wakefield, E., Phillips, R., Matthiopoulos, J., 2009. Quantifying habitat use and preferences of pelagic seabirds using individual movement data: A review. *Marine Ecology – Progress Series*, 393, Doi: 10.3354/meps08203

Wood, S. N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B*, 73: 3–36.

Wynn, R. B., Josey, S., Martin, A., Johns, D., Yésou, P., 2007. Climate–driven range expansion of a critically endangered top predator in northeast Atlantic waters. *Biology Letters*, 3: 529–532, Doi: 10.1098/rsbl.2007.0162

Yen, P. P. W., Sydeman, W. J., Hyrenbach, K. D., 2004. Marine bird and cetacean associations with bathymetric habitats and shallow–water topographies: implications for trophic transfer and conservation. *Journal of Marine Systems*, 50: 79–99, Doi: 10.1016/j.jmarsys.2003.09.015

Yésou, P., 2003. Recent changes in the summer distribution of the Balearic shearwater *Puffinus mauretanicus* off western France. *Scientia Marina*, 67: 143–148.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., Smith, G. M., 2009. GLM and GAM for Count Data. In: *Mixed effects models and extensions in ecology with R. statistics for Biology and Health:* 209–243 (Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, Graham M. Smith, Eds.*)*. Springer, New York. Doi: 10.1007/978-0-387-87458-6_9