A Machine Learning Approach for Animal Trajectory Classification

J.M. Hernández[1], J. P. Rodríguez[2], A. M. M. Sequeira[3] and V.M. Eguíluz[1]

[1]IFISC (CSIC-UIB) Palma de Mallorca – Spain.
[2]IMEDEA (CSIC-UIB), Esporles – Spain.
[3]UWA Oceans Institute, Indian Ocean Marine Research Centre, University of Western Australia – Australia.

jorgemedina@ifisc.uib-csic.es

Abstract

The ocean is the largest ecosystem on Earth where diverse human activities threaten marine life. Thus, knowing how, when, where and why animals move is important for their conservation. As a result of the study of marine animal movement through tracking devices during the past decades, we have collected a large database of around 13000 individual trajectories from more than 100 species, which can be analyzed via data-driven methods. Since its potential remains generally unexplored under these novel techniques, our goal will be to assess their performance and adequateness through the **classification of species associated with spatio-temporal points** (latitude, longitude, time). When shifting the trajectories to a common origin, we find that the initial accuracy of 88% falls to 66%, indicating that while the initial location is a useful feature, the **algorithms are able to extract information from the shape of the trajectory**. Furthermore, **performance is robust to noise** (artificially generated trajectories) and through the **error analysis** we are able to provide insight for **identifying corrupted or inaccurate data**, which can be useful for determining potential flaws in the data collection.

Methodology

Metrics and training procedure

The performance in the multiclass classification task is measured by the accuracy, estimated as the **average accuracy across stratified K-fold cross-validation splits (K=5)**. Thus, we divide the data set in K folds, training iteratively over the K-1 folds (training set) and evaluating the accuracy in the remaining one (test set). Samples are selected from each species in the same proportion they appear in the dataset, ensuring training and test set have similar distributions. Models are trained for 100 epochs and the weights corresponding to the highest accuracy are stored.

Models

Neural network models representative of the state of the art, such as **ResNet, bidirectional LSTM and InceptionTime**. Other models containing convolutional and/or recurrent layers yield similar results.

Feature selection and feature engineering

Shift to common origin

To determine how much information the models extract from the **shape of the trajectory** rather than the geographical location, the initial position can be shifted to a common origin.

Artificial trajectories

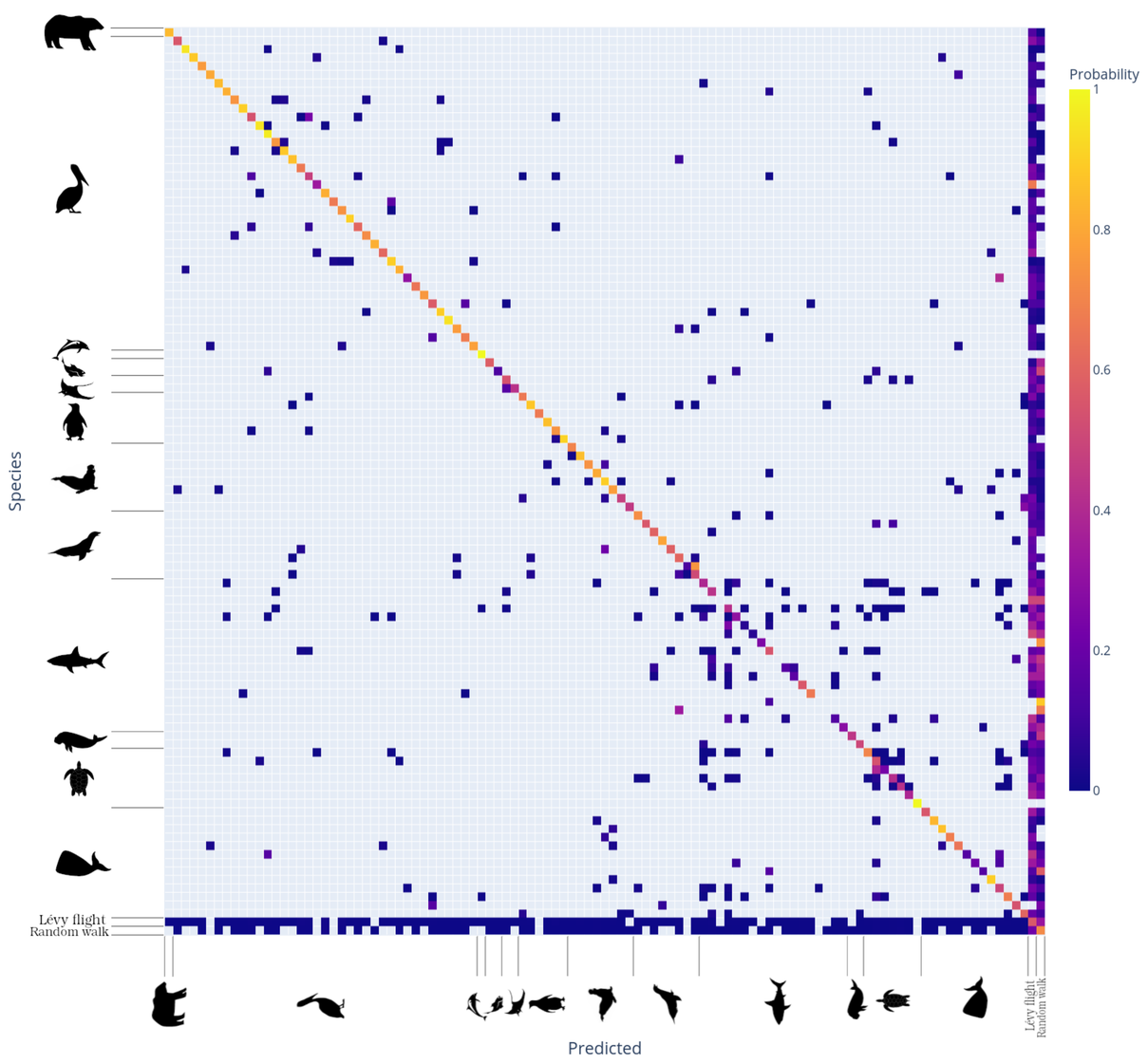
Random walks and Levy flights tuned to mimic original trajectories can be added to the dataset as extra classes. This can be thought of as a form of **noise**.

Environmental variables

Geographical and weather variables can be included. We ensure the absence of multicollinearity, performing a hierarchical clustering of Spearman correlations and a VIF test ($R^2 < 0.8$).

Accuracy

Confusion matrix



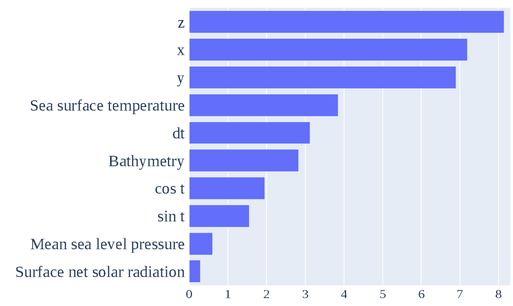
Confusion matrix for the LSTM model. The dataset contains 2500 (~ 20% of the original dataset) artificial trajectories, equally splitted between Levy flights and random walks. The latter have been generated taking a similar origin and sampling over a Levy (Gaussian) distribution with individualized values of scale parameter c (mean μ and std σ) and cutoff to increase the similitude with the animal trajectories. Both types of artificial trajectories share the same median. The artificial trajectories are stratified over the species, i.e. roughly 10% of the trajectories of each species was used to generate them. There is one class for each species' artificial trajectories, (ex. random walk – white shark), which are grouped in the matrix by artificial trajectory type (random walk) for visualization purposes. **Accuracy is high for most animals (86%), but performance significantly drops in the shark taxa.** Furthermore, it is the taxa with the highest probability of being misclassified as an artificial trajectory.

The importance of environmental variables

Classifier	Common origin	Accuracy	Accuracy (E)
ResNet		0.87	0.91
LSTM		0.89	0.88
InceptionTime	$\bar{x}$	0.66	0.85

Accuracy results for several classifiers. (E) indicates the environmental variables have been added. Environmental features can slightly **boost the performance** and contain a significant portion of the information of the **spatial location**, since adding their values evaluated at the initial locations in the common origin setting restores most of the accuracy

Top 10 features by mean absolute SHAP [2,3] value, averaged across all the dataset (the union of the results for the test sets from the k-fold cross validation split (K = 5)). SHAP values indicate the **contribution of each feature to the output of the model**. The only features with impact on the model output of the order of those of the spatiotemporal coordinates are the **sampling period dt** and the environmental variables **sea surface temperature** and **bathymetry**, in agreement with previous results [1],



Assessing data quality through error analysis

To assess why the model fails, we compute **association rules** of the form

$$LHS \longrightarrow \text{Prediction} = \text{wrong}$$

Rules provide insight for the detection of corrupted or inaccurate tracking regarding

Species at specific locations

Tagging systems

Insufficient tracking or low-tech

30% of the misclassifications are explained by rules with confidence **c > 0.95**.

LHS	c	count
{Blue shark, cluster ID 32 }	1	83
{Family Lamnidae, cluster ID 42 }	1	88
{Taxa birds, cluster ID 33 }	0.52	425
{Whales, cluster ID 32}	0.97	104
{Taxa Sharks, tag GLS, years 2006-2009 }	0.97	90
{Unknown sex, tag ARGOS }	0.33	1042
{Tag type PSAT, animals in data set < 54 }	1	132
{tag type SPOT }	0.4	727
{Trajectory data < 79 points }	0.34	1793
{Year < 2002 }	0.36	685

Association rules of the form LHS ⟶ Prediction = wrong, computed using the Apriori algorithm over the ResNet classification results. Total number of trajectories that verify the rule: count × confidence (c). Cluster IDs refer to the geographical location and correspond to clusters computed using HDBSCAN+DBSCAN.

Conclusions

The models are able to achieve **high accuracies** in the marine animal trajectory classification task, with the exception of the shark taxa, likely explained by the difficulty in the tracking process.

Most of it is maintained when we analyze only the shape of the trajectory by shifting all the initial locations to a common origin and results are **robust to the presence of noise** (artificial trajectories).

State-of-the-art algorithms are not only a powerful tool for analyzing animal trajectories, but provide insight to **identify possible flaws in the data collection.**

References

[1] A.M.M. Sequeira et al., *Convergence of marine megafauna movement patterns in coastal and open oceans*, PNAS **115** (2018)
[2] S. I. Lee, and S. M. Lundberg, *A unified approach to interpreting model predictions,* Adv. Neural Inf. Process. Syst. (2018).
[3] S. M. Lundberg et al., *From local explanations to global understanding with explainable AI for trees*, Nat. Mach. Intell., **2** (2020).
Icons from www.flaticon.com