A Machine Learning Approach for Animal Trajectory Classification

Jorge Medina Hernández[1] J. P. Rodríguez[2], A. M. M. Sequeira[3] and Víctor M. Eguíluz[1]
[1]Institute for Cross-Disciplinary Physics and Complex Systems IFISC (CSIC-UIB), 07122 Palma de Mallorca, Spain.
[2]Instituto Mediterráneo de Estudios Avanzados IMEDEA (CSIC-UIB), 07190 Esporles, Spain.
[3]UWA Oceans Institute, Indian Ocean Marine Research Centre, University of Western Australia, Crawley, WA 6009, Australia.

Oceans are environments where a diversity of human activities threaten the 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, there exists now a large database of around 13000 individual trajectories from more than 100 species, susceptible of being 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).

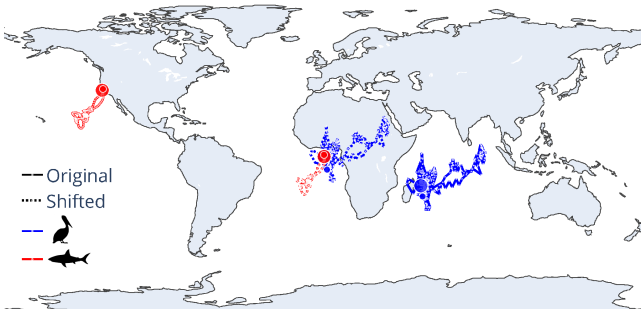

Fig. 1. Original trajectory (solid) and trajectory shifted to the origin $(\theta_0, \phi_0) = (0, 0)$ (dotted) for a wedge-tailed shearwater (blue) and a white shark (red). The initial location is plotted as the biggest point and the final one as the second biggest.

The results in terms of accuracy are shown in Table 1. We find that when trajectories are shifted to a common origin preserving distances and directions (Fig 1), the initial accuracy of 88% falls to 66%, indicating that while the initial location is a useful feature, the algorithms are also able to extract information from the shape of the trajectory. Additionally, we find that including features related to the environment can provide a slight boost in the performance. In particular, the variables with highest impact on the model output (Fig. 2) are the sea surface temperature, the sampling period dt and, in agreement with previous results [1], the bathymetry. Furthermore, these variables 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.

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

Table 1. Accuracy results for several classifiers. (E) indicates the environmental variables have been added.

Lastly, we analyze the errors by computing association rules of the form $LHS \longrightarrow \text{Prediction} = \text{wrong}$ using the Apriori algorithm. We find that approximately 30% of the misclassifications are explained by rules with confidence $c > 0.95$ and involve very specific groups of animals (Table 2). Since the overall accuracy is high, the downfall may be explained by corrupted or inaccurate tracking of the trajectories. This can affect certain species at specific locations (blue shark, whales) and tagging systems (GLS, ARGOS) or types (PSAT, SPOT). Furthermore, the improvements in the tracking systems are reflected on the results: trajectories from 1985 to 2002 account for 1.8% of the data and 14.8% of the errors. Thus, 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.

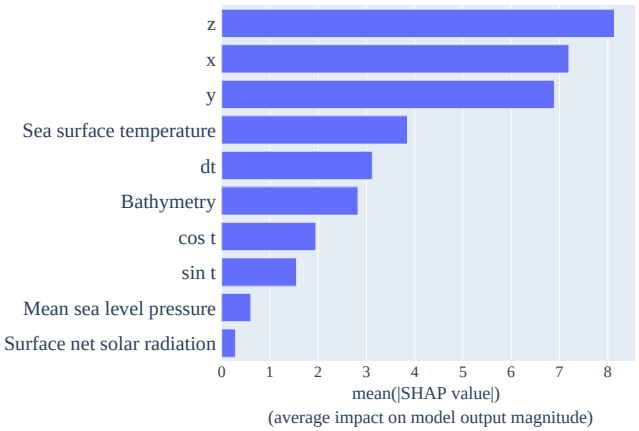

Fig. 2. Top 10 features by mean absolute SHAP [2] value, averaged across all the dataset (the union of the results for the test sets from the k-fold cross validation split ($k = 5$)).

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

Table 2. Several association rules where the right hand side is "Prediction=wrong" for the ResNet classifier, which has an accuracy of 87% (Table 1.) Includes all the dataset (the union of the results for the test sets from the k-fold cross validation split ($k = 5$)). Some rules can provide insight about why model fails in certain trajectories. Total number of trajectories that verify the rule: count $\times$ confidence. Cluster IDs refer to the geographical location and correspond to clusters computed using HDBSCAN+DBSCAN.

[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).