Learning Fishing Information from AIS Data

Gerard Pons Recasens
Universitat Politècnica de Catalunya, BarcelonaTech
Barcelona, Spain
gerard.pons.recasens@estudiantat.upc.edu

Besim Bilalli
Universitat Politècnica de Catalunya, BarcelonaTech
Barcelona, Spain
bbilalli@essi.upc.edu

Alberto Abelló
Universitat Politècnica de Catalunya, BarcelonaTech
Barcelona, Spain
aabello@essi.upc.edu

Santiago Blanco Sánchez
Iberconsa
Vigo, Spain
santiago.blanco@iberconsa.com

ABSTRACT

The Automatic Identification System (AIS) allows vessels to emit their position, speed and course while sailing. By international law, all larges vessels (e.g., bigger than 15m in Europe) are required to provide such data. The abundance and free availability of AIS data has created a huge interest in analyzing them (e.g., to look for patterns of how ships move, detailed knowledge about sailing routes, etc.). In this paper, we use AIS data to classify areas (i.e., spatial cells) of the South Atlantic Ocean as productive or unproductive in terms of the quantity of squid that can be caught. Next, together with daily satellite data about the area, we create a training dataset where a model is learned to predict whether an area of the Ocean is productive or not. Finally, real fishing data are used to evaluate the model. As a result, for blind movements (i.e., with no information about real catches in the previous days), our model trained on data generated from AIS obtains a precision that is 18% higher than the model trained on actual fishing data – this is due to AIS data being larger in volume than fishing data, and 36% higher than the precision of the actual decisions of the ships studied. The results show that despite their simplicity, AIS data have potential value in building training datasets in this domain.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

AIS data, Environmental data, Analytics

1 INTRODUCTION

Finding productive fishing areas in the Ocean is a difficult task. Historically, depending on the targeted species, different kinds of techniques or relevant environmental variables have been used to find potentially productive areas (e.g., with abundance in fish). Typically, once a good location is found, the vessels move towards the same area and tend to fish together, and thus the difficulty lies in finding the initial location with no information about the actual catches in the area for the previous days. We refer to those movements as *blind movements* since they do not rely on any information about the actual catches to decide for their next fishing location. In Figure 1, a graphical representation of real vessel movements (i.e., vessels fishing squids) is shown. As observed, when a good fishing location is found the vessels tend to fish together in the same area, but once the captures fall below a given threshold, the vessels move to other locations. Taking the decision of where to go next is a challenge.

To address the challenge of blind movements and the overall problem of fishery management, recent works have studied the possibility of building predictive presence/absence models using information on physical and environmental conditions. That is, models have been trained on top of environmental variables to predict if fish are present or not in a given area [11, 19]. Typically, the goal is to prevent overfishing, bycatch or providing forecasting solutions that allow for dynamic fishery management [3, 13]. That is, managing fisheries based on real-time predictions of the distribution of marine species. Some works go one step further and study the possibility of predicting the actual catch per unit effort on a fishing area. Thus, building more complex models that are not only capable of predicting the presence of fish, but also the effort required to catch them (e.g., potential quantity) [1]. Regardless of the complexity of the models, one of the biggest challenges these works face is the scarcity of data. Generally, actual real catch data from different ships for different years are used to calculate the catch per unit effort and classify fishing areas as productive or not. However, these data are typically not enough since they often cover a small portion of the Ocean for a small number of years and are difficult to obtain for the entire fleet. To remedy this problem, some works have used satellite imagery data in order to extract fishing information [23]. That is, images have been analyzed to spot the areas where many vessels fish together, with the pretense that those areas are potentially abundant in fish. However, this is a tedious task, not easily scalable and not easily applicable to different types of species.

In this paper, we study how AIS data can be used as a source of information for classifying areas of the Ocean as productive for fishing. AIS data are simple and freely available, yet we show that they provide a valuable source of information which together with environmental data from satellites can be used to build large training datasets for learning models that are capable of predicting areas that are abundant in fish. In particular, we showcase the problem using data about vessels that fish *Illex argentinus*, or the Argentine short-finned squid, which is an important specie in the Patagonian shelf ecosystem in the Southwest Atlantic [18].

Contributions. The main contributions of this paper can be summarized as follows:

- We develop a simple and efficient rule-based method for generating valuable training datasets for fishing predictions from abundantly available AIS data.

The final publication is available at ACM via http://dx.doi.org/10.1145/3557921.3565543

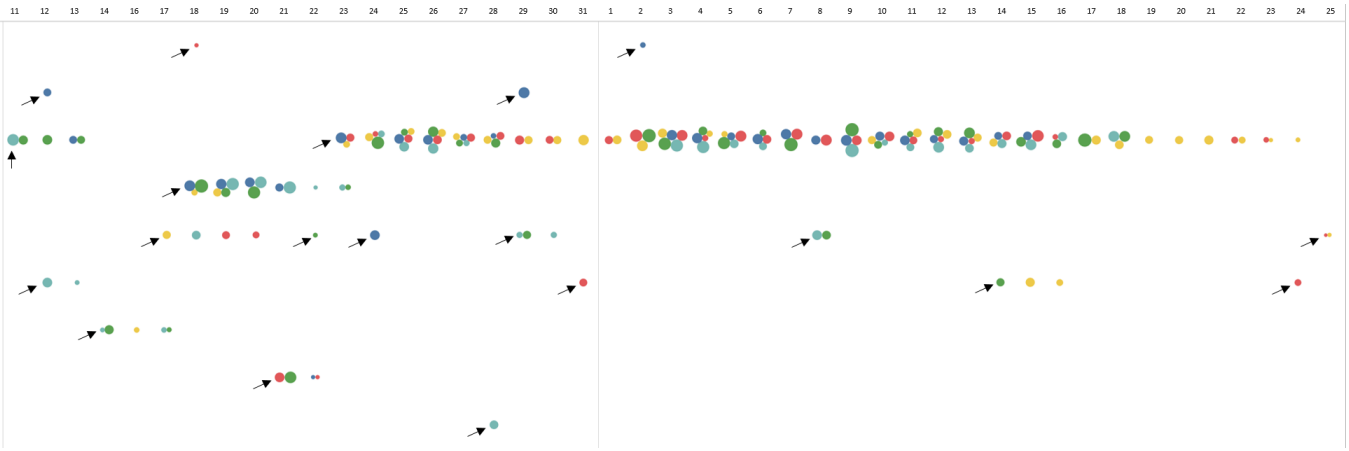

Figure 1: Daily movements of ships for two months. Rows denote fishing areas of size $1^\circ \times 1^\circ$. Columns represent days. Colored circles represent vessels and their size represents the normalized quantity (i.e., kilograms per fishing line) of squid captured. The movements marked by arrows are considered blind movements.

Table 1: Example of the information contained in the fishing reports.

Vessel	Lines	Day	Latitude	Longitude	Species	Size	Quantity (kg)
—	60	02/25/2019	44.52°	63.58°	Squid	M	2339.67

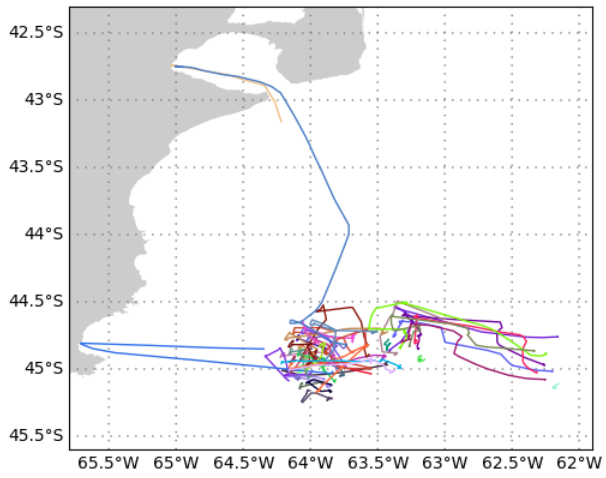

Figure 2: Trajectories of the entire squid vessel fleet sailing/fishing in close proximity on January 22, 2020. Colors indicate vessels.

- We demonstrate that training datasets generated out of AIS data can be effectively used to learn predictive models in a real fishing case in the South Atlantic Ocean. In particular,
  - Models trained over AIS data obtain a precision that is 18% higher than the models trained on actual fishing data.
  - The predictions are 36% more precise than the actual decisions of the studied vessels.

The remainder of this paper is organized as follows. In Section 2, we describe the data sources used in our study. In Section 3, we explain the preprocessing applied over the data. In Section 4, we discuss the developed predictive models and the results obtained after their evaluation. In Section 5, we discuss the Related Work and finally in Section 6, we provide the Conclusions.

2 DATA SOURCES

The different data sources used in this study can be grouped into three categories: Fishing data, AIS data and Environmental data. The first two allow to label geographical locations in terms of fishing productivity while the latter is used to enrich the information of those locations with environmental variables. In particular, the geographical area under study is conceptually viewed as a grid and the data sources are used to enrich each cell of the grid with information from the domain. The Fishing data provide information about real catches in each cell, the AIS data provide vessel movement information through the cells and Environmental data provide the physical conditions of the Ocean in each cell. Hereunder, an overview of the content of the different sources is presented.

2.1 Fishing Data

Fishing reports are provided by the Iberconsa company. Apart from technical details about the vessel, they contain information about the location of fishing sessions and the kilograms caught per specie, per size and per day (see Table 1 for an example). Only the reports from squid vessels were selected, which correspond to six vessels

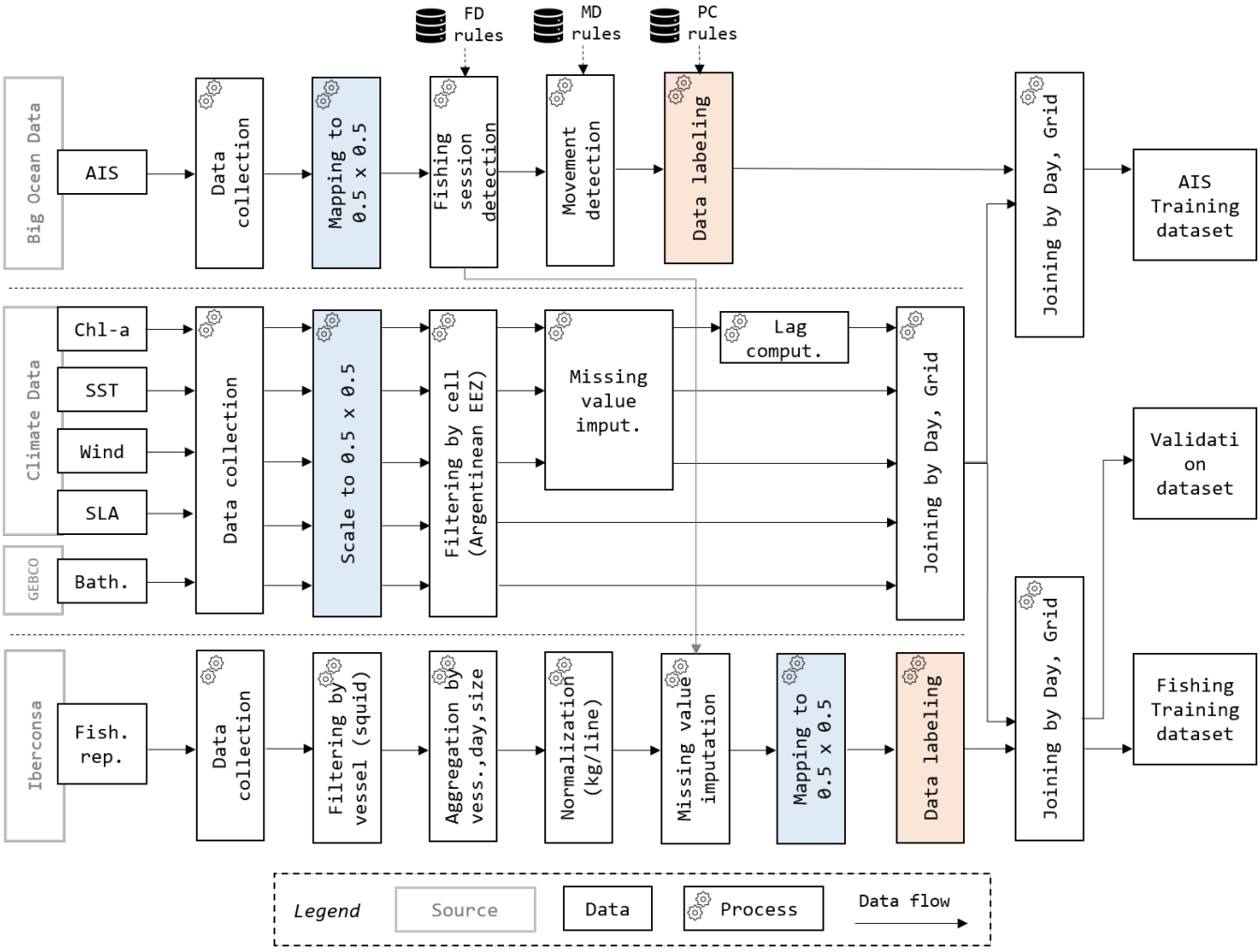

Figure 3: Overview of the preprocessing pipelines for all the data sources. Pipelines for AIS data, Environmental data and Fishing reports are shown on the top, middle and bottom, respectively.

from the entire fleet of the company. Regarding the time period, they extend from 2017 to 2022, and they correspond to the South Atlantic Ocean squid season period, which usually takes place from January to June. Considering the grid view over the geographical area, the fishing data are relevant since they can be used to classify the cells as productive or unproductive. However, note that these reports exist only for a few vessels, hence despite being rich in information they are low in volume and coverage of the vast Ocean.

2.2 AIS Data

AIS is an automatic tracking system for vessels, which provides information about their speed, course and position (i.e., latitude and longitude) through the day in iregular intervals. Data provided by a vessel tracking platform[1], which at the time of the report covers the 2019-2021 period, was used to calculate fishing sessions, detect vessel movements, and impute missing information in the daily catch reports. From it, special attention was paid to the movements of the more than 75 squid vessels that fish each season in the Argentinian Economic Exclusive Zone (EEZ). In Figure 2, we provide a snapshot of the vessel movements for a given day. It is obvious that squid vessels tend to fish in nearby locations. That is, depending on the period of the season, out of all the Argentinian EEZ, fishing is done only around a handful of cells, resulting in little catch information available for other zones.

Considering the grid view and studying movement patterns from AIS data one can use them to classify the cells as productive or unproductive. This may be an approximation, however, given that AIS data are available for the entire fleet,[2] they provide a rich source of information that can compensate for the low volume in terms of fishing reports. The latter is demonstrated in Section 4.

[1]https://www.bigoceandata.com

[2]Here we mean the entire fleet of vessels that fish squid in the Argentinian EEZ, regardless of who owns them.

2.3 Environmental Data

The purpose of using environmental satellite data is to expand the information of each fishing session with potentially relevant environmental variables, aiming at capturing patterns or drawing conclusions from the results in the modeling phase. These data are provided by open source datasets from different international institutions, and as explained in Section 3, they were assessed carefully to know which one satisfies best the study's needs. To select which environmental variables were to be included in the models, domain knowledge was combined with conclusions and practices from scientific and research publications [12, 21]:

- **Chlorophyll a (Chl-a)**: chlorophyll levels are directly associated with phytoplankton levels, which are the base of the marine food chain. Hence, variations of the chlorophyll concentration in different zones could have a relation with the abundance of species.
- **Sea Surface Temperature (SST)**: different fish species have temperature ranges in which they feel more comfortable. Moreover, slight variations in temperature have been reported to increase fish's activity.
- **Bathymetry**: it gives a measure of the depth of the ocean. Various studies have found a direct relationship between bathymetry and the distribution and behavior of species. Contrary to the other variables, this has been considered as atemporal.
- **Sea Level Anomaly (SLA)**: in upwelling zones, water from the sea floor raises up, bringing with it nutrients that could be used by phytoplankton, creating an ideal environment for fish to find nutrients. The contrary situation occurs in downwellings.
- **Wind**: certain wind intensities can induce the upwelling effect, as well as the creation of currents, which provoke the movement of both fish and nutrients.

In addition to these, the **Lunar Phase** was considered too. Due to the techniques of squid fishing, the productivity can be heavily affected by the moon phase: with low luminosity (i.e. new moon) the vessel lamps whose aim is to attract squids could be more efficient; moreover, the moon phase could also have an effect on tides. However, when the final task of modeling is to find productive zones given a day, as the lunar phase is constant for all the cells, it becomes an uninformative variable. Thus, it has not been used as a variable in the final models.

3 DATA PREPROCESSING

Before explaining in detail the data preparation steps, a general view of which are the needs for the modeling process should be presented. On the one hand, fishing reports are used to classify a fishing session as productive or unproductive depending on the kilograms fished. On the other hand, heuristic rules are defined to detect fishing sessions from the AIS data and also label them according to their productivity. Finally, environmental data are needed for all the zones where these fishing sessions took place, both from the fishing sources and the ones created from AIS data. An overview of the entire preprocessing pipeline, whose details are explained in the following sections is depicted in Figure 3. The final goal is to train classification models that predict whether a fishing cell is productive or not. It should be noted that correctly defining the dimensions of the fishing cell is crucial for producing informative outputs. This has been defined in concordance with the suggestions from the expert team regarding vessel velocities and squid detection capabilities and also by taking into account the internal division of the fishing area. Hence, the Argentinian EEZ has been divided into $0.5° \times 0.5°$ cells as shown in Figure 4, which are small enough to address vessels there, but big enough for an adequate number of instances to be registered in all of them, enabling the creation of predictive models.

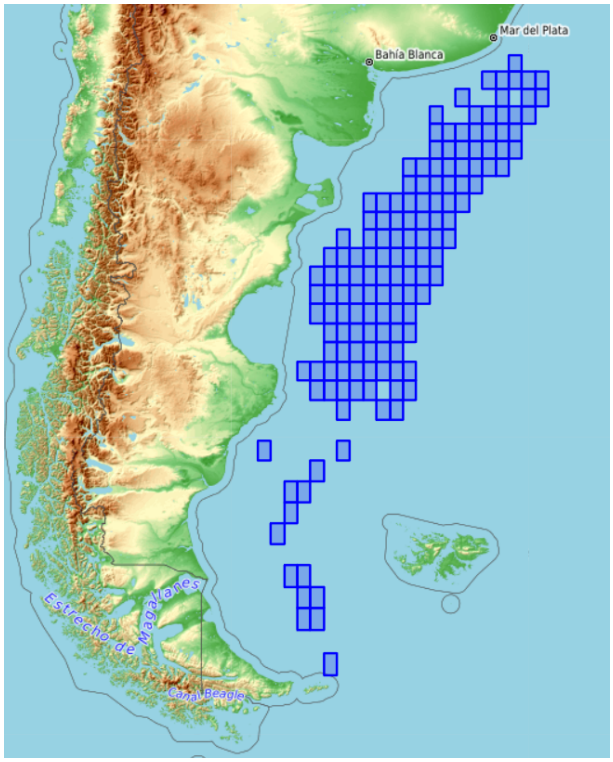

Figure 4: Distribution of cells of size $0.5° \times 0.5°$ for which fishing reports for Illex argentinus exist. Patagonian Shelf, Southwest Atlantic.

3.1 Fishing Data Preprocessing

The preprocessing applied over the Fishing data is depicted in the bottom part of Figure 3. Once the data is collected, vessels are filtered to only those that fish squid. Next, as the daily catches are reported by vessel, species and size, an aggregation per vessel of the daily total kilograms for different sizes has been calculated. Furthermore, as vessels have different technical characteristics, the fishing reports had to be transformed in order to become comparable (i.e., in two areas with the same productivity, a bigger vessel would fish more kilograms than a small one). Hence, a normalization step of calculating the kilograms per fishing line (i.e., the cords with baited hooks attached) for each vessel is performed. Additionally, for the days in which there is missing information about the fishing locations, the results of the Fishing Session Detection explained

in the next subsection are used to impute the position. Next, the locations of all the fishing sessions are mapped into the defined cells and finally the generated data is used for labeling the cells (see Section 3.4).

3.2 AIS Data Preprocessing

As explained before, only the movements of squid vessels are of interest, thus the entire squid fleet (consisting of 78 different vessels on average every year, both from the company and not) are processed accordingly in order to extract valuable knowledge from them. To select the squid fleet vessels, the IMO numbers from the AIS reports are matched with the annual reports provided by the Argentinian National Institute of Fishery Research and Development (INIDEP).[3] The different preprocessing steps depicted on the top part of Figure 3 are the result of different iterations done with the recommendations of domain experts and supported by real fishing data. Once the data is collected, a mapping over the grid with spatial cells of size $0.5^\circ \times 0.5^\circ$ is performed. Next, two important preprocessing steps are applied:

3.2.1 Fishing Session Detection. The first crucial step is to detect when a squid vessel is fishing, for which a collection of rules created alongside the team with an expertise on the domain are defined:

- *FD1*: The velocity of the vessel should be under 1.5 knots (nautical miles per hour), as during the fishing sessions the engines are shot down and the only movement is caused by drift.
- *FD2*: The time of the day should be between dusk and dawn, as the squid fishing technique is based on attracting them with powerful incandescent lights during the night.
- *FD3*: The location should comply with the fishing closure periods (which are different for every year and are decided by governmental institutions) and not be in locations close to harbors.
- *FD4*: The aggregated duration of the session (i.e. the total temporal length of consecutive AIS registers satisfying FD1, FD2 and FD3) should exceed 3 hours, since smaller durations could introduce noise due to being originated by non-fishing related activities.

With this procedure, the AIS reports have been transformed into a collection of periods of certain duration — derived from an ordered list of AIS reports, tagged as *Fishing* or *No Fishing* for each vessel (see Figure 6 for an example), whose locations have also been mapped into the cells shown in Figure 4.

Finally, notice that the data generated here, as noted in Figure 3 and Section 3.1, is also used to impute the missing fishing locations inside the Fishing data.

3.2.2 Movement Identification. Once the trajectory of a vessel is marked as *Fishing* or *No Fishing*, the second step is to identify fishing location changes. To this end, a set of additional rules are defined:

- *MI1*: Both start and end locations must have been previously tagged as Fishing.
- *MI2*: The distance between start and end locations must be greater than 30 miles, since shorter movement distances are not considered as a fishing zone change.

The above mentioned steps generate the base information required for later labeling the cells within the trajectory of the movements as productive or not (see Section 3.4) and also for identifying *blind movements* (see Section 4.1).

3.3 Environmental Data Preprocessing

The environmental variables can be extracted from different data sources, each of them having different spatial resolution and temporal availability. The former does not play an important role in the decision making process, as the defined working resolution ($0.5^\circ \times 0.5^\circ$) is coarser than any of those provided by the studied sources. However, as these data are collected from sensors in satellites or vessels, one should previously analyze them to assess their quality and usefulness. Concretely, the main problem we found were for the Chl-a, SST and Wind variables, as the sensors capturing them are not capable of gathering data through dense clouds, resulting in zones of incomplete information for some days. This problem is shown in Figure 5, where the SST registers in the Argentinian coast are shown for a given day from various sources. The most important problem that can be seen is that as the missing values are clustered, an imputation procedure by spatial interpolation is not possible in the vast majority of the cells. Moreover, one should also take into account that different sources can use their own satellites with different sensors and produce the measurements at different times of the day, hence combining the data from more than one source would not be an appropriate procedure. The final solution used to solve the completeness problem was inspired by techniques used by domain experts, which consists in using in a weighted (w_j) manner (e.g., closer days have higher weight), values (T_j) for the same position from three days before and three days after (j). Hence, in the case of having a missing value for a day:

$$T_i = \begin{cases} T_i, & \text{if } T_i \neq null \\ \frac{\sum_{(i-3)<j<(i+3), j\neq i, T_j \neq null} w_j T_j}{\sum_{(i-3)<j<(i+3), j\neq i, T_j \neq null} w_j}, & \text{otherwise} \end{cases}$$

Regarding the data source selection, the decision was to work with two different sources. The first choice was Climate Data[4], as apart from providing data for Wind, Chl-a, SLA and SST by means of the same API, although separated in different files per variable, it showed a good trade-off between reliability (i.e., some sources provide measurements with their corresponding levels of quality, hence data with low reliability is provided and the decision whether to use it is let to the user), resolution and completeness. For the Bathymetry variable, GEBCO[5] was chosen as it provides very detailed and complete data for all the Ocean.

Furthermore, as explained in more detail in Section 4.2, long range temporal correlations were studied between the environmental variables and the fishing productivity, and a correlation with past Chl-a concentration levels was found (see Figure 7). Consequently, in the final preprocessing stage this lagging information is also

[3] https://www.argentina.gob.ar/inidep

[4] https://cds.climate.copernicus.eu/#!/home

[5] https://www.gebco.net/data_and_products/gridded_bathymetry_data

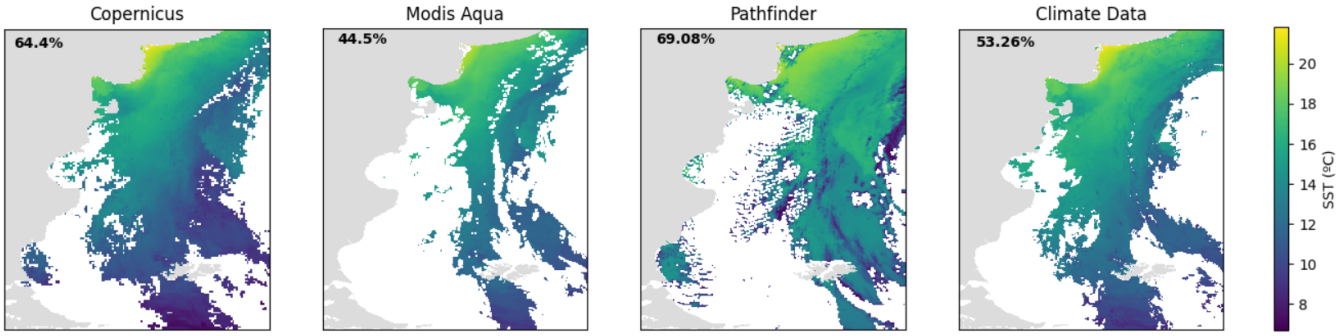

Figure 5: SST registers for the Argentinian coast from different sources for January 6, 2020. The values on the top left corner denote completeness and the white zones correspond to missing values.

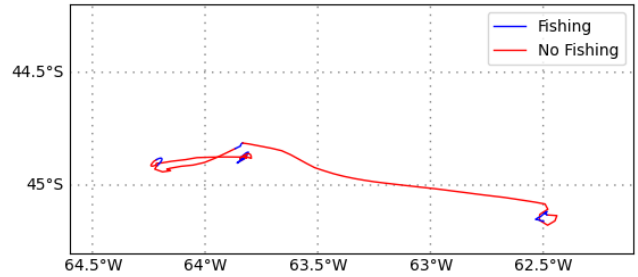

Figure 6: An example of a trajectory of a vessel (from January 19 through Januray 23, 2020) with parts of the trajectory marked as *Fishing* and *No Fishing*, based on the heuristic rules applied.

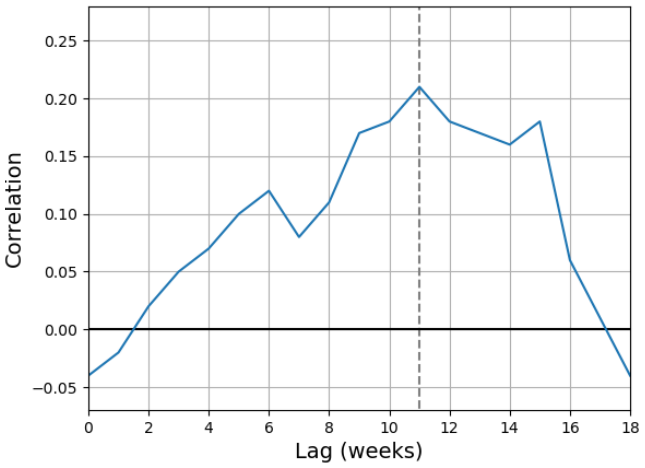

Figure 7: Correlation of the normalized kilograms (kg/line) with different values of Chl-a temporal lag (from 0 to 18 weeks). The dashed line indicates the temporal Chl-a lag used as a model variable, as it corresponds to the maximum correlation value.

computed. To this end, in summary, the environmental preprocessing steps start by scaling the data to the defined grid dimensions ($0.5^\circ \times 0.5^\circ$), selecting the grids corresponding to Argentinian EEZ, solving the completeness problem with the mentioned procedure, calculating the Chl-a lag variable and joining by grid and day the different variables; see Figure 3.

3.4 Data Labeling

With the previous explained transformations applied to the data, they need to be labeled accordingly to be fed to our classification models. Different approaches were taken depending on the origin of the data.

3.4.1 Fishing reports. For the fishing reports, the actual kilograms fished are used to label the cells as productive or not. In particular, a moving threshold for the normalized kilograms fished is defined by calculating a running average for the different weeks of the year using a two week window and the fishing reports from the 2017 to 2021 seasons. This procedure is done because as it can be seen in Figure 8, the squid abundance is not constant throughout the fishing season, hence defining a static threshold would be inappropriate. With it, cells where a vessel fished over the moving threshold are labeled as *Productive* and the cells where they fished below it as *Unproductive*.

3.4.2 AIS data. For AIS data, since the actual kilograms fished are not available, different rules are required to label the cells as productive or not. In particular, rules were defined taking into account movements and fishing sessions described in Section 3.2, and also domain expert suggestions. Thus, a cell is labeled as *Productive* if:

- *PC1*: A vessel moves over the cell complying with the movement criteria defined by rules MI1 and MI2, thus it is fishing there.
- *PC2*: It stays in the same cell for three days or more.

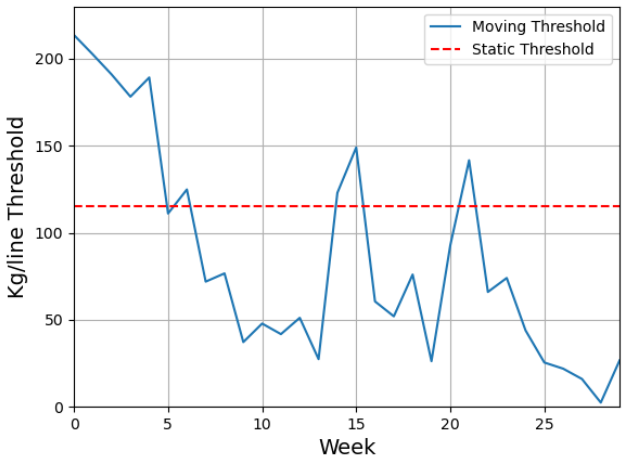

Figure 8: Evolution of the moving threshold (kg/line) throughout the squid season compared with a static threshold defined as the global median.

If *PC1* holds but the vessel moves after just one or two days of being there, the cell is labeled as *Unproductive*. Hence, with these rules it is assumed that a vessel will stay three or more days in a fishing spot if it is productive. Using a lower number of days could result in wrongly labelling the cell, as it is not usual for vessels to change locations immediately after having committed to a long range movement. Note that these rules have been defined using domain knowledge on squid, hence for other species they must be fine tuned depending on the characteristics of the target specie.

4 MODELING

One must notice that after the preprocessing, two labeled datasets are obtained, one from the fishing reports and one coming from the AIS reports (see the final output of Figure 3). The former contains more reliable information, as real quantities fished are used for the labeling, but in terms of volume it is much smaller. On the other hand, for the latter there is a high volume of data but with estimated information, as heuristic rules are used for deriving the labels. Although in machine learning contexts volume usually has a higher weight than quality, two models have been created using the different sources and their performances have been compared. Nevertheless, it must be highlighted that the validation has been done with data from fishing reports from 2022 for both of the models. Thus, they both have been assessed against real fishing data and not against rule derived labels.

The final training datasets are created by joining the labeled data from the respective source (i.e., all the available AIS data for 2019 to 2021 and Fishing data for 2017 to 2021) with the preprocessed environmental variables, and for the test dataset the same procedure has been done but only for Fishing data from 2022. The idea is to assess whether AIS data contain valuable information that can be used to complement the actual fishing data. If this holds, AIS data that are available in abundance can be used not only for detecting the patterns of how vessels move or for obtaining detailed knowledge about sailing routes [2, 14, 15], but they can also be used for building training datasets that can serve other purposes. Moreover, they are not limited to a particular specie and thus can be used for other species too (just adapting the heuristic rules).

Before going into more details on the modeling step, let us recall what is expected from the model and what it will be used for:

- First of all, the model is thought to be used when a *blind movement* is to be attempted (i.e., the vessel has just departed from the port or wants to substantially change locations due to poor fishing results). Moreover, this movement should be uninformed: it is clear that when a vessel captain has information about the presence of squids in his actual or other locations the model suggestions fall in second place (see Section 1, for more details about *blind movements*).
- The models have been evaluated giving greater importance to the precision[6] metric. As the output of the model will suggest productive fishing locations where to move ships, the risk of suggesting a bad spot should be minimized due to its economical impact.
- The model should suggest productive fishing spots throughout all the season, not just clustering them in the best fishing months.

Finally, the data preprocessing and modeling are performed using Python (using *scikit-learn* [17], *pandas* [24] and *netCDF4* libraries) and R (using *raster* [22] library). Yet, the data in its entirety cannot be made publicly available since it contains information about the company. Only the code, data and the models that do not contain/use information about the company are publicly available in Github[7].

4.1 Baseline Definition

As explained before, the models are not to be used when captains have factual information about the presence of squid in certain locations (i.e., the quantity of squid caught at the time of moving towards that location), as it obviously cannot compete against them. Hence, a baseline needs to be decided against which the models can be fairly compared in order to assess their performance. For that, one needs to calculate the ratio of productive movements performed by a vessel. This allows to calculate the precision of decisions over historical fishing data.

To spot a productive movement, a set of simple rules are applied over the fishing reports in combination with the movement identifications:

- *GM1*: A vessel performs a long range movement (i.e. +30 miles) to a cell.
- *GM2*: There is no information available for that cell for the previous day.
- *GM3*: It fishes a quantity over the moving threshold.

Contrarily, if only *GM1* and *GM2* hold, it is considered an unproductive movement.

To this end, for the available historical fishing data, the success rate of finding new productive locations, for the year 2021 (the last year for which data about the entire season is available) is

[6] Given that we are dealing with a binary classification problem of predicting a productive or not productive cell, precision is calculated as the ratio of true productive cells over all locations predicted as productive.
[7] https://github.com/gerardponsrecasens/fishing-information-from-AIS

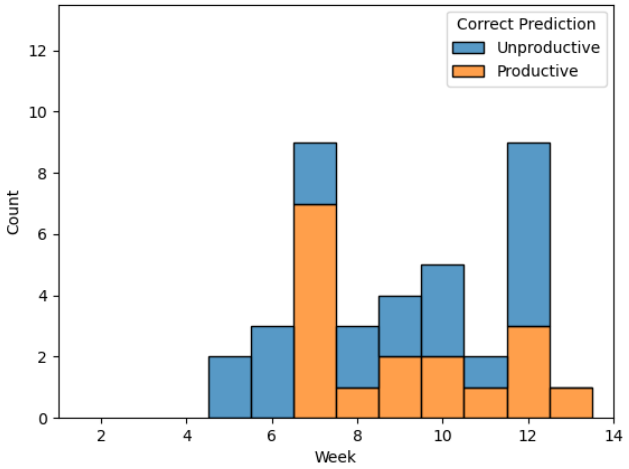

Figure 9: Distribution of the correct predictions when learning a model using Fishing Data.

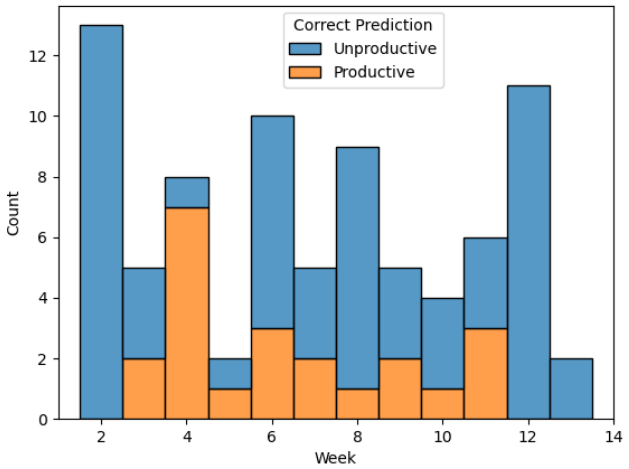

Figure 10: Distribution of the correct predictions when learning a model using AIS Data.

of 50.2%. That is, for *blind movements*, the likelihood of finding a productive cell (i.e., with catches above the threshold) resembles that of tossing a coin. These results illustrate the difficulty of finding new productive areas without a priori information, as explained in Section 1.

Finally, notice that the precision of the decisions (i.e., baseline) can be compared against model predictions, as they are both calculated over movements that do not rely on catch information.

4.2 Modeling with Fishing Data

Obviously, when relying on data about the actual catches (i.e., kilograms caught), one can think of building a model to predict the actual quantity in a given cell (i.e., treating the problem as a regression problem [4]), and then assess whether this translates to a productive cell or not. The other option is to convert the problem into a classification problem [4], where a cell is already labeled as productive or not using data from real catches (see Section 3.4). The latter was our main goal, however, here, we briefly explain our alternative try, since it allowed us to gain some deeper knowledge on the modeling part.

4.2.1 Regression Model. The first naive approach into modeling could be to predict the productivity using regression models trained directly over the normalized fishing reports. Although these would have resulted in very informative models, the lack of data hinders their learning capabilities, and the results are far from acceptable. Yet, performing such an exercise allows to get an understanding of the data and explore patterns and correlations. Concretely, long range correlations of environmental variables with kilograms fished were studied, and although there were no important findings for Wind, SST and SLA, fishing productivity showed a significantly higher correlation with Chl-a concentration with a three months lag, reaching its maximum at concretely eleven weeks prior (see Figure 7). As a consequence, this additional information has been included as an input variable in the different classification models that follow.

4.2.2 Classification Model. Our main goal was to learn a classification model using the labels from the daily fishing reports. For it, different classification models and configurations of them have been studied. The best results were obtained using Random Forest Models [5], for which the precision reached 60.1%. Yet, the model predicting capabilities are limited when it comes to generating correct suggestions for the first weeks of the season. That is, the model is not able to predict enough truly productive cells in the first weeks of the season as shown in Figure 9, where the length of each bar represents the number of correct predictions for *Productive* cells (orange) and *Unproductive* cells (blue). The results shown in Figure 9 were undesirable and thus we followed with developing models over training datasets labeled using AIS data.

4.3 Modeling with AIS Data

In this approach, the data labeled using AIS information was used, and the exact same modeling steps as in Section 4.2.2 were followed. In this case however, the best model found is created with a Voting ensemble [9] of Support Vector Machines (SVM) [8] and Random Forests. With this model the obtained precision is 68.5%, which results in an acceptable increase with respect to the classification model based on Fishing data. Moreover, the correct suggestions are spread out more evenly across the season — see Figure 10, hence resulting to a more informative model. To understand what the model is learning, the feature importances of the Random Forest model used in the ensemble (whose isolated precision was 65.6%) are visualized in Figure 11. It can be observed that the model gave more importance to environmental variables, with the lagging Chl-a variable being the most informative in predicting a productive location. In contrast, for the model obtained in Section 4.2.2, it was the week who took the first place, resulting to the undesirable behavior of producing few (if at all) suggestions of productive cells for certain weeks.

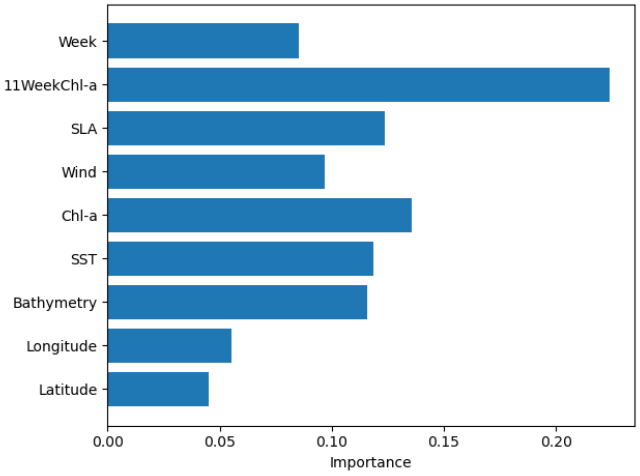

Figure 11: Importance of the different features in the model for the Random Forest Classifier which is part of the Voting Ensemble of the model with AIS data.

4.4 Results

Finally, a comparison of the results of the models with respect to the baseline (50.1%) are shown in Table 2. It can be seen that both models outperform the baseline results, hence they are more precise at suggesting productive fishing locations than the methods and practices currently used when there is no information available. Furthermore, modeling with AIS yields better results than modeling with fishing reports, indicating that in this domain, establishing appropriately defined rules can help overcome the data scarcity problem. In turn, these results suggest that acquiring more fishing data (e.g., getting reports from more vessels or recording them during more years) could help improve the modeling performance, as it contains more precise and valuable information.

Table 2: Summary of the precision of the models and their improvement with respect to the baseline.

Model	Precision	% Imp. over the Baseline
Model Fishing data	60.1%	19.7%
Model AIS data	68.5%	36.4%

5 RELATED WORK

Relationships between species distributions and the physical environment have been observed consistently. They are typically modeled using Species Distribution Models (SDM) [10]. SDMs can be applied across terrestrial, freshwater, and marine environments. In [10], the challenges of SDMs as numerical tools that combine observations of species occurrence or abundance with environmental species are studied. In [20], it is argued that eco-informatics solutions that allow for near real-time prediction of the distribution of highly mobile marine species are an important step towards the maturation of dynamic ocean management and ecological forecasting. Fisheries' observer data (with observer rates of 15%) from the California drift gillnet fishery are used to model the relative probability of occurrence (presence–absence) and catchability (total catch per gillnet set) of broadbill swordfish in the California Current System. In [19], they develop predictive models of Argentine shortfin squid abundance in relation to physical and environmental conditions. Again, fishery and biological data collected by scientific observers aboard commercial trawlers are analysed in relation to physical and environmental factors to establish the spatio-temporal pattern of the species' distribution and quantify the influence of environmental variables.

To prevent overfishing and bycatch, in [11], they create species distribution models for one target specie and three bycatch-sensitive species using both satellite telemetry and fisheries observer data. In [7], using catch per unit effort data of Argentine shortfin squid from squid jigging fleets during the years 2004–2013, they evaluate their variability in abundance and the most possible relationships to migration patterns. In [1], they build models to predict the catch per unit effort in the North-western Adriatic Sea. The area of study is organized in the form of a grid and each cell is given a score of catch per unit effort. Real fishing data and AIS data are combined together to impute the fishing quantity on each cell. Finally, a model is built to learn the relationship between environmental variables and the catch per unit effort. However, the scarce availability of fishing data limits the study to only two years. In [6], a detailed study of how temperature effects the size and maturity of Argentine squid is performed. 20-year time series of fisheries data and monthly temperature data from the key regions associated with Illex argentinus are used.

Generally, data about fishing catches is either scarce or unreliable [16]. Most of the works mentioned above rely on fishery observer data which are limited to small samples or small datasets. There have been studies however, that have looked at the possibility of using imagery satellite data to study the distribution of Argentine squid [23]. Argentine squid is caught by jigging vessels which attract squid using powerful incandescent lights, which are detectable in remotely sensed satellite imagery data. The assumption is that changes in the distribution of the fleet can reflect shifts in the distribution of the squid. To this end, in [23] they have argued for the use of such external data source to label areas of the Ocean as abundant or not. However, this approach is quite complex and we contend that there is a better and simpler alternative of obtaining similar data, that of using heuristics on top of AIS data.

6 CONCLUSIONS

Evidences suggest that catch data updated and disseminated annually by the Food and Agriculture Organization of the United Nations (FAO)[8] on behalf of member countries may considerably underestimate actual fisheries catch [16]. Even more, such data is typically aggregated and thus extrapolating it to the cells of a hypothetical grid over the Ocean is not trivial, if not impossible. A solution to the problem is to get data from individual ships and companies, however such data is not accessible. To this end, we propose a method that exploits the movement data of the vessels

[8] https://www.fao.org/fishery/statistics-query/en/capture/capture_quantity

(i.e., reported via AIS transceivers) in order to label the geographical cells of the Ocean as productive or not at a given point of time. We use data about vessels that catch Argentine short-finned squid in order to evaluate our proposal. In particular, we use heuristic rules on top of AIS data in order to label the geographical cells according to their productivity. Together with environmental variables this provides a training dataset on top of which a predictive model is learned, with the goal of forecasting the most productive cells. Finally, real catch data are used to validate the proposed model and the heuristic rules. As a result, the model trained on top of AIS data performs 19.7% more precisely than the model trained on top of real catch data, and 36.4% more precisely than the currently used methods in the company which is considered as our baseline.

Acknowledgments. We thank the entire Iberconsa team, but in particular Daniel Brunengo for helping us collect the required data and Alvaro Pazos and Jose Dominguez for sharing their valuable knowledge and expertise during the development of the project.

REFERENCES

[1] Pedram Adibi, Fabio Pranovi, Alessandra Raffaetà, Elisabetta Russo, Claudio Silvestri, Marta Simeoni, Amilcar Soares, and Stan Matwin. 2020. Predicting Fishing Effort and Catch Using Semantic Trajectories and Machine Learning. In *Multiple-Aspect Analysis of Semantic Trajectories*, Konstantinos Tserpes, Chiara Renso, and Stan Matwin (Eds.). Springer International Publishing, Cham, 83–99.

[2] Andreas S. Andersen, Andreas D. Christensen, Philip Michaelsen, Shpend Gjela, and Kristian Torp. 2021. AIS Data as Trajectories and Heat Maps *(SIGSPATIAL '21)*. Association for Computing Machinery, New York, NY, USA, 431–434. https://doi.org/10.1145/3474717.3484208

[3] J. R. Beddington, D. J. Agnew, and C. W. Clark. 2007. Current Problems in the Management of Marine Fisheries. *Science* 316, 5832 (2007), 1713–1716. https://doi.org/10.1126/science.1137362 arXiv:https://www.science.org/doi/pdf/10.1126/science.1137362

[4] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Irina Chemshirova, Henk-Jan Hoving, and Alexander Arkhipkin. 2021. Temperature effects on size, maturity, and abundance of the squid Illex argentinus (Cephalopoda, Ommastrephidae) on the Patagonian Shelf. *Estuarine, Coastal and Shelf Science* 255 (2021), 107343. https://doi.org/10.1016/j.ecss.2021.107343

[7] Tsan-Yu Chiu, Tai-Sheng Chiu, and Chih-Shin Chen. 2017. Movement patterns determine the availability of Argentine shortfin squid Illex argentinus to fisheries. *Fisheries Research* 193 (01 Sep 2017), 71–80. https://www.sciencedirect.com/science/article/pii/S0165783617300905

[8] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[9] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *MULTIPLE CLASSIFIER SYSTEMS, LBCS-1857*. Springer, 1–15.

[10] Jane Elith and John R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics* 40, 1 (2009), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159 arXiv:https://doi.org/10.1146/annurev.ecolsys.110308.120159

[11] Elliott L. Hazen, Kylie L. Scales, Sara M. Maxwell, Dana K. Briscoe, Heather Welch, Steven J. Bograd, Helen Bailey, Scott R. Benson, Tomo Eguchi, Heidi Dewar, Suzy Kohin, Daniel P. Costa, Larry B. Crowder, and Rebecca L. Lewison. 2018. A dynamic ocean management tool to reduce bycatch and support sustainable fisheries. *Science Advances* 4, 5 (2018), eaar3001. https://doi.org/10.1126/sciadv.aar3001 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aar3001

[12] Edgar Lanz, Juana Martinez, Manuel Nevarez Martinez, and J.A. Dworak. 2009. Small pelagic fish catches in the Gulf of California associated with sea surface temperature and chlorophyll. *California Cooperative Oceanic Fisheries Investigations Reports* 50 (12 2009), 134–146.

[13] Rebecca Lewison, Alistair J. Hobday, Sara Maxwell, Elliott Hazen, Jason R. Hartog, Daniel C. Dunn, Dana Briscoe, Sabrina Fossette, Catherine E. O'Keefe, Michele Barnes, Melanie Abecassis, Steven Bograd, N. David Bethoney, Helen Bailey, David Wiley, Samantha Andrews, Lucie Hazen, and Larry B. Crowder. 2015. Dynamic Ocean Management: Identifying the Critical Ingredients of Dynamic Approaches to Ocean Resource Management. *BioScience* 65, 5 (03 2015), 486–498. https://doi.org/10.1093/biosci/biv018 arXiv:https://academic.oup.com/bioscience/article-pdf/65/5/486/16648349/biv018.pdf

[14] Giuliana Pallotta, Michele Vespe, and Karna Bryan. 2013. Traffic knowledge discovery from AIS data. In *Proceedings of the 16th International Conference on Information Fusion*. 1996–2003.

[15] Giuliana Pallotta, Michele Vespe, and Karna Bryan. 2013. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy* 15, 6 (2013), 2218–2245. https://doi.org/10.3390/e15062218

[16] Daniel Pauly and Dirk Zeller. 2016. Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining. *Nature Communications* 7, 1 (19 Jan 2016), 10244. https://doi.org/10.1038/ncomms10244

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[18] Graham J. Pierce, Vasilis D. Valavanis, Angel Guerra, Patricia Jereb, Lydia Orsi-Relini, Jose M. Bellido, Isidora Katara, Uwe Piatkowski, João Pereira, Eduardo Balguerias, Ignacio Sobrino, Eugenia Lefkaditou, Jianjun Wang, Marina Santurtun, Peter R. Boyle, Lee C. Hastie, Colin D. MacLeod, Jennifer M. Smith, Mafalda Viana, Angel F. González, and Alain F. Zuur. 2008. A review of cephalopod–environment interactions in European Seas. *Hydrobiologia* 612, 1 (01 Oct 2008), 49–70. https://doi.org/10.1007/s10750-008-9489-7

[19] Sacau, Mar, Pierce, Graham J., Wang, Jianjun, Arkhipkin, Alexander I., Portela, Julio, Brickle, Paul, Santos, María B., Zuur, Alain F., and Cardoso, Xosé. 2005. The spatio-temporal pattern of Argentine shortfin squid Illex argentinus abundance in the southwest Atlantic. *Aquat. Living Resour.* 18, 4 (2005), 361–372. https://doi.org/10.1051/alr:2005039

[20] Kylie L. Scales, Elliott L. Hazen, Sara M. Maxwell, Heidi Dewar, Suzanne Kohin, Michael G. Jacox, Christopher A. Edwards, Dana K. Briscoe, Larry B. Crowder, Rebecca L. Lewison, and Steven J. Bograd. 2017. Fit to predict? Eco-informatics for predicting the catchability of a pelagic fish in near real time. *Ecological Applications* 27, 8 (2017), 2313–2329. https://doi.org/10.1002/eap.1610 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/eap.1610

[21] Allan Stoner. 2004. Effects of environmental variables on fish feeding ecology: Implications for the performance of baited fishing gear and stock assessment. *Journal of Fish Biology* 65 (12 2004), 1445 – 1471. https://doi.org/10.1111/j.0022-1112.2004.00593.x

[22] Robert J. Hijmans Jacob van Etten. 2012. *raster: Geographic analysis and modeling with raster data.* http://CRAN.R-project.org/package=raster R package version 2.0-12.

[23] Claire M. Waluda, Huw J. Griffiths, and Paul G. Rodhouse. 2008. Remotely sensed spatial dynamics of the Illex argentinus fishery, Southwest Atlantic. *Fisheries Research* 91, 2 (2008), 196–202. https://doi.org/10.1016/j.fishres.2007.11.027

[24] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 56 – 61. https://doi.org/10.25080/Majora-92bf1922-00a