



# From multiple aspect trajectories to predictive analysis: a case study on fishing vessels in the Northern Adriatic sea

Bruno Brandoli<sup>1</sup> · Alessandra Raffaetà<sup>2</sup> · Marta Simeoni<sup>2,3</sup> · Pedram Adibi<sup>1</sup> ·  
Fateha Khanam Bappee<sup>1</sup> · Fabio Pranovi<sup>2</sup> · Giulia Rovinelli<sup>2</sup> · Elisabetta Russo<sup>2</sup> ·  
Claudio Silvestri<sup>2</sup> · Amilcar Soares<sup>4</sup> · Stan Matwin<sup>1,5</sup>

Received: 6 August 2021 / Revised: 24 December 2021 / Accepted: 1 February 2022 /  
Published online: 31 March 2022  
© The Author(s) 2022

## Abstract

In this paper we model spatio-temporal data describing the fishing activities in the Northern Adriatic Sea over four years. We build, implement and analyze a database based on the fusion of two complementary data sources: trajectories from fishing vessels (obtained from terrestrial Automatic Identification System, or AIS, data feed) and fish catch reports (i.e., the quantity and type of fish caught) of the main fishing market of the area. We present all the phases of the database creation, starting from the raw data and proceeding through data exploration, data cleaning, trajectory reconstruction and semantic enrichment. We implement the database by using MobilityDB, an open source geospatial trajectory data management and analysis platform. Subsequently, we perform various analyses on the resulting spatio-temporal database, with the goal of mapping the fishing activities on some key species, highlighting all the interesting information and inferring new knowledge that will be useful for fishery management. Furthermore, we investigate the use of machine learning methods for predicting the Catch Per Unit Effort (CPUE), an indicator of the fishing resources exploitation in order to drive specific policy design. A variety of prediction methods, taking as input the data in the database and environmental factors such as sea temperature, waves height and Chlorophyll-a, are put at work in order to assess their prediction ability in this field. To the best of our knowledge, our work represents the first attempt to integrate fishing ships trajectories derived from AIS data, environmental data and catch data for spatio-temporal prediction of CPUE – a challenging task.

**Keywords** Semantic trajectories · Spatio-temporal databases · Fisheries · Machine learning · Predictive models

---

✉ Alessandra Raffaetà  
raffaeta@unive.it

Extended author information available on the last page of the article

## 1 Introduction

The Northern Adriatic Sea area is one of the most exploited areas of the Mediterranean Sea, causing an over-exploitation of the fish resources. Having a clear representation and understanding of the main factors driving such phenomenon is of paramount importance both for ecologists and for local policymakers. In fact, they could use such information for the development of effective fishery management plans, with the aim of making fishing activities sustainable and ensuring a productive and healthy ecosystem.

In this setting, we identify three main tasks that we address in the paper with the goal of improving the knowledge on the North Adriatic sea. The first task is to have a clear and sound representation of the fishing activities by defining and implementing a specific spatio-temporal database of the area of interest. The second task is to analyze the obtained database to gain a deep knowledge about the fishing activities in the Northern Adriatic basin, evaluate the effectiveness of the current fishery management, and detect the spatial distribution of commercial fishery catches. The third task is to investigate prediction methods for fishing activities forecast to drive specific policy design. This requires the ability to predict fish catches - a challenging task. Theoretically, if the sea was completely filled with fish, it would be sufficient to know the trajectory of the fishing vessel and the capacity of its fishing gear to estimate the catch. However, fish - grouped in schools - are distributed sparsely and unevenly. There is no dataset that will give us the location of the fish schools in a given basin. Fishers' experience often tells them where the fish is likely to be at a given time, and therefore knowing a fishing ship trajectory in a given area is a highly relevant data resource for estimating the likely quantity of fish in that area. There are, however, other factors relevant to the availability of fish, such as the biological and atmospheric conditions of the environment in which the fish live. In particular, chlorophyll concentration and sea temperature seem to be important driving factors for fish availability [44, 51].

In our work we rely on three data sources: trajectories of fishing vessels obtained from AIS data, environmental data and fishing catch reports. We use such data to perform future catches prediction. While such data is unique today, we believe that as the fishing is being infused with information technology, such data will become commonplace, thus allowing for extensive applications of our analysis technique in the future.

Literature on the use of predictive methods to forecast fish catch in space and time is limited. Recent paper [46] is perhaps the closest to our work as it presents a correlative method to predict spatio-temporal presence of fish for small-scale fisheries using environmental and VMS (Vessel Monitoring System) data. One significant difference with our work is that [46] uses VMS instead of AIS. VMS data have some limitations, such as long time between the transmission of two consecutive signals (low temporal resolution), as well as the difficulty to obtain data. Moreover, catch reports are not available in [46], so spatio-temporal catch prediction is not considered.

To accomplish the first task, namely the creation of a spatio-temporal database, we start from two complementary data sources covering four years, from January 2015 to December 2018. The first data source is the set of terrestrial Automatic Identification System (AIS) data, i.e., the AIS data sent by ships and received by ground stations on the Italian coast of Northern Adriatic sea. In particular, we focus on the AIS data of the fishing vessels. The second data source is the fish catch reports of the Chioggia fish market, which is the primary market of the Northern Adriatic basin. Such reports contain the quantity and type of fish caught by all vessels selling their landings at the Chioggia fish market.

We present all the phases of the database creation. First, trajectories are reconstructed by linear interpolation of the raw AIS data: we clean the data and we detect the fishing vessels trips. As a second step the resulting trajectories are enriched with additional information concerning the activities and anomalies occurring during their trips. Finally, relying on the landing reports of the Chioggia fish market, we add a further valuable semantic aspect to the trajectories, annotating each trajectory segment of the fishing vessel with the quantity of fish caught in that part of the fishing trip. In order to distribute the total catches during a given trip along the segments of the trajectory of that trip, we define two different approaches, which are put into action and compared through specific analyses. We first consider an approach based on a *uniform distribution*, i.e., the catch of each given species is uniformly distributed along the fishing segments of the corresponding trajectory. Concretely, each fishing segment is associated with a portion of the total amount of fish, proportional to its length. The uniform distribution is clearly a simplification of reality. It is refined in a second approach which is based on a *weighted distribution*, whose underlying idea is that the areas where more vessels are fishing during a given time period are more likely to have higher catch rates, and thus catches are distributed in a way that privileges locations with a higher concentration of vessels.

We also provide a prototype implementation of our spatio-temporal data-base using MobilityDB [52], an open source geospatial trajectory data management and analysis platform, specifically developed to support the representation and the analysis of moving objects. On the one hand, the implementation in MobilityDB allows us to perform various analyses on the dataset and assess the appropriateness of the conceptual framework. On the other hand, it reveals the potentialities of MobilityDB for the reconstruction and management of trajectories enriched with semantic information.

We use the implemented spatio-temporal database to accomplish the second task, that is, gaining knowledge about the fishing activities in the Northern Adriatic sea. First, we check the AIS coverage to detect areas where there are transmission problems. Then, we map the fishing activities of some key species, highlighting all the interesting information and inferring new knowledge that will be useful for fishery management. The analyses show that spatializing the distribution of catches allows one to single out the fishing grounds and their seasonal and annual variation. This can be useful for explaining the fishers' behavior and better understanding the seasonal migration of the target species.

Finally, to address the third task, the spatio-temporal database and some relevant environmental data are used to explore a variety of prediction methods to forecast the so called Catch per Unit Effort (CPUE), an indicator intended to quantify the exploitation of fishing resources. Similar data have been used in [1] to develop early results on the use of machine learning techniques to predict the future CPUE from the past data. The work in [1], however, had some limitations, mainly related to the short temporal horizon – only two years, 2015 and 2016 – of the landing and AIS data. This, in fact, turned out to be a serious problem for the application of prediction methods: using the first year for training and the second one for testing, was not sufficient to assemble a robust model. The novel database that originates from the present work, thanks to the availability of the data sources for two additional years, significantly improves the results and helps in exploring the prediction of CPUE further.

In summary, the contributions of this work are the following:

- We examine and integrate two complementary data sources, i.e., terrestrial AIS data and fish catch reports data. We also incorporate environmental information. Thanks

to such data we reconstruct multiple aspect trajectories and we employ MobilityDB to create a spatio-temporal database of the Northern Adriatic Sea.

- We analyze the obtained database in order to extract useful information about AIS coverage and to detect the spatial distribution of target commercial species.
- We conduct an in-depth experimental analysis of a broad range of predictive models to predict CPUE and evaluate their performance using several measures. To the best of our knowledge, no work in the literature uses a combination of AIS, domain knowledge for fishing activity, fishing catch reports, and environmental variables to forecast CPUE.

The paper is organised as follows: Section 2 discusses some related works, Section 3 describes the trajectory reconstruction and enrichment and the creation of a spatio-temporal database by means of MobilityDB. Section 4 reports and illustrates the results of some specific analyses performed with MobilityDB on the obtained database. Section 5 describes the machine learning methods and the corresponding predictive model results. Finally, we draw some concluding remarks in Section 6. This is an extended version of the workshop paper [39].

## 2 Related work

In this section, we discuss some related work regarding (i) the integration of sea data with heterogeneous sources and the creation of semantic trajectories; and (ii) fishing activities forecast, which is the final goal of our predictive model.

### 2.1 Data fusion of sea data and semantic trajectories

Handling the fusion of ship movements with contextual and semantic information in the maritime domain is a recognized challenge [12, 35]. Several strategies were proposed to properly deal with the fusion of heterogeneous ocean data. For example, the papers [13, 43] show a platform in the maritime vessel traffic domain for discovering real-time traffic alerts by querying and reasoning across numerous streams (e.g., AIS, weather, ice, etc.). The authors use semantic web technologies to integrate heterogeneous data sources. In [8], the authors propose a model for the integration and analysis of data for vessel movement in a real-time maritime situation awareness system, also using semantic web techniques and tools. They introduce an ontology to model the maritime domain and to provide a common view on the different data sources. In particular, they define a movement ontology in which each position of a trajectory is modeled as being a move or a stop and they enrich trajectories also with information coming from the linked open data cloud, such as GeoNames<sup>1</sup> or DBPedia<sup>2</sup> or OpenStreetMap<sup>3</sup>. The queries are posed by using SPARQL and the ontology-based data access system *Ontop* and its extension *Ontop-spatial* are employed to map the relational data to the ontology and to translate queries to SQL queries. This approach has been proved worthwhile to detect routine traffic of vessels and abnormal vessel behaviour.

---

<sup>1</sup> <http://geonames.org>

<sup>2</sup> <http://dbpedia.org>

<sup>3</sup> <http://openstreetmap.org>

In [49] a Semantic Model of Ship Behavior (SMSB) is proposed to represent and reason on the meaning of the behaviors. Their steps include building a semantic network based on maritime traffic rules and detection methods to identify basic ship behaviors in various maritime scenes (e.g., dock, anchorage, traffic lane, etc), and a dynamic Bayesian network (DBN) to reason about potential ship behaviors. Their results show that basic behaviors and potential behaviors in all typical scenes of any harbor can be obtained accurately and expressed conveniently using SMSB. In [41] a framework named SPARTAN is presented. SPARTAN allows for real-time semantic integration of big mobility data with other data sources, aiming at providing enriched trajectories which are exploited by higher-level analysis tasks. The design and implementation of SPARTAN use well-known big data technologies (Apache Flink and Kafka), and their experimental evaluation shows the efficiency and scalability of the framework using maritime and aviation data.

All aforementioned papers use specific semantic models that are focused on the data integration component for detecting anomalies [8, 43], discovering spatio-temporal links between entities [41], or finding particular behaviors at harbours [49]. Other more general semantic models are *stops* and *moves* [33], *CONSTANT* [5] and *MASTER* [29]. They have been proposed as general approaches that can be applied to any application domain involving moving objects. Several concepts of our application cannot be modeled simply and directly as stops and moves patterns (e.g., ship entering or leaving a port to perform fishing activities or the association of environmental variables with a given pattern). The *CONSTANT* model is limited to a subset of aspects related to subtrajectories or the entire trajectory (e.g., activities performed by the object, the means of transportation, the visited POIs, the trip's goal, and some behavior-specific patterns).

We have chosen the *MASTER* model since it is more flexible and expressive and allows for the representation of heterogeneous features, ranging from simple labels to complex objects. In particular, it introduces the novel concept of *aspect* which consists of “a real-world fact that is relevant for the trajectory data analysis” [29]. Different kinds of aspects are modeled: (i) *volatile* aspects, usually associated with the trajectory points, since they vary during the object movement; (ii) *long-term* aspects, which do not change during an entire trajectory, and hence they are associated with the whole trajectory; (iii) *permanent* aspects holding during the whole life of an object, thus they are connected to the moving object and not to the trajectory. Based on this notion, a *multiple aspect trajectory* is defined as a sequence of spatio-temporal points of a moving object with a (possibly empty) set of long-term aspects. Each point can have a set of volatile aspects and the moving object can be related to a set of permanent aspects. It is important to highlight that our modeling is an instance of *MASTER*, so we followed its recommendations in the design and created a concrete implementation of it for a particular goal: integrating AIS, fish catch reports and environmental variables to represent, analyze and predict the fishing activities in the Northern Adriatic basin.

## 2.2 Fishing activities forecast

The literature on fishing activities forecast is broad and can be decomposed in several ways. From a fishing management view, works like [32] propose a seasonal forecast system that combines environmental and fish habitat data (e.g., collected by fish tagging) to predict tuna distribution. The authors in [31] integrate satellite data and statistical models output to examine the relationship between sea surface temperature and chlorophyll-a concentration. They also define simple methods to forecast potential fishing grounds. The work of [14] tries to forecast 1-month

catches considering only the anchovy catches in past months as inputs. In [25, 26] the authors use complementary data such as landings, auction prices, regulatory data and AIS, to assess the spatio-temporal distribution and intensity of fishing activity. The focus is on mapping dredge gear fishing grounds using fishing intensity estimates based on AIS data. The fishing/not fishing activity is inferred using the vessels speed. Similarly to [32] and [31], we use environmental data (e.g., chlorophyll-a and sea surface temperature) to predict fish distributions. Also, similarly to [14] we use fish catch information to predict future catches. The objectives of the works [25, 26] are related to ours but are not the same. The predicted variable in [25, 26] is different since they try to assess the spatio-temporal distribution and intensity of fishing activity while our work is focused on predicting Catch Per Unit Effort (CPUE). Besides, the dataset used in our work is several orders of magnitude larger, we consider several environmental variables, and we explore a wide range of machine learning models to forecast CPUE. Unlike all of them, we are the first to use wave height as an environmental variable in our model.

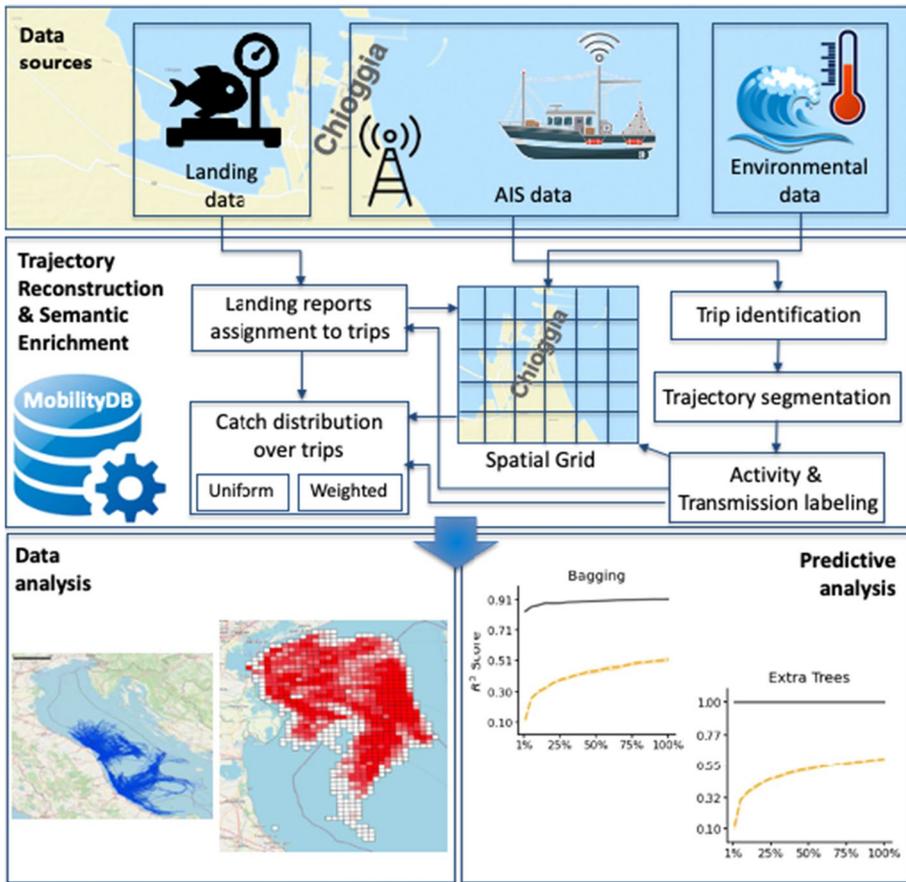
From a viewpoint that considers the geolocation technology used to track ships, some works use Vessel Monitoring System (VMS) [27], satellite images [31] or AIS [15, 17, 45, 48]. Most of these works focus on training models to forecast when a vessel performs a fishing activity. Different types of fishing ships (e.g., long-liners, purse-seiners, etc.) have different movement patterns. Predicting these patterns depends on the training data given to the machine learning model [45], or the domain specialist's ability to create rules that reflect these patterns [30]. In this work we use domain knowledge from specialists to determine the activity of vessels (e.g., fishing or not) on their trajectory segments. Based on the knowledge of ranges of fishing speed for different types of fishing gears (e.g., trawlers, long-liners, etc.), we encode the specific rules to detect vessel activities. By exploiting this information, we can compute in a very accurate way the area swept by vessels while fishing, thus allowing for a more realistic estimate of fishing effort and CPUE.

From the viewpoint of the analysis of models for time-series forecasting, many studies used the Autoregressive Integrated Moving Average (ARIMA), e.g., [4, 22, 28, 38, 50], and the Seasonal version (SARIMA), e.g., [3, 37] to forecast fish landings but without considering the spatial distribution of the resources. Authors in [44] integrated chlorophyll concentration, derived from remote sensing satellite, and sea surface temperature images to generate a fishery forecast. Recently, authors in [51] applied a model technique for optimal fishing, by using fishing location data, chlorophyll-a and sea surface temperature, to forecast the spatio-temporal distribution of the Indian mackerel. Also, in [46] a correlative modelling approach, combining VMS and environmental variables, was used to identify potential fishing grounds of small-scale fishery. Finally, authors in [16] applied statistical and process-based models to predict the changes in fish abundance and distribution correlated to climate change.

Summing up, the works discussed in this section may have used similar techniques, but most were applied for a different goal or used fewer environmental variables when compared to our work. To the best of our knowledge, no work in the literature uses a combination of AIS, domain knowledge for fishing activity, fishing catch reports, and environmental variables to forecast CPUE. Besides, this work tests a wide range of machine learning models on a larger scale (i.e., over 4 years).

### 3 Multiple aspect trajectories

In this section we illustrate the steps we followed to produce a spatio-temporal database of fishing vessels trajectories in the Northern Adriatic sea, enriched with landing data from the Chioggia market. We start by describing the data sources (Section 3.1) of

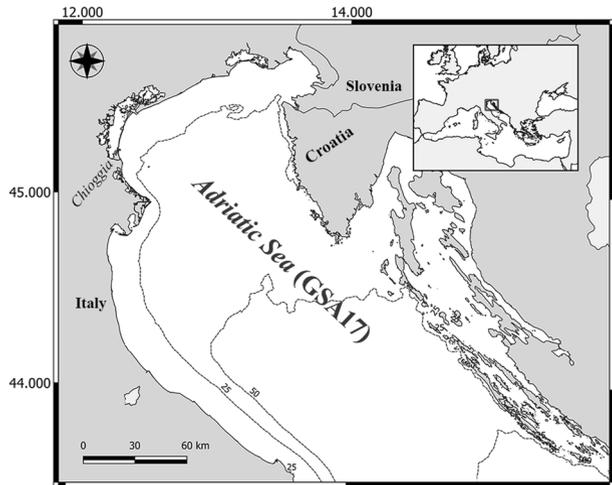


**Fig. 1** Overview of the process consisting of four main steps: data sources, reconstruction and enrichment of trajectories, data analysis and machine learning-based estimation

our case study, that is, the terrestrial AIS data of the Northern Adriatic sea, the landing reports of the Chioggia fish market, and the environmental data. Next, we explain how trajectories can be reconstructed by linear interpolation of the raw AIS data (Section 3.2). In this step, we clean the data, we detect the trips performed by the fishing vessels and we enrich the resulting trajectories with additional information concerning the activities occurring during the trips. Then we illustrate how to assign landing reports to trajectories (Section 3.3) and we formalise the two different techniques to distribute the fish catches along the trajectories and we define how the fishing effort and the catch per unit effort (CPUE) are computed (Section 3.4).

The overall view of the process is depicted in Fig. 1: starting from the raw terrestrial AIS data of the fishing vessels and from the landing reports (amounts of fish by species, caught by a given ship in a single trip) of the Chioggia’s market, we build up on top of MobilityDB a spatio-temporal database of multiple aspect trajectories that enables us to perform analyses on the spatio-temporal and semantic features of the trajectories. Moreover, we use environmental information, i.e., sea temperature, chlorophyll-a, and wave

**Fig. 2** Map of the North Adriatic Sea



height, to enrich the spatial grid we employ to partition the study area and these data are given as input to the predictive model presented in Section 5.

### 3.1 Data sources

#### 3.1.1 Automatic identification system (AIS)

AIS raw data, provided by the Italian Coast Guard, were obtained for the trawl fishing vessels operating in the Northern Adriatic Sea from January 2015 until December 2018. The covered area is about 1.5 million km<sup>2</sup>. Figure 2 shows a map of the North Adriatic Sea.

A total of 70 (2015), 77 (2016), 82 (2017) and 81 (2018) trawlers, with a length overall above 15m and operating from the Port of Chioggia (45.219643; 12.278885), were taken into consideration in this study: in particular, small and large bottom otter trawl (SOTB and LOTB), Rapido, one specific kind of beam trawl (RAP), and midwater pair trawl (PTM). The identification of the vessels was performed by matching the data present in the AIS (MMSI code, vessel name and the call sign) with those of the European Fleet Register, which supplies specific information on the vessels (i.e., primary and secondary gear, length overall, gross tonnage, etc.). All the data given by the AIS (i.e., data position, speed, time, MMSI) were used to identify the fishing tracks and analyze the fishing activities (fishing, not fishing).

#### 3.1.2 Daily landing reports

Landing dataset was obtained from the Chioggia's Fish Market, whose harbor hosts one of the main fishery fleets of the Adriatic Sea. This dataset consists of daily landings (catch amounts in kilogram) for 104 commercial species caught during four years, from January 2015 to December 2018 in the Northern Adriatic Sea. The records pertain around 80 fishing vessels, and contains a total of 278078 transactions over the four years, as detailed in Table 1.

**Table 1** Dimension of the landing report dataset over a time series of 4 years

Year	No. of vessels	No. of transactions
2015	71	64,180
2016	79	70,017
2017	80	71,716
2018	76	72,165

### 3.1.3 Environmental data

We also incorporated daily measures of sea temperature (in kelvin), chlorophyll-a (in  $mg/m^3$ ), and wave height (in meters) over the same four years considered. Data have been taken from Copernicus<sup>4</sup>.

The sea surface temperature and the chlorophyll-a influence the species distribution, while the wave height affects the fishers behavior. Hence, adding such semantic information could be relevant for a more accurate prediction of the CPUE indicator. Moreover, the utilization of the sea surface temperature can be helpful to evaluate the effect of climate changes on fishing activities, a hot topic to be considered.

### 3.2 Trajectory reconstruction and enrichment

Since boat positions are recorded every 10–20 seconds, that correspond to a small spatial displacement of the boat, trajectories are reconstructed by linear interpolation of the raw AIS data. While performing the reconstruction raw data are cleaned: all the points implying movements that are not physically feasible due to a maximum possible boat speed are removed. In case positions are recorded less frequently, it is possible to use other interpolation techniques following an approach similar to the one adopted by the authors of [24] to deal with sparse AIS data (Lagrange interpolation) or other state-of-the-art interpolation methods like the one described in [20]. Next, in order to organize the data into distinct trajectories followed by the fishing boats, we apply two criteria: a new trip begins *a*) when the vessel is inside a port area and there is no transmission for longer than a fixed time, or *b*) there is an AIS datum outside a port area and the immediate previous AIS datum is inside a port area and the time period between the two AIS data is greater than 20 minutes. The first condition corresponds to the fact that the vessel ends a trip, it switches off the AIS, it is docked at the port and after a while it starts a new trip. The second one corresponds to a situation in which a vessel leaves out of the port and then it starts transmitting when it is outside the port (20 minutes is the minimum time a vessel takes to leave the port). A detailed analysis reveals that some fishing vessels, after entering the port area at the end of a trip, continue to transmit their position. In this way, none of the above criteria is met. This causes a wrong trip reconstruction in which two or more trips are considered as a unique trip with a duration of several days. Hence, to avoid this phenomena we remove the AIS data transmitted inside the port when the vessel returns to a port. In Table 2 we report the dimension of the original AIS datasets and the resulting number of trajectories.

Once the reconstruction is carried out, the trajectory is a sequence of segments obtained by connecting consecutive AIS points. Each trajectory contains the following information: MMSI or boat identifier, trip duration (in hours), trip length (in meters), total time of fishing activity (in hours), total length of the fishing activity (in meters), date and time

<sup>4</sup> <https://www.copernicus.eu/en>

**Table 2** Raw AIS data vs trajectories

Year	No. of vessels	AIS data	No. of trajectories
2015	70	29,757,601	11,280
2016	77	38,519,864	11,130
2017	82	21,247,207	35,335
2018	81	25,098,120	9,549

of the trip departure and conclusion, total number of segments with more than 30 minutes between two consecutive AIS transmissions, and the attribute *anomaly*, a code specifying whether the trip presents an anomaly or not and the kind of anomaly.

The last attribute highlights some strange behaviour of the fishing vessel. Possible anomalies are: the time interval between two consecutive AIS data is longer than 30 minutes outside the port, suggesting some points could be missing (*anomaly* is set to 1); a boat remains inside a port area for the whole trip (*anomaly* is set to 2); the duration of the trip exceeds the 24 hours (*anomaly* is set to 3). If none of the above cases occurs, the trip is considered as normal and *anomaly* is set to 0.

It is worth noting that through the MMSI, we can obtain further information on the vessel, such as its name, the fishing gear, the length overall. Each segment in the trajectory is in turn annotated with: speed, position of the segment with respect to the port areas, activity of the boat within the segment, length of the segment, time spent in the segment and transmission.

The *activity* attribute describes what the vessel is doing: 0 the vessel is in the port, 1 means exiting from the port, 2 is about entering to port, 3 is about fishing and 4 corresponds to navigation. The *in port*, *exiting from port* and *entering to port* situations can be deduced from the position of the extremes of the segment w.r.t. the port area. If none of the previous cases applies, the fishing or navigation activities are established on the basis of the average speed of the boat. More precisely, if the average speed is in the range of the fishing speed of the gear the boat is equipped with, the boat is assumed to be in a *fishing* phase; otherwise, it is assumed to be in a *navigation* phase. The considered gears and their minimum and maximum speed during the fishing activity are reported in Table 3.

The attribute *transmission* records whether the end points of the segment have a time distance greater than 30 minutes. If this happens the attribute is set to 1, otherwise to 0. As explained above, the presence of segments with transmission set to 1 allows for the detection of an anomalous behaviour of the trajectory.

These trajectories are modeled as a *multiple aspect trajectory*, following MASTER model [29]. Actually, as minimum granularity to attach semantic information, we do not consider a single spatio-temporal point as in the original MASTER model, but segments. This is motivated by the fact that we want to highlight the presence of homogeneous trajectory portions, which are the appropriate granularity level for our analyses. According to the MASTER model classification, the information listed above can be classified as *long-term aspects*, (those associated with the full trajectory), *volatile aspects* (those associated with the segments) and *permanent aspects* (those associated with the fishing vessel, derived from the MMSI).

**Table 3** Gears and their minimum and maximum fishing speed (in km/h)

ID	Gear description	Min speed	Max speed
SOTB	Small bottom otter trawl	3.704	8.334
LOTB	Large bottom otter trawl	3.704	8.334
PTM	Pelagic pair trawl	3.704	10.186
RAP	Rapido	7.408	12.964

By using the MASTER model we are able to represent different aspects of our trajectories in a uniform and simple way. Moreover, this representation allows us to perform complex queries merging spatial, temporal and semantic features. In the rest of the paper, we denote by  $T$  the resulting set of multiple aspect trajectories.

### 3.3 Catch distribution

We next describe how to merge the trajectories of the fishing vessels with the daily landing reports provided by the Chioggia fish market. The latter dataset contains information about each trading transaction, including the landing date, MMSI of the seller, the species, and the quantity of fish. Note that we work on a subset of the set of reconstructed multiple aspect trajectories. In fact, we exclude from our analysis, fishing vessels that do not sell their fish in Chioggia, trajectories that do not leave the port area, and trajectories that do not have any fishing activity. In order to perform the merge we need to associate each fish market transaction with a trajectory of the vessel having the specified MMSI. To accomplish this task, for each transaction, we select the vessel trip with the most recent arrival in the port (before 4 PM of the landing date). Arrivals after 4 PM are associated with transactions occurring the next day. The quantity (weight) of fish assigned to a trajectory is called a *catch*.

To distribute the fish associated with a trajectory over the trajectory's fishing segments we follow two different approaches: (1) uniform distribution, and (2) weighted distribution.

In the first case, the catch is uniformly distributed along the fishing segments of the corresponding trajectory. Each fishing segment of the trajectory is associated with a fraction of the total amount of fish, proportional to its length. We consider separately each species that the fishing vessel caught.

**Definition 1 (Uniform distribution)** Let  $tr$  be a trajectory and let  $catch$  the record containing the quantities of the different species associated with the trajectory  $tr$ . Given a segment  $s$  belonging to  $tr$  with activity set as *fishing* and a species  $sp$ , the *uniform catch* for segment  $s$  and species  $sp$  is defined as

$$d_U(s, sp) = \frac{s.len}{tr.len\_fishing} * catch.sp \quad (1)$$

where  $tr.len\_fishing$  is the attribute storing the total length of the fishing activity for the trajectory  $tr$ ;  $s.len$  is the length of the segment;  $catch.sp$  selects the quantity of a certain species  $sp$ .

Clearly the assumption of uniform catch distribution is a simplification of reality. We consider also a refinement based on a so called *weighted distribution*. The idea is that the areas where more vessels are fishing, during a given time period, are more likely to have higher catch rates. In order to implement this technique, we need to suitably partition the fishing area of interest because it becomes crucial to evaluate the number of fishing vessels present in a certain zone. We will use a square grid whose size will be influenced by two elements. On the one hand, it must take into account the dimension of the fishing vessels and their behaviour during the fishing activity, a knowledge provided by the environmental scientists. On the other hand, it will depend on the kind of analysis to be performed: generally speaking it should be large enough to include an amount of data adequate for the analysis.

The introduction of the grid leads to a further segmentation of the trajectories. In fact, each segment that spatially crosses one or more cells of the grid needs to be split into smaller segments in such a way that each portion is completely inside a single cell. Moreover, since we deal with a spatio-temporal grid, all segments spanning over two days are split into two smaller segments by taking as extra point the interpolated position at midnight.

In order to compute the weighted distribution, we associate a coefficient with each spatio-temporal cell of the grid.

**Definition 2 (Fishing Coefficient)** Let  $c$  be a spatio-temporal cell and  $sp$  a species. The *fishing coefficient* of cell  $c$  for the species  $sp$  is defined as follows:

$$\alpha(c, sp) = |\{tr \in T \downarrow sp \mid tr \cap c \neq \emptyset\}| * \sum_{tr \in T \downarrow sp} \sum_{s \in tr \cap c, s.activity=fishing} s.len \quad (2)$$

where  $T \downarrow sp$  is the set of trajectories having a landing report with the species  $sp$ ;  $tr \cap c$  returns the intersection between the trajectory  $tr$  and the cell  $c$ ;  $s.activity$  and  $s.len$  are respectively the attributes of segment  $s$  storing the activity and the length of the segment.

The coefficient  $\alpha(c, sp)$  combines the number of fishing vessels and the amount of fishing activity they perform in the cell, hence it provides a measure of the fishing activity in the cell. Note that the coefficient depends on the species. Hence, for each species  $sp$ , we select only the trajectories having a landing report for the given species  $sp$ .

Since it is natural to expect that vessels will mostly concentrate in fishy areas, the intuition is that cells where the fishing coefficient is higher will have higher catch rates. This leads to the idea, formalised below, of using such coefficient as a weight when distributing catches over a trajectory.

**Definition 3 (Weighted distribution)** Let  $tr$  be a trajectory and let *catch* the record containing the quantities of the different species associated with the trajectory  $tr$ . Given a segment  $s$  belonging to  $tr$  with activity set as *fishing* and a species  $sp$ , the *weighted catch* for segment  $s$  and species  $sp$  is defined as

$$d_w(s, sp) = \frac{\alpha(s.cell, sp) * s.len}{\sum_{s' \in tr \wedge s'.activity=fishing} (\alpha(s'.cell, sp) * s'.len)} * catch.sp \quad (3)$$

where  $s.cell$  is the unique cell the segment  $s$  belongs to.

When distributing the catch over the segments of the trajectory  $tr$ , again only segments which are classified as fishing are considered. The difference is that in this case each segment  $s$  receives a weight which is proportional not only to the length  $s.len$  of the segment but also to the fishing coefficient  $\alpha(s.cell, sp)$  of the cell the segment belongs to.

### 3.4 Computation of the fishing effort over the grid

After the creation of the multiple aspect trajectories, we proceed with the computation of the *fishing effort*, an essential indicator for monitoring the fishing pressure on an area of interest over time. As mentioned above, we partition the Northern Adriatic Sea into a

regular grid. The fishing effort for a spatio-temporal cell is defined as the ratio between the area of the cell “swept” by vessels while fishing during the associated time period and the total area of the cell itself. The swept area depends on the employed gear which can be recovered from a specific dataset where each vessel, identified by its MMSI, is associated with its gear.

In the following we will denote by  $c$  a generic spatio-temporal cell in the area and period of interest, and by  $g$  a gear (small and large bottom otter trawl, Rapido and mid-water pair trawl).

**Definition 4** Let  $c$  be a spatio-temporal cell and  $g$  a gear. The *fishing effort* wrt the gear  $g$  in the cell  $c$  is defined as follows:

$$fe(c, g) = \frac{(\sum_{tr \in T, gear(tr)=g} len(tr \cap c)) * gear\_width(g)}{area(c)} \tag{4}$$

where  $T$  is the set of multiple aspect trajectories;  $len(tr \cap c)$  returns the sum of the lengths of the fishing segments of trajectory  $tr$  falling in the spatio-temporal cell  $c$ ;  $gear\_width(g)$  is the width of the net of gear  $g$ ;  $area(c)$  is the total area of the spatial component of the cell  $c$ .

It is worth noting that we can obtain the total fishing effort in a spatio-temporal cell  $c$  by summing up the fishing effort for each gear.

Thanks to the reconstruction and the semantic enrichment of trajectories we can compute the lengths of the fishing segments falling in each cell. This allows a more accurate and realistic estimate of the swept area and therefore of the fishing effort.

### 3.4.1 Catch per unit effort (CPUE)

Catch per unit effort (CPUE) is an indicator of the species abundance in the assessment of fishery resources. This index represents a valid method to evaluate the population trends where, a decrease of CPUE indicates a situation of over-exploitation, a steady CPUE value points out sustainable exploitation of the fishery resources, and an increase of its value corresponds to a healthy and growing population.

In order to compute this indicator, we need the quantity of fish caught in each spatio-temporal cell by boats having a particular gear  $g$ .

**Definition 5** Let  $c$  be a spatio-temporal cell and  $g$  a gear, the fish *catch* wrt to the gear  $g$  in cell  $c$  is defined as follows:

$$catch(c, g) = \sum_{tr \in T, gear(tr)=g} quantity(tr, c) \tag{5}$$

where  $T$  is the set of multiple aspect trajectories;  $quantity(tr, c)$  returns the sum of the fish quantities in kilograms associated with the fishing segments of trajectory  $tr$  falling in the spatio-temporal cell  $c$ .

Note that the function  $quantity$  can be computed by using either the uniform or the weighted distributions, and this will produce different values for  $catch$  that we denote by



**Fig. 3** Trajectory visualisation as a sequence of spatio-temporal points (left), as a *continuous* function (center), and as a semantic object where the *activity* attribute is highlighted (right)

$catch_U(c, g)$  and  $catch_W(c, g)$ , respectively. Correspondingly, we obtain the *uniform* CPUE index and the *weighted* CPUE index defined as follows.

**Definition 6** Let  $c$  be a spatio-temporal cell and  $g$  a gear, the *catch-per-unit-effort* (CPUE) wrt to the gear  $g$  in cell  $c$  is defined as follows:

$$cpue_X(c, g) = \frac{catch_X(c, g)}{fe(c, g)} \quad \text{for } X \in \{U, W\} \quad (6)$$

CPUE is, therefore, a key indicator for fisheries management since it gives information on the sustainability of the fishing activities in the area of interest. As a consequence, an accurate forecast of CPUE could help decision makers to maintain a sustainable fishing business by adapting the fisheries management plans based on its forecasted values.

### 3.5 Implementation

To construct and store the set of multiple aspect trajectories, we used MobilityDB [52], an open source extension to the PostgreSQL database system<sup>5</sup> and its spatial extension PostGIS<sup>6</sup>. It provides temporal types and spatio-temporal operators that ease the management of moving objects.

One main feature of MobilityDB is that it offers a construct for representing the evolution of a value during a sequence of time instants. The values between successive instants are interpolated using a linear function. Clearly, this construct perfectly suits the representation of trajectories, which are reconstructed from a sequence of spatio-temporal data. In our case, the spatio-temporal points are the AIS data aggregated on the basis of the trajectory *id*. We created a set of objects of type `tgpoint`, which is a temporal type modelling a point changing its position along a time period.

Next, the function `trajectory` is applied to these objects, and a `geometry` value is returned. In this way the trajectory can be visualized. In our work, for visualizing trajectories and the result of our analyses, we used QGIS<sup>7</sup>, an Open Source GIS that supports viewing, editing, and analysis of geospatial data. For instance, Fig. 3(left) shows the sequence of AIS data, i.e., the sequence of spatio-temporal points, related to the trip of a

<sup>5</sup> <https://www.postgresql.org/>

<sup>6</sup> <https://postgis.net/>

<sup>7</sup> <https://qgis.org/en/site/>

fishing boat, whereas Fig. 3(center) illustrates the *continuous* representation of the same trip obtained by using the MobilityDB construct. The linear interpolation is internally implemented by the system, with the dual advantage of raising the user from this task and simplifying queries and analyses. Note that using a different interpolation technique is possible but it would require an explicit implementation.

MobilityDB provides a lot of spatio-temporal operators to handle trajectories. For instance, `startTimestamp` and `endTimestamp` return respectively the first and last time instant among a set of time instants and this can be useful to extract the beginning and ending points of a trajectory; `getValue` returns a value at a particular time instant. There are operators to check topological relations between trajectories, like `tintersects`, `t disjoint`, and others to compute distances. Interestingly the results of these operators are values changing in time. In fact, it can happen that at certain time periods trajectories enjoy the relations whereas at other ones they do not, and the distance between the objects can vary depending on the movement of the objects themselves. For instance, the user can check whether a fishing vessel respects the rule that it can fish only at a distance greater than three nautical miles from the coast and eventually detect where and when the ban has not been observed.

MobilityDB allows for an easy representation of multiple aspect trajectories where semantic attributes can be modelled as temporal types. This means that we can model in a single table both the sequence of spatio-temporal points forming a trajectory and information associated with the whole trajectory itself, such as the duration and length of the trajectory. Moreover, a trajectory can be segmented and each segment can be stored as a temporal type. Even in this case we can add other attributes modelling features of the segment itself, such as the speed, the activity, the transmission and the quantity of caught fish. In Fig. 3(right) the different colours describe the activities of the fishing vessel. They allow the user to immediately detect where the vessel is fishing and also the shape of the movement. For instance, the figure highlights several circular movements and the experts have confirmed that they are typical of this kind of fishing activity.

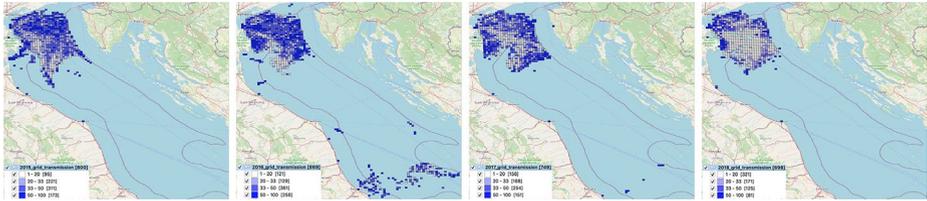
Finally, MobilityDB provides support for the GiST (Generalized Search Tree) and SP-GiST (Space-Partition GiST) indexes, which can be created for table columns of temporal types. We used such indexes for accelerating spatial, temporal and spatio-temporal queries.

## 4 Exploratory data analysis

This section presents some analyses performed with MobilityDB on the obtained spatio-temporal database of the Northern Adriatic sea. For these analyses, a suitable cell size for the spatial grid resulted in being 3x3 km.

The first analysis aims at visualizing the regions where there are transmission problems. We exploit the *anomaly* attribute, and in particular, we investigate trajectories having this attribute set to 1. In Fig. 4 we show for each cell the percentage of trajectories that got disconnected from the AIS for a time period greater than 30 minutes while crossing that cell with respect to the total number of trajectories passing through the cell.

Looking at Fig. 4, it is evident that the no-transmission anomaly has decreased a lot from 2015 to 2018. In fact, in 2015 the area where this percentage is over 50% is extensive, and it covers almost the whole fishing zone. Instead, in 2018 this phenomenon is localized



**Fig. 4** Spatial distribution of the no-transmission anomaly, years 2015, 2016, 2017 and 2018, respectively

in few areas, i.e., close to the coasts and along the borders of the territorial waters. Moreover, in 2018 there are also some isolated cells in the southern part.

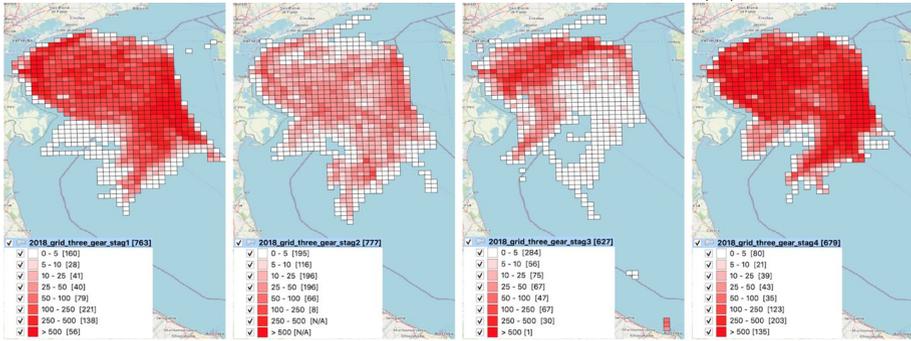
The low spatial coverage of AIS is a well-known issue, and the amount of missing data can vary substantially between vessels as discussed in [42]. Our analysis reveals that data from 2018 are more reliable and can be helpful in detecting areas where the AIS signal is not received well, like the isolated cells in the southern portion of the sea area under investigation.

This analysis is an example of how the semantic knowledge hidden in a single attribute, such as the *anomaly* attribute, can be useful to improve the general spatio-temporal knowledge of the domain of interest significantly. On the one hand, the progressive low-coverage reduction of AIS data is *per se* a piece of highly valuable information for ecologists and policymakers since this ensures the reliability of the collected data. On the other hand, the proposed implementation allows the experts to continuously monitor the degree of coverage and eventually decide to add further terrestrial AIS receivers.

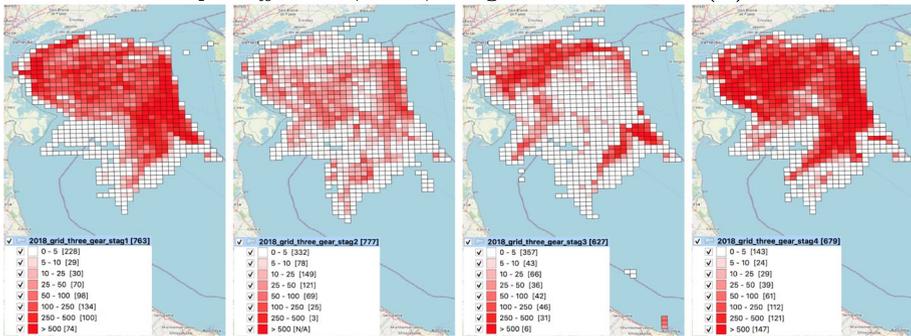
The second and third analyses take advantage of the catches distribution and infer some knowledge on key species in the area. In fact, spatializing the distribution of catches has several important applications. For instance, it allows us to obtain knowledge about the seasonal variation of the fishing grounds. This, in turn, is useful for explaining the fisher behavior and better understanding the seasonal migration of a target species. Figure 5 reports the seasonal spatial distribution of cuttlefish, *Sepia officinalis*, aggregated by fishing gears (SOTB, LOTB and RAP) in 2018. Cuttlefish is one of the main target species of the Adriatic Sea; hence it is an ideal case study for showing seasonal migratory behavior. It is worth noting that the most productive seasons were autumn and winter, with two high-density areas, one nearer the coast and the other one more offshore, at the border with the Croatian waters. In spring, the catches resulted more scattered, while in summer, the catch area was more defined and localized closer to the Italian coast. This is in line with the general ecological knowledge about the behavior of the species; hence, the catches data correctly reflect cuttlefish seasonal spatial distribution behavior. Figure 5 also reports the comparison between the uniform (A) and the weighted (B) distribution maps of cuttlefish *Sepia officinalis* in 2018. It is evident that the maps obtained with the weighted distribution (B) are more defined, allowing the identification of the fishing grounds of cuttlefish better.

Another important application of the spatial distribution of catches is detecting different fishing grounds over the years. As an example, Fig. 6 shows the spatial distribution of anchovies catches in fishing grounds recorded in winter 2015, 2016, 2017, and 2018 and distributed according to the weighted distribution of the catch. The maps clearly show how the fishing grounds and, consequently, the distribution of anchovies changed over the years. In particular, a gradual reduction of the fishing grounds is observed from 2016 to 2018. This is clearly a piece of relevant information for both ecologists and policymakers:

*Sepia officinalis*, 2018, Uniform distribution (A)



*Sepia officinalis*, 2018, Weighted distribution (B)



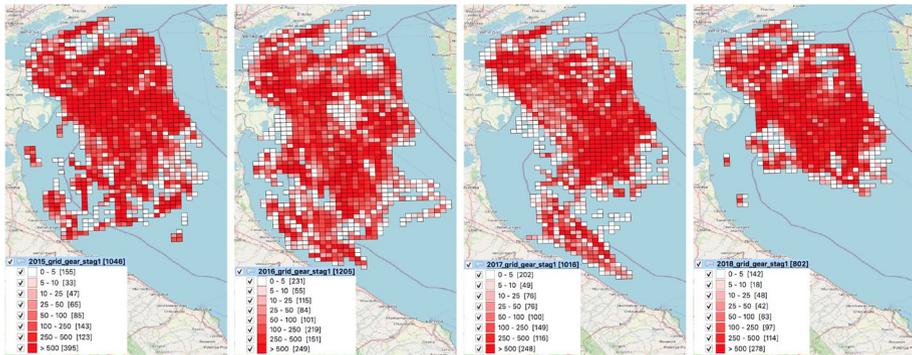
**Fig. 5** Comparison between uniform (A) and weighted (B) distribution of cuttlefish *Sepia officinalis*, aggregated by seasons (winter, spring, summer and autumn 2018)

if the fishing ground reduction results from overexploitation of the species, they can adopt appropriate countermeasures.

Finally, we would like to point out that these are only a few examples of the analyses that can be performed using the dataset of multiple aspect trajectories. For instance, we can focus on vessels equipped with specific fishing gear (i.e., LOTB, SOTB, RAP, and PTM) and determine their fishing grounds and the corresponding degree of exploitation. This fine-grained analysis could help to reveal different efficiency degrees of fisheries that, in turn, could constitute a basis to implement specific management actions for these activities. Moreover, we can vary our analysis according to different periods and consider only certain sea areas. For instance, one could focus on protected areas, like the Pomo Pit or the Sole Sanctuary. We can also select the behavior of single trajectories satisfying complex conditions concerning both their movements and their semantic annotations by using the operators available in MobilityDB.

### 5 Experimental results with machine learning models

This section presents the experiments using machine learning models for regression or, simply, regressors. Our objective is to estimate the CPUE, the catch-per-unit effort introduced in Section 3.4.1, and evaluate the data fitting of the models using learning curves. Recall that we need to fix a spatial grid, which, for these analyses, consists of cells of size 5x5 km.



**Fig. 6** Spatial distribution of anchovy *Engraulis encrasicolus* in winter, years 2015, 2016, 2017 and 2018, respectively

We run our experiments with ten well-known machine learning regressors, as implemented in the *Scikit learn* library [9]: Extra Trees [18], Random Forests [7] XGBoost [11], Bagging [6], Cat Boost [36], KNeighbors [2], LGBM [21], HistGBoost [34], Adaptive RF [19], MLP [40]. Among them, Extra Trees [18] and Random Forest (RF) [7] have gained considerable attention because of their optimal regression performance. Extra Trees or Extremely Randomized Trees is a robust ensemble learning algorithm. In particular, it is an ensemble of decision trees similar to Random Forest. The Extra Trees algorithm creates a large number of decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees. This method uses the whole original sample, unlike RF. RF uses bootstrap replicas. On the other hand, for selecting cut points, RF selects the optimum split while Extra Trees selects a split point at random.

## 5.1 Evaluation metrics

To compare the performance of the regressor models, we use the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination ( $R^2$ ). The MAE, RMSE range between 0 and 1, where 0 is the best case. However, the  $R^2$  is defined between 1 and -1, where 1 is the best-case scenario and any value lower than zero points out arbitrarily worse results. These three metrics are formally defined as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - \hat{T}_i| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \hat{T}_i)^2} \quad (8)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - \hat{T}_i)^2}{\sum_{i=1}^n (\bar{Y}_i - \hat{T}_i)^2} \tag{9}$$

where  $\hat{Y}$  and  $\bar{Y}$  are the predicting values and the average of observed values respectively, while  $\hat{T}$  stands for the ground truth.

### 5.2 Attribute description and data splitting

The season information was converted into numeric values by using one-hot-encoding of the attribute regarding the seasons of each year. Hot-encoding essentially transforms categorical into numeric attributes. The list of attributes used for modeling the fishing prediction is described in Table 4. They are divided into three categories: environmental, spatial, and temporal.

Concerning the work of Adibi et al. [1], we have significantly extended the steps of data preparation and preprocessing, including the tasks for removing outliers, handling missing values, and data standardization by scaling the attributes of the dataset. The standardization was needed to set all features in the same range and transform data into a Gaussian-like continuous distribution. That was carried out by using the approximation of Kolgomorov-Smirnov. For the target value estimation, we calculated the uniform and weighted CPUE indexes according to the definition provided in Section 3.4.1. Table 5 shows the results for the top ten regressors considering the selected metrics, i.e., MAE, RMSE and  $R^2$ . The data split used here was a standard 5-fold cross-validation (i.e., dataset is split into 5 folds, and each fold is used as a testing set eventually in the validation procedure) with the error of the five estimates averaged as the final performance measurement.

The use of machine learning models usually involves careful tuning of learning parameters and model hyperparameters. The strategy used in this article for parameter tuning was grid-search, varying uniformly the number of parameters for each machine learning method using the standard intervals of the Scikit learn package. However, after selecting the interval of parameters for each model, we also used a Gaussian Process to speed up the optimization process. The Neural Network architecture was the Multi-layer Perceptron (MLP) with 10-100 layers, with activation function ReLu and optimizer Adam and the learning rate was not fixed.

**Table 4** List of attributes used for modeling the fishing prediction

	Attribute description	Symbol	Type	Unit
Environmental	daily chlorophyll-a concentration	chl	float	<i>mg/m<sup>3</sup></i>
	daily sea surface temperature	sst	float	<i>kelvin</i>
	daily spectral significant wave height	vhm0	float	<i>meter</i>
Spatial	latitude of grid cell centre	lat	float	<i>degree</i>
	longitude of grid cell centre	lon	float	<i>degree</i>
Temporal	day of year (1-365)	doy	int	
	month of year (1-12)	moy	int	
	week of year (1-53)	woy	int	
	season (1-4)	season	int	

**Table 5** Results of the regressors using three evaluation metrics - MAE, RMSE and  $R^2$  score - over the dataset prepared for the experiments

Regressor	Uniform CPUE			Regressor	Weighted CPUE		
	MAE	RMSE	$R^2$		MAE	RMSE	$R^2$
Extra Tree [18]	0.34	0.59	0.65	Extra Tree [18]	0.41	0.62	0.60
Random Forest [7]	0.37	0.62	0.61	Random Forest [7]	0.44	0.66	0.56
Bagging [6]	0.40	0.65	0.57	Bagging [6]	0.47	0.69	0.51
Cat Boost [36]	0.45	0.69	0.52	Cat Boost [36]	0.50	0.71	0.49
XGBoost [11]	0.45	0.69	0.52	XGBoost [11]	0.50	0.72	0.48
kNeighbours [2]	0.44	0.71	0.49	LGBM [21]	0.52	0.74	0.44
LGBM [21]	0.47	0.72	0.47	kNeighbours [2]	0.50	0.74	0.44
HistGBoost [34]	0.47	0.73	0.47	HistGBoost [34]	0.52	0.74	0.44
Adaptive RF [19]	0.49	0.77	0.40	MLP [40]	0.55	0.78	0.38
MLP [40]	0.50	0.77	0.40	Adaptive RF [19]	0.54	0.78	0.38

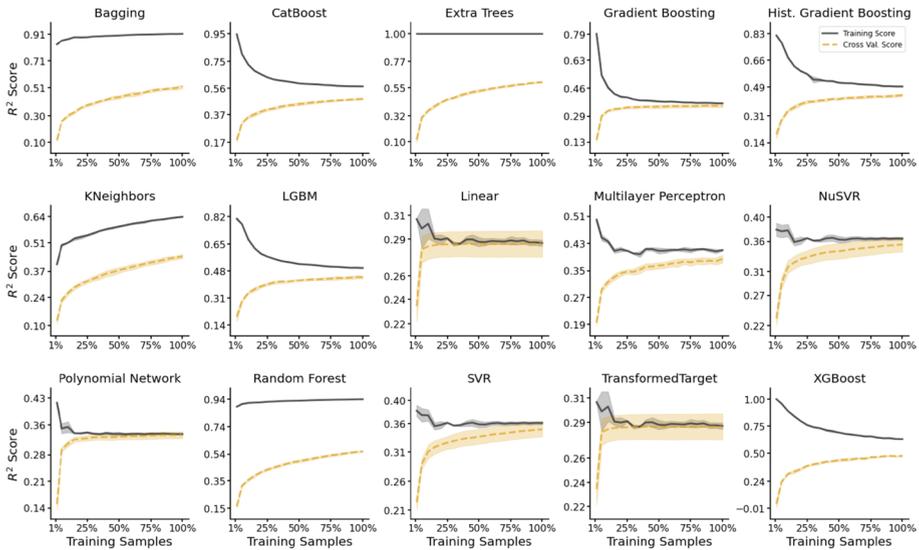
Values are ordered according to the  $R^2$  Score and RF corresponds to Random Forest

### 5.3 Results

For the uniform CPUE, the Extra Tree regressor achieved the highest  $R^2$  score with 0.65, followed by the classical Random Forest regressor with 0.61, and then one ensemble-based bagging method with 0.57. Error-based metrics also reflect the same trend in terms of results found with the  $R^2$  score. Although state-of-the-art methods XGBoost and CatBoost usually outperform traditional regressors, the results underperformed Random Forest for this particular dataset. Another classical regressor worth mentioning is the Support Vector Regressor (SVR) [47] and variations which also scored poorly when compared with the other regressors; hence it is not included in the table. This table also shows the weighted CPUE results, which is considered a more realistic fishing index. The top-5 regressors are similar to the uniform CPUE, corroborating our findings that the best regressors are viable for suggesting fishing catch prediction for the next seasons. However, estimation based on uniform CPUE performs better than the weighted CPUE.

We also analyzed the learning curves assessing the generalization power by increasing the number of training samples. This is very relevant to show the variance of the regressors relating to the model sensitivity when the training set varies. Figure 7 shows the learning curves of fifteen regressors using different splitting in the x-axis standing for the size of data sampling during the training phase with the step of 25% in size. The corresponding evaluation metric  $R^2$  score is shown on the y-axis. It is worth observing that all attributes were used for the estimation without performing attribute selection or reduction. The learning curves illustrate the data fitting learned by the model. Additionally, we find out how much we benefit from adding more training data and whether the estimator suffers more from a variance error or a bias error.

For the top-three regressors, Extra Trees, Random Forest, and Bagging, both the validation score and the training score converge to a better score value with the increasing size of the training set. Thus, we will probably benefit much from more training data with series from extra years. It is worth noticing that linear regression performs low in the score if the sampling size is larger. This shows that the problem is not easily solved with a linear



**Fig. 7** Learning curves of the regressors assessed with the coefficient of determination and by varying the size of data training sampling. The black line stands for the training estimation, while the yellow dashed line is the validation score using 5 fold cross-validation

approach. In contrast, for small amounts of data, the training score of the regressors, Gradient Boosting, Linear Regressor, nuSVR [10], and SVR are much greater than the validation score. Adding more training samples will most likely increase generalization.

## 6 Conclusion

This work built and analyzed a spatio-temporal database of fishing vessel trajectories in the Northern Adriatic sea. We started from the terrestrial AIS data of the area of interest and the fish reports of the primary fish market, Chioggia, for the years 2015, 2016, 2017, 2018. We determined the trajectories and introduced semantic attributes capable of unveiling interesting information and aspects of the original data themselves. Moreover, we gave a formal definition of two different catch distribution techniques, the *uniform* and *weighted*, and we put them at work and compared their behavior. We implemented the spatio-temporal database using MobilityDB, which provides a suitable environment for storing, querying, and visualizing trajectories of moving objects.

The ecological experts proposed some analyses on the obtained database. We first analyzed the AIS transmission anomalies – stored as a new semantic feature – that allowed us to acknowledge a concrete and progressive improvement of the data completeness in the years 2015–2018, thanks to the growing use of the AIS system in the fishing vessels and the increasing AIS data receiving coverage. We proceeded with the analysis of the two proposed distribution techniques. It resulted that the weighted distribution is a more realistic index with respect to the uniform one, able to better define the fishing ground of the species of interest. Besides, we showed how multiple aspect trajectories could assess the fishing activities, capturing spatial and temporal patterns.

Furthermore, we have built predictive models on the available dataset. Our results indicate that Machine Learning is a viable data analysis technique for fisheries and fish ecology applications. In particular, a large number of regressors were tested, aiming to predict CPUE. We cannot compare our work with the results reported in [1] because our approach used different data preparation methods and a significantly larger dataset (4 years instead of 2). Besides, additional regressors (10 in total) were also adopted in this study. Results based on three metrics, including error-based and coefficient of determination, achieve a score of 0.65 out of 1 for  $R^2$ . The result is considered a good achievement because the problem is challenging, given the fact that the dataset contains multivariate and spatio-temporal aspects to cope with.

As future work, we intend to use a more granular time component (e.g. months) than the currently used seasons. Also, a more fine-grain prediction based on fishing gear would be worthwhile to be performed as soon as more data are available. Another issue is the train-test regime. Ideally, when data is time-stamped, cross-validation should be avoided as it will not be sensitive to latent concept drift almost always present in real data. In our case, one should train on a dataset representing several years and test on the following year to model the real deployment situation. This regime will be applied when we have more years of data at our disposal. It will be interesting to compare such longitudinal train-test regime with cross-validation results on the total data available at that time. Furthermore, anomaly behavior of different species, as investigated in [23, 35], might be another interesting future direction. Also, as chlorophyll-a and sea surface temperature have been proved to be important driving factors for fish availability [44, 51], it would be interesting to use them to define a more refined approach to catch distribution. Finally, as a longer-term goal we will investigate how fishing predictions may change when the models we build are informed by climate change models.

**Acknowledgements** This paper is supported by the MASTER project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska Curie grant agreement N. 777695. Stan Matwin is thankful for the support of the Natural Sciences and Engineering Research Council of Canada. This work was also supported by the CHIST-ERA grant CHIST-ERA-19-XAI-010, by MUR (grant No. not yet available), FWF (grant No. I 5205), EPSRC (grant No. EP/V055712/1), NCN (grant No. 2020/02/Y/ST6/00064), ETAg (grant No. SLTAT21096), BNSF (grant No. KP-06-D002/5).

**Author Contributions** AR, MS, SM and FP have developed the initial idea. ER, PA collected the data. GR and CS have performed the spatio-temporal analyses and BB has conducted the last set of experiments involving machine learning algorithms. BB described the results involving machine learning. ER, FP, AR, AS and SM interpreted the results. AR, MS, AS, BB, SM wrote the first draft of the manuscript. All the authors have contributed to the manuscript revision and have read and approved the submitted version.

## Declarations

**Conflicts of interest** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Adibi P, Pranovi F, Raffaetà A, Russo E, Silvestri C, Simeoni M, Soares A, Matwin S (2019) Predicting fishing effort and catch using semantic trajectories and machine learning. In: Multiple-aspect analysis of semantic trajectories - first international workshop, MASTER 2019, Lecture notes in computer science, vol 11889. Springer, pp 83–99
2. Altman N (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician* 46(3):175–185. <https://doi.org/10.1080/00031305.1992.10475879>
3. Amiri K, Shabanipour N, Eagderi S (2018) Forecasting the catch of kilka species (*Clupeonella* spp.) using time series SARIMA models in the southern caspian sea. *Caspian Journal of Environmental Sciences* 16(4):349–358
4. Anuja A, Yadav VK, Bharti VS, Kumar NR (2017) Trends in marine fish production in Tamil Nadu using regression and autoregressive integrated moving average (ARIMA) model. *Journal of Applied and Natural Science* 9(2):653–657. <https://doi.org/10.31018/jans.v9i2.1252>
5. Bogorny V, Renso C, de Aquino AR, de Lucca Siqueira F, Alvares LO (2014) Constant-a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS* 18(1):66–88
6. Breiman L (1996) Bagging Predictors. *Machine Learning* 24(2):123–140
7. Breiman L (2001) Random forests. *Machine Learning* 45:5–32
8. Brüggemann S, Bereta K, Xiao G, Koubarakis M (2016) Ontology-based data access for maritime security. In: European semantic web conference. Springer, pp 741–757
9. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G (2013) API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: languages for data mining and machine learning, pp 108–122
10. Chang CC, Lin CJ (2002) Training nu-support vector regression: theory and algorithms. *Neural Computation* 14(8):1959–1977
11. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16, p 785–794. Association for computing machinery, New York, NY, USA. <https://doi.org/10.1145/2939672.2939785>
12. Claramunt C, Ray C, Camossi E, Jousselet A, Hadzagic M, Andrienko GL, Andrienko NV, Theodoridis Y, Vouros GA, Salmon L (2017) Maritime data integration and analysis: recent progress and research challenges. In: Proceedings of the 20th international conference on extending database technology, pp 192–197
13. Dividino R, Soares A, Matwin S, Isenor AW, Webb S, Brousseau M (2018) Semantic integration of real-time heterogeneous data streams for ocean-related decision making. In: Big data and artificial intelligence for military decision making. STO
14. Estrada J, Silva C, Yáñez E, Rodríguez N, Pulido-Calvo I (2007) Monthly catch forecasting of anchovy *Engraulis ringens* in the north area of Chile: Non-linear univariate approach. *Fisheries Research* 86(2):188–200
15. Etemad M, Etemad Z, Soares A, Bogorny V, Matwin S, Torgo L (2020) Wise sliding window segmentation: A classification-aided approach for trajectory segmentation. In: Canadian conference on artificial intelligence. Springer, pp 208–219
16. Fernandes JA, Rutterford L, Simpson SD, Butenschön M, Frölicher TL, Yool A, Cheung WWL, Grant A (2020) Can we project changes in fish abundance and distribution in response to climate? *Global Change Biology* 26(7):3891–3905
17. Ferrà C, Tasseti AN, Grati F, Pellini G, Polidori P, Scarcella G, Fabi G (2018) Mapping change in bottom trawling activity in the Mediterranean Sea through AIS data. *Marine Policy* 94:275–281
18. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning* 63(1):3–42. <https://doi.org/10.1007/s10994-006-6226-1>
19. Gomes HM, Bifet A, Read J, Barddal JP, Enembreck F, Pfahringer B, Holmes G, Abdesslem T (2017) Adaptive random forests for evolving data stream classification. *Machine Learning* 106:1–27. <https://doi.org/10.1007/s10994-017-5642-8>
20. Jie X, Chaozhong W, Zhijun C, Xiaoxuan C (2017) A novel estimation algorithm for interpolating ship motion. In: 2017 4th International conference on transportation information and safety (ICTIS). IEEE, pp 557–562
21. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp 3149–3157
22. Kehinde O, Joseph G (2018) Time series modelling for forecasting artisanal fish production in Nigeria. *Int J Adv Multidiscip Res* 5(7):10–17

23. Kontopoulos I, Chatzikokolakis K, Zissis D, Tserpes K, Spiliopoulos G (2020) Real-time maritime anomaly detection: detecting intentional AIS switch-off. *Int J Big Data Intell* 7(2):85–96
24. Kontopoulos I, Varlamis I, Tserpes K (2021) A distributed framework for extracting maritime traffic patterns. *International Journal of Geographical Information Science* 35(4):767–792
25. Le Guyader D, Ray C, Brosset D (2018) Identifying small-scale fishing zones in France using AIS data. In: *Advances in shipping data analysis and modeling: tracking and mapping maritime flows in the age of big data*, chap 15. Taylor & Francis
26. Le Guyader D, Ray C, Gourmelon F, Brosset D (2014) Defining high-resolution dredge fishing grounds with automatic identification system (AIS) data. *Aquat Living Resour* 30:39. <https://doi.org/10.1051/alr/2017038>
27. Maina I, Kavadas S, Somarakis S, Tserpes G, Stratis G (2016) A methodological approach to identify fishing grounds: A case study on Greek trawlers. *Fisheries Research* 183:326–339
28. Mehmood Q, Sial M, Sharif S, Hussain A, Riaz M, Shaheen N (2020) Forecasting the fisheries production in Pakistan for the year 2017–2026, using Box-Jenkin's methodology. *Pakistan Journal of Agricultural Research* 33(1):140–145
29. Mello RDS, Bogorny V, Alvares LO, Santana LHZ, Ferrero CA, Frozza AA, Schreiner GA, Renso C (2019) MASTER: A multiple aspect view on trajectories. *Transactions in GIS* 23(4):805–822
30. Mills CM, Townsend SE, Jennings S, Eastwood PD, Houghton CA (2006) Estimating high resolution trawl fishing effort from satellite-based vessel monitoring system data. *ICES Journal of Marine Science* 64(2):248–255
31. Nurdin S, Ahmad Mustapha M, Lihan T, Abd Ghaffar M (2015) Determination of potential fishing grounds of Rastrelliger kanagurta using satellite remote sensing and GIS technique. *Sains Malaysiana* 44(2):225–232
32. Paige Eveson J, Hobday A, Hartog J, Spillman C, Rough K (2015) Seasonal forecasting of tuna habitat in the Great Australian Bight. *Fisheries Research* 170:39–49
33. Parent C, Spaccapietra S, Renso C, Andrienko G, Andrienko N, Bogorny V, Damiani ML, Gkoulalas-Divanis A, Macedo J, Pelekis N et al (2013) Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)* 45(4):42
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
35. Petry LM, Soares A, Bogorny V, Brandoli B, Matwin S (2020) Challenges in vessel behavior and anomaly detection: from classical machine learning to deep learning. In: *Advances in artificial intelligence - 33rd canadian conference on artificial intelligence, Lecture notes in computer science*, vol 12109. Springer, pp 401–407
36. Prokhorenkova L, Gusev G, Vorobev A, Dorigush AV, Gulina A (2018) CatBoost: unbiased boosting with categorical features. In: *Proceedings of the 32nd international conference on neural information processing systems, NIPS'18*. Curran Associates Inc., Red Hook, NY, USA, p 6639–6649
37. Raman RK, Das BK (2019) Forecasting shrimp and fish catch in chilika lake over time series analysis. *Time Series Analysis - Data, Methods, and Applications*
38. Raman RK, Sathianandan TV, Sharma AP, Mohanty BP (2017) Modelling and forecasting marine fish production in odisha using seasonal ARIMA model. *National Academy Science Letters* 40(6):393–397. <https://doi.org/10.1007/s40009-017-0581-2>
39. Rovinelli G, Matwin S, Pranovi F, Russo E, Silvestri C, Simeoni M, Raffaetà A (2021) Multiple aspect trajectories: a case study on fishing vessels in the Northern Adriatic sea. In: *BMDA 2021: 4th International workshop on big mobility data analytics, CEUR workshop proceedings*, vol 2841. CEUR-WS.org
40. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536. <https://doi.org/10.1038/323533a0>
41. Santipantakis GM, Glenis A, Patroumpas K, Vlachou A, Doulkeridis C, Vouros GA, Pelekis N, Theodoridis Y (2020) Spartan: Semantic integration of big spatio-temporal data from streaming and archival sources. *Future Generation Computer Systems* 110:540–555. <https://doi.org/10.1016/j.future.2018.07.007>
42. Shepperson JL, Hintzen NT, Szostek CL, Bell E, Murray LG, Kaiser MJ (2018) A comparison of VMS and AIS data: the effect of data coverage and vessel position recording frequency on estimates of fishing footprints. *ICES Journal of Marine Science* 75(3):988–998
43. Soares A, Dividino R, Abreu F, Brousseau M, Isenor AW, Webb S, Matwin S (2019) CRISIS: Integrating AIS and ocean data streams using semantic web standards for event detection. In: *International conference on military communications and information systems*

44. Solanki H, Dwivedi R, Nayak S, Somvanshi V, Gulati D, Pattnayak S (2003) Fishery forecast using OCM chlorophyll concentration and AVHRR SST: validation results off Gujarat coast. India. *International Journal of Remote Sensing* 24(18):3691–3699
45. de Souza EN, Boerder K, Matwin S, Worm B (2016) Improving fishing pattern detection from satellite AIS using data mining and machine learning. *PLoS One* 11(7):e0158248
46. Torres-Irineo E, Salas S, Euán-Ávila JI, Palomo LE, Quijano Quiñones DR, Coronado E, Joo R (2021) Spatio-temporal determination of small-scale vessels fishing grounds using a vessel monitoring system in the southeastern gulf of Mexico. *Frontiers in Marine Science* 8:542
47. Vapnik VN (1995) *The nature of statistical learning theory*. Springer-Verlag, New York Inc
48. Vespe M, Gibin M, Alessandrini A, Natale F, Mazzarella F, Osio GC (2016) Mapping EU fishing activities using ship tracking data. *Journal of Maps* 12:520–525
49. Wen Y, Zhang Y, Huang L, Zhou C, Xiao C, Zhang F, Peng X, Zhan W, Sui Z (2019) Semantic modelling of ship behavior in harbor based on ontology and dynamic Bayesian network. *ISPRS International Journal of Geo-Information* 8(3). <https://doi.org/10.3390/ijgi8030107>
50. Yadav AK, Das KK, Das P, Raman RK, Kumar J, Das BK (2020) Growth trends and forecasting of fish production in Assam, India using ARIMA model. *Journal of Applied and Natural Science* 12(3):415–421
51. Yusop SM (2021) Determination of spatio-temporal distribution of *Rastrelliger kanagurta* using modelling techniques for optimal fishing. *Journal of Coastal Conservation* 25(1):15–15
52. Zimányi E, Sakr M, Lesuisse A (2020) MobilityDB: A Mobility Database Based on PostgreSQL and PostGIS. *ACM Trans Database Syst* 45(4):1–42



**Bruno Brandoli** holds a B.Sc. in computer engineering and received his M.Sc. and Ph.D. degrees from the University of São Paulo, the latter in 2016. He was an Associate Professor in Brazil for 6 years. In 2019, he joined Dalhousie University as a Faculty member teaching AI-related courses. His current research interests include complex networks and deep learning with applications lately focused on ocean sciences.



**Alessandra Raffaetà** MSc (1994) and PhD (2000) in Computer Science, University of Pisa - is an assistant professor at Ca' Foscari University of Venice (Italy). Her research interests include Data warehouses, GISs, spatio-temporal reasoning, design and formal semantics of programming languages and constraint logic programming. She published over 60 papers on international journals and conferences. She was member of the program committee of several international conferences and she participated to several national and international research projects.



**Marta Simeoni** received her master degree in Computer Science at the University of Udine in 1995 and the PhD in Computer Science at the University of Rome “La Sapienza” in 2000. Since 2000, she is Assistant Professor at Ca’ Foscari University of Venice. Her present research interests are in Bioinformatics and Ecoinformatics in the field of Biological/Ecological Systems Modeling and Analysis.



**Pedram Adibi** is interested in applications of Machine Learning to promote sustainability. He obtained his master’s degree in Computer Science with a focus on Machine Learning from Dalhousie University in 2020. Since then, he has been working as a ML engineer helping build solutions for a more sustainable aquaculture industry.



**Fateha Khanam Bappee** is a postdoctoral fellow in the Faculty of Computer Science at Dalhousie University. She received the PhD degree in computer science from Dalhousie University in 2021 under the supervision of Stan Matwin. Her research interests include machine learning, data science, big data analytics, spatial and spati-temporal analytics.



**Fabio Pranovi** full professor in Ecology at the Ca' Foscari University (Environmental Sciences, Informatics and Statistics Dept.); main fields of interest: ecosystem approach in the management of renewable resources exploitation in marine environment and possible effects of climate changes on the structure and functioning of marine ecosystems; authors of about 130 scientific papers.



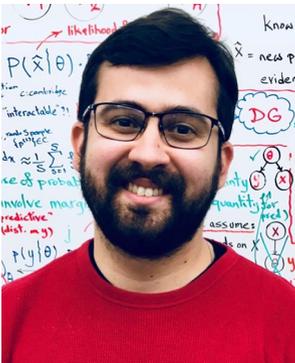
**Giulia Rovinelli** is a second-year Master's degree student of Computer Science at Ca' Foscari University. She received the Bachelor's Degree in Computer Science from Ca' Foscari University in 2020 (summa cum laude). In her studies she is focusing on artificial intelligence, machine learning and data science.



**Elisabetta Russo** is a marine biologist and ecologist with a PhD in Environmental Science. She is a Post-Doctoral Researcher at Ca' Foscari University of Venice since 2020. Her main research interest is focused on marine ecology, and in particular on the assessment of the fishery activities at sea and the fishing stocks status, through the application of up-to-date methodologies such as the use of AIS data for the spatialization of fishing effort and landings.



**Claudio Silvestri** received his M.S. degree in Computer Science in 2002 and a Ph.D. degree in Computer Science in 2006. He is an Assistant Professor at the Department of Environmental Sciences, Informatics and Statistics of Ca' Foscari University of Venice and a researcher at the European Center For Living Technology inter-university consortium. His research interests are in the areas of databases, distributed and high performance computing, spatio-temporal data processing, privacy in mobile computing. He published over 50 papers on peer reviewed international journals and conferences.



**Amilcar Soares** is an Assistant Professor at the Memorial University of Newfoundland at the Department of Computer Science. His research interests include spatiotemporal data segmentation, classification, enrichment, and visualization. He holds a Ph.D. in computer science from Federal University of Pernambuco. He has been involved in several research projects funded by the Natural Sciences and Engineering Research Council of Canada (NSERC), Department of Fisheries and Oceans (DFO), Transport Canada (TC), and Defence Research and Development Canada (DRDC).



**Stan Matwin** is the Director of the Institute for Big Data Analytics at Dalhousie University, Halifax, Nova Scotia, Canada where he is a Professor and Canada Research Chair in Interpretability for Machine Learning. He is also a Distinguished Professor of the University of Ottawa and a Professor at the Institute of Computer Science of the Polish Academy of Sciences. His main research results are machine learning, particularly applied to data derived from the world's oceans, text mining, applications of machine learning, and in data privacy. Stan Matwin is a EurAI Fellow, a Fellow of the Canadian AI Association and a recipient of its Lifetime Achievement Award.

## Authors and Affiliations

**Bruno Brandoli**<sup>1</sup> · **Alessandra Raffaetà**<sup>2</sup>  · **Marta Simeoni**<sup>2,3</sup> · **Pedram Adibi**<sup>1</sup> · **Fateha Khanam Bappee**<sup>1</sup> · **Fabio Pranovi**<sup>2</sup> · **Giulia Rovinelli**<sup>2</sup> · **Elisabetta Russo**<sup>2</sup> · **Claudio Silvestri**<sup>2</sup> · **Amilcar Soares**<sup>4</sup> · **Stan Matwin**<sup>1,5</sup>

Bruno Brandoli  
brunobrandoli@dal.ca

Marta Simeoni  
simeoni@unive.it

Pedram Adibi  
pedram.adibi@dal.ca

Fateha Khanam Bappee  
ft487931@dal.ca

Fabio Pranovi  
fpranovi@unive.it

Giulia Rovinelli  
867381@stud.unive.it

Elisabetta Russo  
elisabetta.russo@unive.it

Claudio Silvestri  
silvestri@unive.it

Amilcar Soares  
amilcarsj@mun.ca

Stan Matwin  
stan@dal.ca

<sup>1</sup> Institute for Big Data Analytics, Dalhousie University, B3H 1W5 Halifax, Canada

<sup>2</sup> Università Ca' Foscari Venezia, Venezia, Italy

<sup>3</sup> European Centre for Living Technology (ECLT), Venice, Italy

<sup>4</sup> Department of Computer Science, Memorial University of Newfoundland, St. Johns, Canada

<sup>5</sup> Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland