











# Predicting Fishing Effort and Catch Using Semantic Trajectories and Machine Learning

Pedram Adibi<sup>1</sup>, Fabio Pranovi<sup>2</sup>, Alessandra Raffaetà<sup>2</sup>,  
Elisabetta Russo<sup>2</sup>, Claudio Silvestri<sup>2</sup>, Marta Simeoni<sup>2</sup>,  
Amilcar Soares<sup>1</sup>, and Stan Matwin<sup>1,3</sup>

- <sup>1</sup> Institute for Big Data Analytics, Dalhousie University, Halifax, Canada  
{pedram.adibi,amilcar.soares,stan}@dal.ca
- <sup>2</sup> Dipartimento di Scienze Ambientali, Informatica e Statistica,  
Università Ca' Foscari Venezia, Venezia, Italy  
{pranovi,raffaeta,russo,silvestri,simeoni}@unive.it
- <sup>3</sup> Polish Academy of Sciences, Warsaw, Poland

**Abstract.** In this paper we explore a unique, high-value spatio-temporal dataset that results from the fusion of three data sources: trajectories from fishing vessels (obtained from terrestrial Automatic Identification System, or AIS, data feed), the corresponding fish catch reports (i.e., the quantity and type of fish caught), and relevant environmental data. The result of that fusion is a set of semantic trajectories describing the fishing activities in Northern Adriatic Sea over two years. We present early results from an exploratory analysis of these semantic trajectories, as well as from initial predictive modeling using Machine Learning. Our goal is to predict the Catch Per Unit Effort (CPUE), an indicator of the fishing resources exploitation useful for fisheries management. Our predictive results are preliminary in both the temporal data horizon that we are able to explore and in the limited set of learning techniques that are employed on this task. We discuss several approaches that we plan to apply in the near future to learn from such data, evidence, and knowledge that will be useful for fisheries management. It is likely that other centers of intense fishing activities are in possession of similar data and could use the methods similar to the ones proposed here in their local context.

**Keywords:** Spatio-temporal data · Fisheries · Machine Learning · Semantic trajectories · AIS

---

The authors would like to thank NSERC (Natural Sciences and Engineering Research Council of Canada) for financial support. This work was partially supported by project MASTER (Marie Skłodowska-Curie agreement N. 777695), which has received funding from the EU Horizon 2020 Programme.

## 1 Introduction

In this paper, we present early results from an ongoing international research project in which mobility data researchers and fishery ecologists collaborate closely. In our project, we explore a unique, high-value dataset that results from the fusion of three data sources: trajectories from fishing vessels, the corresponding fish catch reports (i.e., the quantity and type of fish caught), and relevant environmental data. The goal of this project is to predict the future Catch Per Unit Effort (CPUE) from the past data. CPUE is an indicator of fishing resources exploitation that allows for assessing the pressure of these activities at the ecosystem level. Intuitively, a decrease of CPUE indicates a situation of over-exploitation, a steady CPUE value points out a sustainable exploitation of the fishery resources and an increase of its value corresponds to a healthy and growing population. CPUE is therefore a key indicator for fisheries management since it could help to define the sustainability of the fishing activities in the area of interest: an accurate forecast of CPUE could help decision makers to obtain a sustainable fishing business by adapting the fisheries management plans on the basis of the forecast results. Here we discuss and present early results from the use of Machine Learning techniques to predict the CPUE in the Northern Adriatic Sea.

We believe this research demonstrates the opportunities provided by mobility data analysis to gain insights and evidence that can guide fisheries management decisions. Such decisions will have significant environmental and economic consequences at the regional, national, and eventually global level. Our results are preliminary, both in the temporal data horizon that we are able to explore, and in the broader set of techniques that could be employed on this task. It is likely that other centers of intense fishing activities are in possession of similar data and could use the methods similar to the ones proposed here in their local context.

The Northern Adriatic Sea area, on which our work is based, needs tools and models that can assist fisheries management at the macro scale. This area, known for its very high productivity, is recognized to be one of the most exploited area of the Mediterranean Sea, causing an over-exploitation of the fish resources. In this context, the development of effective fishery management plans is needed to make fishing activities sustainable and ensure a productive and healthy ecosystem. Currently, different management measures are used in the Northern Adriatic Sea (e.g., the permanent ban of trawling activities within 3 nm of the coast, the seasonal biological rest period for trawlers). In this context, forecasting the fishing activities and their catches in space and time represents a step forward to assess the efficiency of these measures and develop new ones.

Recently, several works report the use of mobility-tracking technologies, such as Automatic Identification System (AIS) to monitor fishing activities. In its inception, AIS was primarily designed as a navigational aid to avoid vessel collisions, but nowadays it has become - often due to its open nature - the primary source of data about fishing-related activities. In our setting, we have access to terrestrial AIS data, i.e., AIS data sent by ships and received by ground stations

on the Italian coast of Northern Adriatic. Vessels transmit their position at a variable rate, from 2s up to two minutes. We use AIS data to reconstruct, in time and space, the fishing trajectories. The latter have been enriched with the available environmental data, such as daily surface temperature, chlorophyll-a, and wave height. A further valuable piece of information available for this work is the landing reports of the Chioggia’s fishing market, which is the primary market of the Northern Adriatic basin.

The two main research questions that guide this work are: (i) How can we improve our knowledge of the spatio-temporal aspects of the fishing activities in the Northern Adriatic Sea?; and (ii) How can we predict the CPUE for next year?

Data fusion, management, and Machine Learning techniques from the mobility data analysis are applicable to provide evidence-based answers to these questions. In this paper, we focus on the use of semantic trajectories of fishing vessels and predictive modeling using spatio-temporal Machine Learning techniques, and we address mainly the last of the above questions.

The contributions of this work are the proposal of a framework that: (i) integrates the heterogeneous data sources; (ii) extracts knowledge from the integrated data using semantic trajectory modeling; and (iii) applies Machine Learning (e.g., Random Forest) to learn from those semantic trajectories a model for forecasting the CPUE.

This paper is structured as follows: Sect. 2 describes the related work in the literature concerning semantic trajectories and fishing activities forecast. Section 3 illustrates the architecture of the developed system and describes in details the data sources, how the raw data have been fused and incorporated in a semantic model, and the model developed for prediction analysis. Section 4 reports the predictive model results and Sect. 5 draws some concluding remarks and illustrates possible future developments.

## 2 Related Works

In this section, we discuss related works regarding (i) sea data fusion and semantic trajectories, which is the concept used for enriching complex objects like fishing ships with relevant information; and (ii) fishing activities forecast, which is the final goal of our predictive model.

### 2.1 Data Fusion of Sea Data and Semantic Trajectories

Combining the AIS, trading transactions, and environmental variables from the vessels into a single representation is challenging. Several strategies were proposed to deal with the fusion of heterogeneous ocean data properly. For example, the paper [18] shows 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 [6], the authors propose a model for

integration and analysis of data for vessel movement in a real-time maritime situation awareness system, also using semantic web techniques and tools. Unlike the previous methods, we model our trajectory data with a semantic model. By considering data sources such as AIS, environmental variables, and landing reports, the trajectory of every fishing vessel becomes a complex object with several data dimensions that are contextual to the movement.

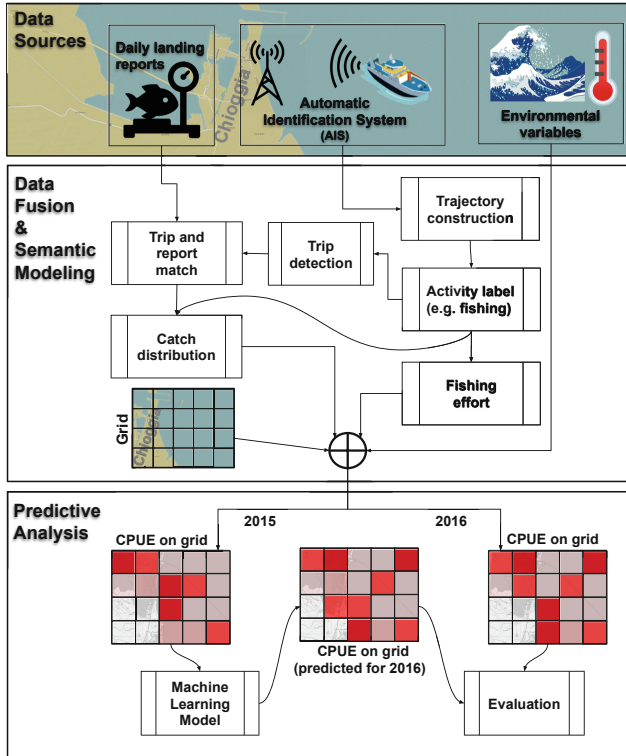
Several semantic models for trajectory data have been proposed, such as the *stops* and *moves* [17], CONSTANT [2], and recently MASTER [13]. This last model is more flexible and expressive since it allows for enriching trajectories with complex objects and it provides not only a conceptual model but also a logical schema in the Resource Description Framework (RDF) and a triplestore based on NoSQL databases for maintaining RDF data.

By following the MASTER semantic model, the trajectory of fishing vessels can be represented as a *multiple aspect trajectory*. The AIS data constitute the sequence of spatio-temporal points. Moreover, the MASTER model introduces the concept of *aspect* which consists of “a real-world fact that is relevant for the trajectory data analysis” [13]. It distinguishes between *volatile* aspects—which are usually associated with the trajectory points, since they vary during the object movement—and *long term* aspects—which do not change during an entire trajectory, and hence they are related to the whole trajectory. For instance, for vessel trajectories, the speed is a volatile aspect, whereas the fishing gear type is a long term aspect.

## 2.2 Fishing Activities Forecast

The literature on fishing activities forecast is vast and can be analyzed in several ways. From the fishing management perspective, works like [16] propose a seasonal forecast system that combines environmental and fish habitat data (collected by fish tagging) to predict tuna distribution. In [15], the authors, integrate satellite data and statistical models output to investigate the relationship between sea surface temperature and chlorophyll-a concentration, and also define simple methods to forecast potential fishing grounds. The work of [9] tries to forecast 1-month catches considering only the anchovy catches in previous months as inputs. Similarly, to [15,16], we use environmental data (e.g., chlorophyll-a and sea surface temperature). Also similarly to [9] we use fish catch information to predict future catches. Unlike all of them, we use wave height as an environmental variable in our model.

From a perspective that considers the geolocation technology used to track ships, some works use Vessel Monitoring System (VMS) [12,19], satellite images [15] or AIS [10,20,22]. Most of these works focus on training models to forecast when a vessel is performing a fishing activity or not. Different types of fishing ships (e.g., long-liners, purse-seiners, etc.) have different types of movement patterns. Predicting these patterns depends on the training data given to the machine learning model [19,20], or the domain specialist ability to create rules that reflect these patterns [14]. In this work, we use domain knowledge from specialists to determine the activity of vessels (e.g., fishing or not) on their



**Fig. 1.** An overview of all the steps of the framework for predicting fishing catches.

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. To the best of our knowledge, no work in the literature uses a combination of AIS, fishing catch reports, and environmental variables to forecast CPUE.

### 3 A Framework for Predicting CPUE

In this section, we present the bird’s eyes view of the architecture of the system we have developed. We then discuss the individual data sources—AIS data, catch data (landing reports), environmental data—(Sect. 3.1), and the spatio-temporal mapping of data as well as the fusion of the individual sources (Sect. 3.2). The section is rounded up with a brief discussion of the Machine learning method we have selected to build the prediction model (Sect. 3.3). Schematics of the system is shown in Fig. 1.

### 3.1 Data Sources

**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 2016. A total of 70 (2015) and 77 (2016) trawlers, with a length overall above 15 m and belonging to the Chioggia navy, 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 the ones 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).

**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 the biennium 2015–2016 in the Northern Adriatic Sea. The records pertain to 82 fishing vessels, and a total of 17921 fishing trips over the two years.

A graph of total monthly landings for the two years with the contribution of the five most harvested species is shown in Fig. 2. Seasonality of the data is evident by visual comparison of the annual trends. The graph shows zero landing in August for both years, which reflects the fishing ban. It is also visible from the graph that 2015 landing amounts were higher than 2016 for almost all the months.

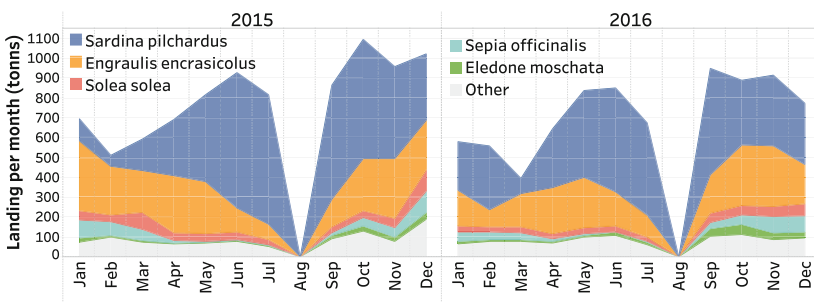


Fig. 2. Chioggia’s total monthly landing and species contribution in 2015–16

**Environmental Data.** We considered the following environmental information in order to enrich the trajectories of the fishing vessels:

- Sea Surface Temperature (in kelvin) [7]
- Sea Daily Chlorophyll-a Concentration (in  $mg/m^3$ ) [7]
- Spectral significant wave height (in meters) [7]

The sea surface temperature and the chlorophyll-a influence the species distribution, while the wave height affects the fishermen 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 Data Fusion and Semantic Modeling

**Activity Labeling and Trip Detection from AIS Data.** Trajectories have been reconstructed by linear interpolation of the raw AIS data. While performing the reconstruction, all implausible points have been discarded. In particular, all the movements that were not physically feasible concerning a maximum possible boat speed were removed.

A trajectory is therefore defined as a sequence of segments, and it is enriched with the following data: MMSI (boat identifier), departure time of the trip (exit from harbour area), departure port, position of the segment with respect to the ports areas, average speed, activity of the boat within the segment and fishing gear (this can be obtained through the MMSI).

The activity attribute is an integer value distinguishing among the following situations: (0): *in port*; (1): *exiting from*; (2): *entering to port*; (3): *fishing*; (4): *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. In case none of the previous situations occurs, 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 equipped gear, the boat is assumed to be in a *fishing* phase; otherwise, it is assumed to be in a *navigation* phase.

This trajectory can be modeled as a *multiple aspect trajectory*, following MASTER model [13]. Indeed, as minimum granularity to attach semantic information, we do not consider a single spatio-temporal point, but we annotate segments since we want to highlight the presence of homogeneous trajectory portions, which are the appropriate granularity level for our analyses. Hence the pieces of information we listed above can be classified as

- *long-term aspects*: MMSI, departure time of the trip, departure port and the gear used for fishing because they do not change during the entire trajectory;
- *volatile aspects*: average speed and activity of the boat since they frequently vary during the object movement and they can be associated with a segment.

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 together spatial, temporal and semantic

features. In the rest of the paper, we denote with  $T$  the resulting set of multiple aspect trajectories.

**The Spatial Grid.** Some of the concepts described later in this section (i.e., fishing effort and CPUE) are defined over an area. In order to calculate those values, the area under study is partitioned into a square grid with  $5 \times 5$  km cell size. Fishing effort and CPUE are then calculated over the individual grid cells. In addition, the grid is used in the calculation of weighted catch distribution (described later), and as the data format for prediction modeling (Sect. 3.3).

**Calculation of Fishing Effort over Grid Cells.** After reconstructing the 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 cell during a fixed period of time is defined as the ratio between the area of the cell “swept” by vessels while fishing during the given 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 cell in the area of interest, by  $p$  a time period (could be day, month, etc.) and by  $g$  a gear (small and large bottom otter trawl, Rapido and midwater pair trawl).

**Definition 1.** Let  $c$  be a cell,  $p$  a time period and  $g$  a gear. The *fishing effort* wrt the gear  $g$  in the cell  $c$  during the time period  $p$  is defined as follows:

$$fe(c, p, g) = \frac{(\sum_{tr \in T, gear(tr)=g} len(tr, c, p)) * gear\_width(g)}{area(c)} \quad (1)$$

where

- $T$  is the set of multiple aspect trajectories;
- $len(tr, c, p)$  returns the sum of the lengths of the fishing segments of trajectory  $tr$  falling in cell  $c$  during time period  $p$ ;
- $gear\_width(g)$  is the width of the net of gear  $g$ ;
- $area(c)$  is the total area of the cell  $c$ .

It is worth noting that we can obtain the total fishing effort in a cell  $c$  during a time period  $p$  by summing up the fishing effort for each gear. Indeed, for our analyses in Sect. 4 we will use the fishing effort for a particular gear, i.e., Rapido.

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.



**Assigning Landing Reports to Trips.** The landing dataset provided by the Chioggia’s fish market contains information about each trading transaction, including the landing date, MMSI of the seller, the species, and the quantity of fish. We have 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). Such a trip has to respect some constraints: it has to last at least 1 h, and have a minimum length of 2 km, from which, at least 100 m have to be classified as fishing. Arrivals after 4 PM are associated with transactions occurring the next day. Assignment of the quantity (weight) of the fish to a vessel is called a *catch*.

**Catches Distribution over Trips.** The association of fish catches with trajectories allows us to add a further *volatile* aspect to our multiple aspect trajectories. In fact, we can distribute the fish associated with a trajectory along with its fishing segments. In particular, we can employ two different techniques:

- *uniform* distribution, or
- *weighted* distribution.

In the first case, for each trading transaction, the amount of fish is *uniformly* distributed along with the fishing segments of the corresponding trip. Each fishing segment of the trajectory is associated with a portion of the total amount of fish, proportional to its length.

Of course, the assumption of uniform catch distribution is a simplification of reality. As an improvement, a *weighted* distribution of catches is also considered. The idea behind this approach is that the areas where more vessels are fishing, during a given time period, are more likely to have higher catch rates. A preliminary method based on this idea is implemented as follows. First, the number of distinct vessels that were detected fishing in each grid cell in a time period is computed. Then the amounts of catch over each segment, derived using the uniform distribution, is weighted by the vessel counts in the cell containing the segment. The weights are normalized by the sum of vessel counts in cells that cover all the fishing segments in a trip, so they add up to 1.

Based on this piece of information, we can compute the quantity of fish caught in each cell during a period of time by boats having a particular gear  $g$ .

**Definition 2.** Let  $c$  be a cell,  $p$  a time period and  $g$  a gear, the fish *catch* wrt to the gear  $g$  in cell  $c$  during the time period  $p$  is defined as follows:

$$catch(c, p, g) = \sum_{tr \in T, gear(tr)=g} quantity(tr, c, p) \quad (2)$$

where

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

**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.

**Definition 3.** Let  $c$  be a cell,  $p$  a time period and  $g$  a gear, the *catch-per-unit-effort (CPUE)* wrt to the gear  $g$  in cell  $c$  during the time period  $p$  is defined as follows:

$$cpue(c, p, g) = \frac{catch(c, p, g)}{fe(c, p, g)} \quad (3)$$

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.

### 3.3 Predictive Modeling

The objective of the modeling procedure described in this section is the prediction of average *monthly* CPUE values for individual grid cells. First, we describe the modeling data, which consists of *daily* values for the model attributes (or variables) per grid cell. Then, we follow with a brief background on the chosen machine learning method—Random Forest (RF). Finally, we explain how we adjust the temporal granularity to obtain the desired *monthly* output from the model, which is built using the *daily* model attributes.

**Modeling Data.** Modeling data maps the data onto a spatio-temporal grid, producing *records* (or *instances*) each of which corresponds to a *date*, and a *spatial grid cell* (grid described in Sect. 3.2). Each *record* is comprised of the response attribute—CPUE—and a set of predictive attributes, all pertaining to the same date and grid cell. The predictive attributes are as follows.

- Environmental attributes (described in Sect. 3.1): daily chlorophyll-a concentration, daily sea surface temperature, and daily spectral wave height; each attribute re-sampled over the grid cells.
- Temporal attributes that preserve seasonality: month of year (1–12), day of year (1–365), week of year (1–53), seasons (four quarters starting in January).
- Spatial attributes: latitude and longitude of the grid cell centres.

CPUE is calculated using fishing effort, which in turn, depends on the type of fishing gear. Among the four fishing gears described in Sect. 3.1, Rapido has the largest share in the dataset with 48% of the records. For this reason, we have limited our presentation in this paper to the model trained and tested on data

for the vessels with the Rapido gear. Also, in this study, CPUE is calculated based on the total catch amounts for all species in the landing dataset.

As described in Sect. 3.2, CPUE is defined on grid cells with some fishing activity in a given period. Consequently, for a given day, the dataset only includes the fished grid cells. It follows that the size of the modeling dataset is obtained by  $\sum_d |C|_d$  where  $|C|_d$  is the number of fished cells on date  $d$ . Size of the dataset for Rapido vessels amounts to 51262 records over the two year period.

**Machine Learning Method.** The task of prediction modeling of CPUE presents a regression problem. RF [5] was chosen as the regression method due to the following considerations. It does not require any assumption about the distribution of the model attributes. It can take numeric and categorical attributes, and it does not require scaling of the attributes. Moreover, RF does not result in instability of output values when presented with predictive attributes with values outside the range of the training data. Besides RFs, we have experimented with several other regression methods, e.g. linear regression with LASSO and with the Support Vector Machines. RF, however, outperformed the other methods.

RF is an ensemble learning method based on decision trees, introduced by Breiman. Ensemble learning methods can improve accuracy, and reduce bias and variance by combining outputs of many base learners [8]. In the case of regression RF, numerous regression decision trees are trained, and the model output is obtained by averaging the outputs of the individual trees. RF uses bootstrap aggregating (Bagging) to construct individual trees that are trained independent of each other. Bagging is an ensemble learning method that trains its base learners on bootstrap samples—samples that are randomly drawn with replacement from a dataset of same size [4]. RF introduces further randomization in the construction of the trees by taking a random subset of the predictive attributes at each node, and selecting one from the random subset to split on.

**Adjusting Temporal Granularity.** Even though the prediction of average *monthly* CPUE is the main interest of this study, the predictive environmental attributes (e.g. wave height and water surface temperature) affect the fishing activity on a *daily* basis. Therefore, aggregating such attributes on a monthly basis prior to modeling would result in losing the information pertaining to the daily cause and effects of those attributes. To preserve that information, first the RF regression method is performed using the *daily* training data consisting of the 2015 dataset, which produces *daily* predictions for individual grid cells for the year 2016. Then, *monthly* predictions for each cell is calculated by averaging the *daily* predictions for the respective cell over the month. Averaging is used to aggregate the predicted CPUE values, as opposed to summation, because predicted CPUE values are ratios and do not produce a meaningful sum.

The resulting average monthly predictions for individual cells are considered to be the model output. Evaluation of the model is then done against similarly calculated *monthly* CPUE averages per grid cell, using the actual 2016 data. Equation (4) shows the calculation of actual and predicted average CPUE for a

given cell  $c$  over a given period  $p$  (e.g. month)<sup>1</sup>, respectively denoted by  $y_{c,p}$  and  $\hat{y}_{c,p}$ ,

$$y_{c,p} = \frac{1}{|D_{c,p}|} \sum_{d \in D_{c,p}} cpue(c, d) \quad \text{and} \quad \hat{y}_{c,p} = \frac{1}{|D_{c,p}|} \sum_{d \in D_{c,p}} \widehat{cpue}(c, d) \quad (4)$$

where  $D_{c,p}$  indicates the set of all days  $d$  in period  $p$  for which cell  $c$  has a CPUE value.  $cpue(c, d)$  and  $\widehat{cpue}(c, d)$  are respectively the actual and predicted daily CPUE values for cell  $c$  on day  $d$ .

## 4 Experiments and Results

Two models are built and evaluated in this experiment: one for CPUE values calculated using the *uniform* catch distribution, and one for the *weighted* distribution. The distinction between the two distributions is described in Sect. 3.2. The models are trained on year 2015 (training data), then prediction and evaluation are done for year 2016 (test data) of the two year dataset. The reason for splitting the data by year for training and testing is the highly seasonal nature of the data set, as described in Sect. 3.1.

**Baseline Model.** The baseline, which is used as the benchmark to compare with our RF model, is to simply use the last observed value as the forecast (prediction)—known as naïve forecasting. This baseline is chosen rather than results from other regression models, because it is a standard practice in forecast modeling [11], and it provides a consistent baseline that is independent of the choice of the regression model. In particular, for this experiment, the baseline average monthly prediction of CPUE for a given cell in 2016 is the respective average monthly CPUE for that cell from 2015. Equation (5) illustrates the baseline prediction

$$\hat{y}_{c,p}^* = y_{c,p \downarrow 1} \quad (5)$$

where  $y_{c,p \downarrow 1}$  is the actual value for cell  $c$  at the same period moved one year backward. For instance, if  $p = June2016$  then  $p \downarrow 1 = June2015$ .

**Evaluation Metrics.** The metrics used for evaluation are as follows.

- Mean Absolute Error (MAE) is calculated for each period  $p$  (i.e. month) as the mean of absolute errors of the predicted average CPUE for all cells in that period. MAE for period  $p$  is shown in Eq. (6)

$$MAE_p = \frac{1}{|C_p|} \sum_{c \in C_p} |\hat{y}_{c,p} - y_{c,p}| \quad (6)$$

<sup>1</sup> As pointed out in Sect. 3.3 we consider CPUE only for the gear Rapido. Hence, instead of writing  $cpue(c, p, Rapido)$ , we simply use  $cpue(c, p)$ , omitting the gear name.

where  $C_p$  denotes the set of all cells with a CPUE value in period  $p$ ; and  $\hat{y}_{c,p}$  and  $y_{c,p}$  are respectively predicted and actual CPUE values for cell  $c$  and period  $p$ . MAE for the baseline prediction is calculated similarly and it is denoted by  $MAE^*$ .

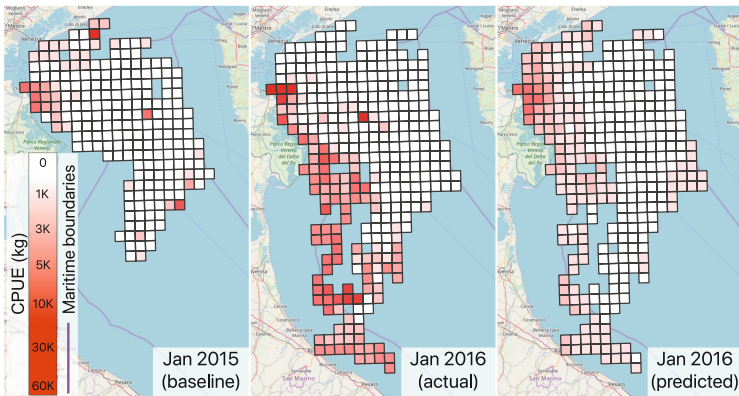
- Normalized Mean Absolute Error (nMAE) is calculated for each period as the MAE for that period divided by the mean of the actual CPUE for that period. nMAE for period  $p$  is shown in Eq. (7).

$$nMAE_p = \frac{MAE_p}{\mu_p}; \quad \mu_p = \frac{1}{|C_p|} \sum_{c \in C_p} y_{c,p} \quad (7)$$

- Relative Absolute Error (RAE) is a measure of model performance relative to the baseline model; it is calculated as the ratio of model MAE to the baseline MAE\* for a given period [1], as shown in Eq. (8).

$$RAE_p = \frac{MAE_p}{MAE_p^*} \quad (8)$$

Model evaluation metrics for uniform and weighted catch distributions are shown in Table 1. Since interpreting MAE requires information about the magnitude of CPUE, mean of CPUE for each period is also included. Due to the variation of CPUE means, MAEs for different periods cannot be directly compared. nMAE allows for direct comparison of the model performance for different periods, since it is normalized by the period mean. Similarly, RAE allows for comparison of model performance against the baseline for different periods. If RAE equals 1, the model is performing as well as the baseline for that period. RAE of less than 1 indicates that the model performs better than the baseline, and vice versa.



**Fig. 3.** CPUE over grid cells for January. Actual values for 2015 are the baseline for Jan. 2016 (left). Actual values for Jan. 2016 (middle) are used in the evaluation of values predicted by our model for Jan. 2016 (right).

**Results and Discussion.** The RAE values in Table 1 are less than 1 for every period, indicating that the RF model consistently performs better than the baseline. RAE on the last row for both models show a 13% improvement compared to baseline for the average results over all months. This naive model performance is largely due to the limited temporal extent of the data as described below, and also because the model is unaware of the spatial and temporal autocorrelation in the data. Incorporating the autocorrelations into the model is a subject for future work.

**Table 1.** Model evaluation metrics for monthly CPUE

	Month (2016)	MAE		CPUE mean (actual)	nMAE		RAE
		RF	baseline		RF	baseline	
Uniform catch distribution	1	2065.28	2473.68	2423.60	0.85	1.02	0.83
	2	1758.83	2473.26	1728.77	1.02	1.43	0.71
	3	812.98	881.96	983.87	0.83	0.90	0.92
	4	622.05	663.76	732.72	0.85	0.91	0.94
	5	862.11	948.41	936.90	0.92	1.01	0.91
	6	675.91	886.13	815.72	0.83	1.09	0.76
	7	2333.37	2377.42	2419.54	0.96	0.98	0.98
	8 <sup>†</sup>	na	na	na	na	na	na
	9	3078.92	3168.13	3379.25	0.91	0.94	0.97
	10	1430.51	1705.33	1733.98	0.82	0.98	0.84
	11	2101.59	2295.61	2372.60	0.89	0.97	0.92
	12	1113.45	1213.80	1237.16	0.90	0.98	0.92
	All	1490.18	1707.45	1658.45	0.90	1.03	0.87
Weighted catch distribution	1	1947.78	2188.87	2193.30	0.89	1.00	0.89
	2	1474.28	1963.45	1593.04	0.93	1.23	0.75
	3	658.65	740.02	782.88	0.84	0.95	0.89
	4	441.24	486.75	529.31	0.83	0.92	0.91
	5	580.32	643.32	641.55	0.90	1.00	0.90
	6	436.84	637.21	542.09	0.81	1.18	0.69
	7	1474.76	1491.21	1523.94	0.97	0.98	0.99
	8 <sup>†</sup>	na	na	na	na	na	na
	9	2239.63	2336.55	2461.78	0.91	0.95	0.96
	10	1124.27	1242.47	1312.44	0.86	0.95	0.90
	11	1297.63	1660.43	1526.47	0.85	1.09	0.78
	12	778.46	876.93	868.05	0.90	1.01	0.89
	All	1116.69	1290.72	1254.27	0.89	1.03	0.87

<sup>†</sup> No data available due to the fishing ban.

Figure 3 shows the baseline, actual, and predicted monthly CPUE for January. Presence of a grid cell on the map indicates that fishing activity occurred at least once during the whole month, and vice versa. Comparing the actual data for 2015 and 2016 in the figure, it is obvious that the fished area in January of 2016 is much larger than 2015. Since the model only uses the 2015 data for making predictions for 2016, it has no information about the areas that were not fished in that period of 2015. This presents a limitation which is imposed on the model due to the temporal restriction of the data. In other words, the model performance can be improved by having data for a longer period of time.

## 5 Conclusion and Future Work

In this paper, we explored a spatio-temporal dataset resulting from the fusion of the trajectories of the fishing vessels of the Northern Adriatic sea for 2015 and 2016, the landing report of the primary fish market of the area, and relevant environmental data. The landing reports represent quite a unique semantic feature to be associated with the fishing trajectory. Also, the utilization of environmental attributes influencing the species distribution (sea surface temperature, chlorophyll-a) and the fishermen behaviors (wave heights), represent an additional key element. Moreover, the utilization of the sea surface temperature can be helpful to evaluate the effect of climate changes on the fishing activities, a current and heavy issue needed to be addressed.

We applied to the dataset the Random Forest Machine Learning method, with the goal of predicting the CPUE indicator that could be helpful to improve the fisheries management plans for sustainable exploitation of the fishing resources. The forecast results—surpassing the baseline prediction by approximately 13%—indicate the value of the use of Machine Learning for this task, while clearly leaving a lot of room for improvement. Firstly, the task itself is difficult, and clearly a number of variables not currently captured—both environmental and latent (e.g., economic) factors, as well as fishermen behavior— influence the capabilities of the prediction model. On the one hand, this is due to the short temporal horizon of the landing and AIS data: two years, one for training and the other one for testing, are not sufficient. We expect that with access to 2017 data, the results will improve. On the other hand, other prediction techniques, such as the use of lag variables [21], or an alternative approach using modern time series prediction, should be exploited as we continue the project.

In data modeling, more sophisticated methods of catch distribution into grid cells, such as habitat selection [3], need to be looked at. A combination of both data modeling and machine learning extensions to this early research could also help to overcome the difficulties deriving from the short time period.

## References

1. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: empirical comparisons. *Int. J. Forecast.* **8**(1), 69–80 (1992)

2. Bogorny, V., Renso, C., de Aquino, A.R., de Lucca Siqueira, F., Alvares, L.O.: Constant—a conceptual data model for semantic trajectories of moving objects. *Trans. GIS* **18**(1), 66–88 (2014)
3. Bonanno, A., et al.: Habitat selection response of small pelagic fish in different environments. Two examples from the oligotrophic Mediterranean Sea. *PLoS ONE* **9**(7), e101498 (2014)
4. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
5. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
6. Brüggemann, S., Bereta, K., Xiao, G., Koubarakis, M.: Ontology-based data access for maritime security. In: Sack, H., Blomqvist, E., d’Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) *ESWC 2016. LNCS*, vol. 9678, pp. 741–757. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-34129-3\\_45](https://doi.org/10.1007/978-3-319-34129-3_45)
7. Copernicus: Europe’s eyes on Earth. <https://www.copernicus.eu/en>
8. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) *MCS 2000. LNCS*, vol. 1857, pp. 1–15. Springer, Heidelberg (2000). [https://doi.org/10.1007/3-540-45014-9\\_1](https://doi.org/10.1007/3-540-45014-9_1)
9. Estrada, J., Silva, C., Yáñez, E., Rodríguez, N., Pulido-Calvo, I.: Monthly catch forecasting of anchovy *Engraulis ringens* in the north area of Chile: non-linear univariate approach. *Fish. Res.* **86**(2), 188–200 (2007)
10. Ferrà, C., et al.: Mapping change in bottom trawling activity in the Mediterranean Sea through AIS data. *Mar. Policy* **94**, 275–281 (2018)
11. Hyndman, R.J., Athanasopoulos, G.: *Forecasting: Principles and Practice*, 2nd edn. OTexts, Melbourne (2018)
12. Maina, I., Kavadas, S., Somarakis, S., Tserpes, G., Stratis, G.: A methodological approach to identify fishing grounds: a case study on Greek trawlers. *Fish. Res.* **183**, 326–339 (2016)
13. Mello, R.d.S., et al.: MASTER: a multiple aspect view on trajectories. *Trans. GIS* (2019, to appear). <https://doi.org/10.1111/tgis.12526>
14. Mills, C.M., Townsend, S.E., Jennings, S., Eastwood, P.D., Houghton, C.A.: Estimating high resolution trawl fishing effort from satellite-based vessel monitoring system data. *ICES J. Mar. Sci.* **64**(2), 248–255 (2006)
15. Nurdin, S., Ahmad Mustapha, M., Lihan, T., Ghaffar, M.A.: Determination of potential fishing grounds of *Rastrelliger kanagurta* using satellite remote sensing and GIS technique. *Sains Malays.* **44**(2), 225–232 (2015)
16. Paige Eveson, J., Hobday, A., Hartog, J., Spillman, C., Rough, K.: Seasonal forecasting of tuna habitat in the Great Australian Bight. *Fish. Res.* **170**, 39–49 (2015)
17. Parent, C., et al.: Semantic trajectories modeling and analysis. *ACM Comput. Surv. (CSUR)* **45**(4), 42 (2013)
18. Soares, A., et al.: CRISIS: integrating AIS and ocean data streams using semantic web standards for event detection. In: *International Conference on Military Communications and Information Systems* (2019)
19. Soares Júnior, A., Moreno, B.N., Times, V.C., Matwin, S., Cabral, L.d.A.F.: GRASP-UTS: an algorithm for unsupervised trajectory segmentation. *Int. J. Geogr. Inf. Sci.* **29**(1), 46–68 (2015)
20. de Souza, E.N., Boerder, K., Matwin, S., Worm, B.: Improving fishing pattern detection from satellite AIS using data mining and machine learning. *PLoS ONE* **11**(7), e0158248 (2016)
21. Tyrallis, H., Papacharalampous, G.: Variable selection in time series forecasting using random forests. *Algorithms* **10**, 114 (2017)
22. Vespe, M., Gibin, M., Alessandrini, A., Natale, F., Mazzarella, F., Osio, G.C.: Mapping EU fishing activities using ship tracking data. *J. Maps* **12**, 520–525 (2016)



**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), 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 license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license 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.

