








A workflow for standardizing the analysis of highly resolved vessel tracking data

T. Mendo ^{1,*}, A. Mujal-Colilles ^{2,*}, J. Stounberg³, G. Glemarec ³, J. Egekvist ³,
E. Mugerza ⁴, M. Rufino ^{5,6}, R Swift⁷, M. James ⁷

¹School of Geography and Sustainable Development, University of St. Andrews, KY16 9AL St. Andrews, UK

²Barcelona School of Nautical Studies, Universitat Politècnica de Catalunya, 08003 Barcelona, Catalunya

³National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet, DK-2800 Kgs. Lyngby, Denmark

⁴AZTI, Sustainable Fisheries Management, Basque Research and Technology Alliance (BRTA), Txatxarramendi Ugarteia z/g, 48395 Sukarrieta, Bizkaia (Basque Country), Spain

⁵Portuguese Institute for the Sea and the Atmosphere (IPMA), Division of Modelling and Management of Fisheries Resources, Av. Dr. Alfredo Magalhães Ramalho, 6, 1495-165 Lisboa, Portugal

⁶Centre of Statistics and its Applications (CEAUL), Faculty of Sciences, University of Lisbon, 1649-004 Lisboa, Portugal

⁷Scottish Oceans Institute, University of St Andrews, East Sands Fife KY16 8LB, UK

*First authorship and author's correspondence shared between these authors. Tania.Mendo@st-andrews.ac.uk; anna.mujal@upc.edu

Abstract

Knowledge on the spatial and temporal distribution of the activities carried out in the marine environment is key to manage available space optimally. However, frequently, little or no information is available on the distribution of the largest users of the marine space, namely fishers. Tracking devices are being increasingly used to obtain highly resolved geospatial data of fishing activities, at intervals from seconds to minutes. However, to date no standardized method is used to process and analyse these data, making it difficult to replicate analysis. We develop a workflow to identify individual vessel trips and infer fishing activities from highly resolved geospatial data, which can be applied for large-scale fisheries, but also considers nuances encountered when working with small-scale fisheries. Recognizing the highly variable nature of activities conducted by different fleets, this workflow allows the user to choose a path that best aligns with the particularities in the fishery being analysed. A new method to identify anchoring sites for small-scale fisheries is also presented. The paper provides detailed code used in each step of the workflow both in R and Python language to widen the application of the workflow in the scientific and stakeholder communities and to encourage its improvement and refinement in the future.

Keywords: small-scale fisheries; geospatial data; fisheries management; marine spatial planning

Introduction

Ecosystem-based fisheries management has revealed the need for much more detailed spatial information about fish distribution and fishing effort at the vessel level, to enable the implementation of fine-scale spatial management (Wilén 2004, Stelzenmuller et al. 2008, Parnell et al. 2010). This need has become exacerbated in the last few decades due to the increased pressure from human activities in the marine environment. Indeed, within the 'blue economy' agenda, coastal and marine regions are seen as grounds for new economic opportunities, such as energy generation, mining, tourism, aquaculture, and fisheries, increasing the pressure exerted on the marine environment (Bennett et al. 2019). The designation of marine protected areas also represents a spatial constraint and the pressure to expand these designations is increasing.

Knowledge on the spatial and temporal distribution of the activities carried out in the marine environment is key to manage available space optimally. However, frequently, little or no information is available on the distribution of the largest user of the marine space, namely small-scale fishers (Trouillet 2019). Technological developments have enabled the collection of vessel's tracking data to represent fisher's activities in space. Among the systems used to track vessels, Vessel Monitoring Systems (VMS) are mandatory for all EU vessels

>12 m [EU Fisheries Control Regulation (EC 1224/2009)] and transmit the location of the vessel every 2 h. Automated Identification Systems (AIS), Electronic Monitoring (EM) systems, and other high-resolution vessel-tracking systems report the position of a vessel at intervals from seconds to minutes (Lee et al. 2010, Gerritsen and Lordan 2011, Burgos et al. 2013, Natale et al. 2015, James et al. 2018, Behivoke et al. 2021, Mujal-Colilles et al. 2022, Navarrete Forero et al. 2017). These higher-frequency data allow prediction of vessel activity with much higher precision than using VMS. This is especially important for fisheries that display complex fishing patterns (Muench et al. 2018), in areas where spatial constraints are dense or in small-scale fisheries (SSF), where fishing operations are relatively short in distance and/or duration (Katara and Silva 2017, Mendo et al. 2019a).

Tracking systems can produce a significant amount of data, creating challenges for data transmission, processing, and analysis. For VMS data, standardized methods and tools have been developed to process, analyse, and visualize vessel location (Hintzen et al. 2012, Russo et al. 2014). However, for highly resolved data, only *ad-hoc* tools (developed as needed) have been used thus far, making it difficult to replicate analyses. Moreover, most of the methods currently in use focus on large-scale fishing vessels (e.g. Global Fishing Watch), and

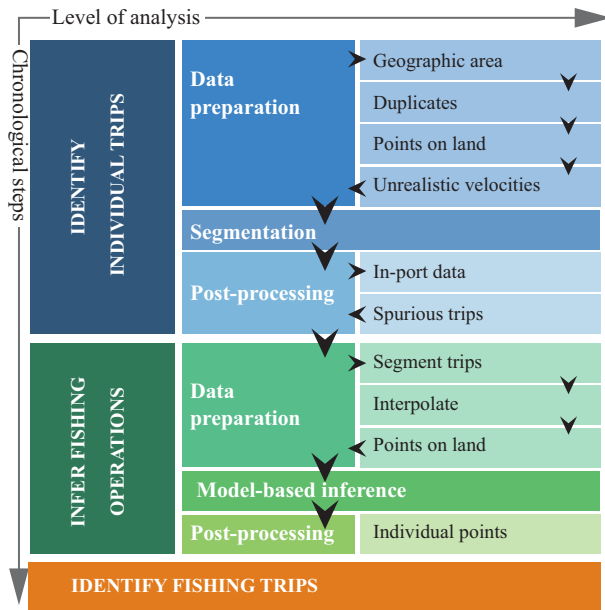


Figure 1 Workflow showing sequential steps to identify fishing trips.

cannot be used directly in SSF because these operate in very different ways. This poses further challenges to developing a standardized method for highly resolved geospatial data. This standardization is particularly relevant at present, as mandatory tracking of all fishing vessels has been implemented in several countries. For example, England and Wales are starting to roll out mandatory tracking for all fishing vessels (MMO 2022). In Portugal and Denmark, tracking of bivalve dredge fisheries is mandatory. Mandatory tracking of SSF is under active consideration in the European Union (EU) (European Commission 2018). The Trilogue (informal interinstitutional negotiation that brings together representatives of the EU Parliament, Council, and Commission) met in May 2023 and reached a provisional agreement for the mandatory implementation of EU-wide tracking for SSF vessels by all EU Member States (European Parliament 2023).

The aim of the current work is to describe a workflow to identify individual vessel trips and infer fishing activities from highly resolved geospatial data, focusing on SSF. Due to the highly variable nature of the trips conducted by SSF vessels, alternative approaches are proposed to allow adaptation to the variability of the fishing operations. This workflow offers flexibility for each user to parametrize the variables based on *a priori* knowledge of the fishery. This paper is organized as follows: First, we describe the workflow, including each step and the rationale behind it. Next, we illustrate the framework with specific case studies using different workflow pathways according to the fleet behaviour and data collected (and associated tracking systems, i.e. AIS and geospatial trackers). Finally, we discuss the results, contextualize the work, and identify future areas of research.

Materials and methods

The proposed workflow consists of a sequence of steps that allow the user to prepare vessel tracking data, differentiate individual vessel trips, infer fishing operations, and differentiate between non-fishing and fishing trips (Fig. 1). It was conceptu-

Table 1. Description of the user-input variables.

Parameter	Units	Definition
cum_dist	metres	Cumulative distance from trip extreme points
crs_wgs84	EPSG	EPSG code for WGS84 CRS
crs_utm	EPSG	EPSG code for UTM projected CRS
d_interp	metres	Minimum distance between points to interpolate
dist_travelled	kilometres	Minimum length of a trip
lat_max, lat_min	decimal degrees	Latitude boundaries
lon_max, lon_min	decimal degrees	Longitude boundaries
n_obs	number	Minimum observations for a valid trip
port_buffer	metres	Distance within the port where points will be discarded
speed_filter	knots	Maximum allowed speed for vessels
time_travelled	hour	Minimum duration of a trip
t_interp	minutes	Time interpolation between points

alized as a result of the experience gained by dealing with data obtained through several types of tracking devices, namely AIS, EM, and geospatial trackers to gather geospatial data in different fishing fleets. Each step will be described in detail in the following subsections.

Due to the huge volumes of data acquired when tracking vessels at high temporal frequencies, it is fundamental to develop a data management system capable of processing, storing, and sharing the data. PostgreSQL and its spatial extension (PostGIS) were selected due to their capacity to deal with any volume of data, and to share it with common software such as Excel, GIS, or R (Urbano and Cagnacci 2014). The code to process the data was developed both in R programming language and in Python, which are both free and open-source (R Core Team 2022, Python Software Foundation 2023). R is widely used in fisheries and ecological research and several R packages exist to manage and analyse spatial data. However, due to the large amounts of data generated from highly resolved vessel tracking devices, the code was also developed in Python, which is generally acknowledged to have faster computing times (Python Software Foundation 2023). The R and Python scripts are available at [Supplementary Materials S1](#) (for R) and [S2](#) (for Python).

Input data

Data should be provided in tabular format, with at least, the following columns: vessel ID, time stamp, and position. Position should be given as decimal latitude and longitude coordinates, using the world geodetic Coordinate Reference System (CRS) WGS84. The EPSG code for a projected CRS in Universal Transverse Meters (UTM), should also be provided, which will depend on the location of the user's data set.

Other input data needed area map of the coastline of the study area (a shapefile or a geopackage), also using CRS WGS84. The map resolution will be key to the computational speed of the data analysis; therefore, users will have to compromise between map resolution and computational resources available.

Input parameters are needed at different stages in the workflow (Table 1). They relate to information i.e. specific to each fishery and require expert knowledge on how the fishery

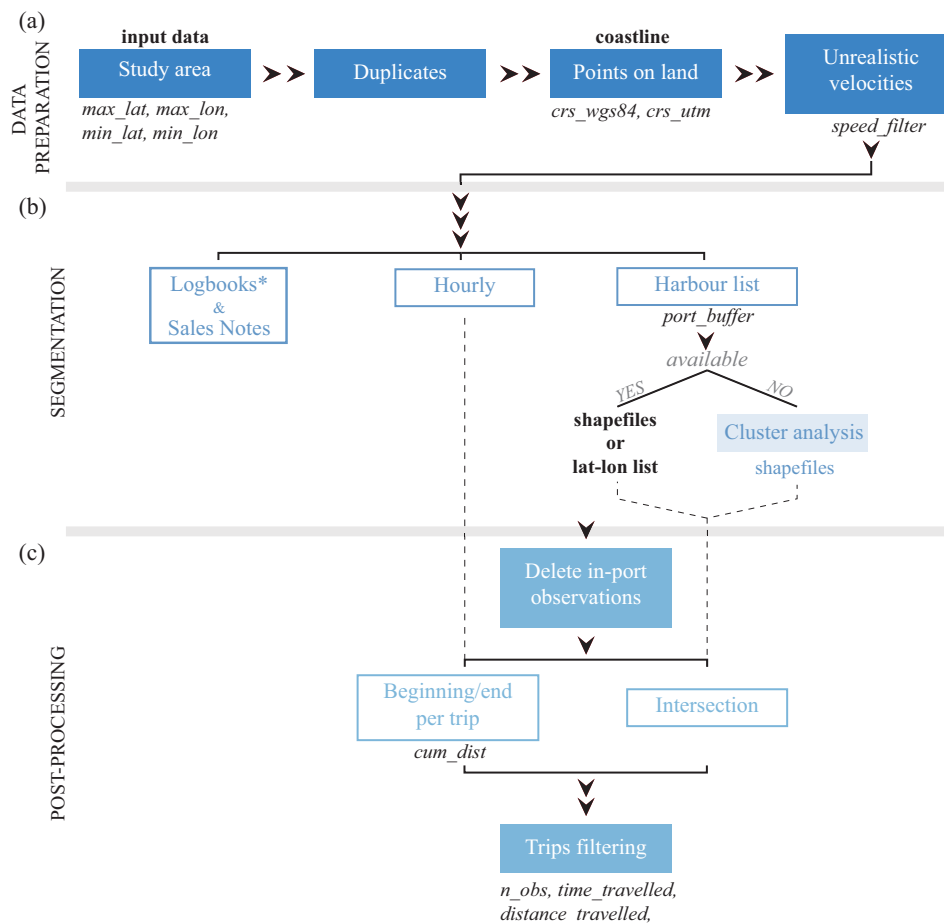


Figure 2 Workflow showing sequential steps used to identify individual trips. Italic text indicates user-required parameters; bold text indicates input data. * for a detailed workflow to identify fishing trips using logbook data, see Hintzen et al. (2012).

operates. The input parameters needed throughout the workflow are:

Identify individual trips

Trips are defined as the time-series of geospatial data between two positions, a point of departure and a point of return, which can be in the same geographical space or at different locations. This section of the methodology consists of segmenting or splitting the time series and identify whether the segments represent an individual trip (see [Supplementary Material](#), Section 1—Identify_individual_trips for corresponding R or Python codes). This is done by preparing the data first, then segmenting the time series, and finally post-processing the identified individual trips by removing spurious points and trips (Fig. 2). Each step is described in detail in the following sub-sections.

Data preparation

The first section of the workflow in Fig. 2a includes standard procedures to clean positional data: (i) defining the study areas and removing erroneous positions; (ii) removing duplicate records; (iii) removing points on land; and (iv) removing data points associated with unrealistic velocities. It is important to clarify here that the code presented in the supplementary materials deletes the data during these cleaning procedures but

saves the resulting data as a different file, thus preserving the raw data. It is highly recommended that a backup copy of the raw data in the database (e.g. PostgreSQL) is retained.

As a first step, erroneous lat-lon positions defined as unrealistic points recorded outside the operating area, depending on the device used to record the tracks, are identified and removed. None of the tracking systems used by the authors in this study (AIS, EM, and geospatial trackers) was entirely free of these errors. These erroneous positions stemmed from the occasional loss of satellite/receiver connection, or from the reconnection of the system after having been turned off, either automatically to maintain battery life, or manually if the fisher needed to do so. For example, AIS provides 0,0 lat-lon positions when the device has lost satellite signal. To clean this type of errors, a boundary box was specified *a priori* (maximum and minimum latitudes and longitudes) and positions outside this box were removed prior to further analyses.

The second step of the pre-processing section consisted of removing from the dataset duplicate points in time and space. Duplicate records are common in geospatial datasets, especially when using AIS data, as more than one land-based receiver may record the same position and timestamp.

In the third step, inland positions are removed by using a high-resolution coastline map. This map needs to be accurate enough to adequately represent fishing, harbours, anchorages,

bays, and fishing areas. The resolution of the map will determine the processing speed.

The fourth step consists of removing points associated with unrealistic speeds. This step implicitly includes the removal of spurious lat-lon points that were not detected in the previous steps. For example, some tracking devices create spurious points when the GPS signal is lost by duplicating the last lat-lon recorded point in a timestamp 0.1 s prior to the recovery of data. This creates a track segment with unrealistic velocities, larger than the maximum expected in the specific fishing fleet. The threshold used as the speed limit requires expert knowledge but can also be informed by visualizing a histogram of the estimated speed between consecutive positions.

Segmentation

Following the data preparation steps, the dataset will consist of a time series of spatial positions for each vessel, which needs to be further segmented into individual vessel trips. A fishing trip is further defined as any vessel trip during which at least one fishing activity occurs (e.g. setting or hauling, trawling, and dredging). We recognize three approaches to identify trips, depending on the information available for each fishing fleet.

(i) Based on logbooks or sale note information

If available, logbooks or sales notes provide additional (non-geospatial) information on the vessels' fishing activity that can be used to discriminate fishing from non-fishing trips, e.g. using Hintzen et al. (2012) (Fig. 2b).

(ii) Typical fishing activity of the fleet (hourly method)

If the fishery operates with a distinct pattern (e.g. only during daylight hours, as is the case for many SSF), we can assume that each trip starts with each new day, so that midnight can be used to segment data into individual trips. Another alternative would be to use the duration of the 'resting period' between fishing trips to segment the data (Rufino et al. 2023).

(iii) Based on harbours or anchorage sites

Another approach to define the start/end of a vessel trip is by identifying points lying within ports or harbours, mooring sites, or, for small-scale fishing vessels, a mooring/anchorage site on a bay, a beach, or a small pier (Fig. 2b). Here, we describe the latter method, which we developed specifically for this paper.

If there is a known list of harbours, moorings, or anchorage sites (hereafter referred to as anchorage sites) coordinates, a spatial buffer can be created around each of these sites to help identify when a vessel is leaving/entering a harbour and to segment the data into individual trips. For some fisheries, vessels can start trips from exclusive anchorage sites that might or might not be associated with a particular jetty or harbour infrastructure. For example, in many small-scale fisheries, vessels are often launched from the beach. This can also be the case in some countries with a large number of unreported/unknown anchorage sites. When a complete list of latitudes and longitudes for these jetties, mooring, or anchorage sites is not available, the list can be inferred from vessel activity data using a cluster analysis (see [Supplementary Material—Section 4—potential anchorage sites](#)). The goal of the cluster analysis is to obtain a list of the most likely anchorage sites regardless of whether there is infrastructure associated to them.

The workflow designed for this task must be applied for each vessel separately (Fig. 3). Ideally, a time series of data

Single Vessel

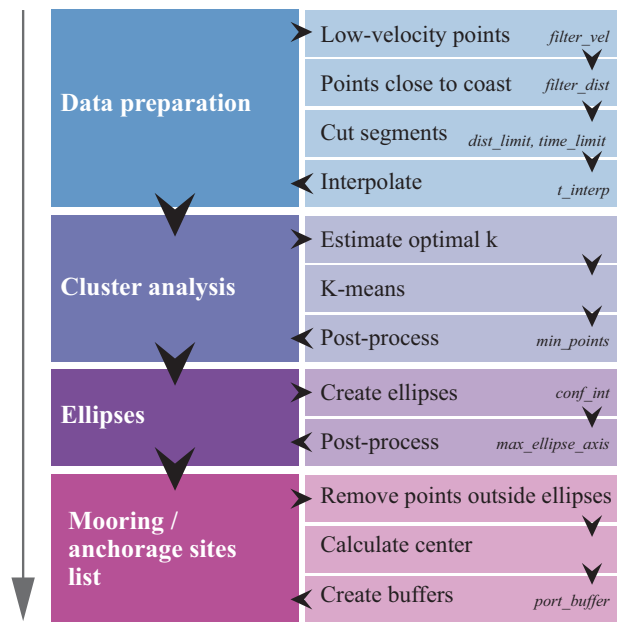


Figure 3 Workflow to identify anchorage sites. Italics show the name of the user-input parameters in each step.

is used as input that can capture all different anchorage sites used during a representative period (e.g. for the Scottish fleet example described below, we used 1 year of data for each vessel). Once the script has been run individually for each vessel, the polygons obtained from this procedure should be joined to define each anchorage site and thus to avoid redundant information.

In order to prepare the data for cluster analysis, the data have to be cleaned by going through the general data preparation procedure in Fig. 2. The data are then filtered using two different criteria: speed (*filter_vel*) and land proximity (*filter_dist*), as seen in Fig. 3. We would not expect vessels to be moving fast when in port or at an anchorage site located close to the coastline. For the devices reporting position with a lower frequency while in port (when the vessel is not moving or reduces speed), an interpolation (up sampling) procedure to a time interval defined by the user (*t_interp*) is required at this stage to augment the number of positions in the areas close to shore. To avoid unrealistic positions that would result from the interpolation process, the interpolation is conducted in groups of data. These groups are based on a maximum distance between observations and maximum time difference between consecutive observations (*dist_limit*, *time_limit*). For devices that gather large amounts of data while in port, e.g. AIS, then the interpolation is used to down sample the data. The main goal of this interpolation phase is to improve the clustering accuracy on identifying and classifying the clusters (if up sampling) or to reduce computing times (if down sampling).

Once the data are prepared, we propose using an automatic method, such as the *Silhouette* method (Rousseeuw 1987) to determine the optimal number of clusters (k) representing the anchorage sites in the data. It is important to bear in mind that in cluster analyses, a minimum of two clusters are set by default. Once the optimal k is inferred, the k -means method (Forgy 1965) is applied to assign the most likely cluster to

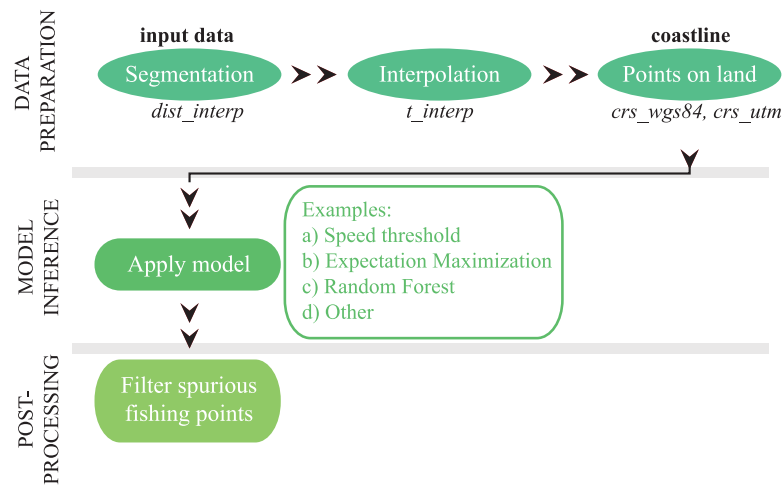


Figure 4 Workflow for inferring fishing operations from highly resolved data. Italic text indicates user-required parameters; bold text indicates input data needed.

each position. A post-processing step eliminates clusters with less than three points (*min_points*).

The following step in Fig. 3 (Cluster Analysis > Post-Process) aims at improving the location of the centroid of each cluster and to detect spurious clusters created by real fishing activity located very close to the coast. Thus, a confidence region is created for each cluster by drawing a covariance error ellipse, given a pre-defined confidence interval. The ellipse axis length and orientation depend on the selected confidence interval and the covariance matrix (eigenvalues and eigenvectors, respectively) (Draper and Smith 1998). In the present examples (Supplementary Material S2), we used a confidence interval of 95%. Once ellipses are created for each vessel, the sizes of the ellipses will be used to remove specific clusters; for vessels that conduct their fishing operations close to land, a cluster might be defined for an area that does not represent an anchorage site. Clusters associated with this activity usually have been found with our examples to have a large spatial extension. A threshold size for a potential anchorage site (*max_ellipse_axis*) should thus be defined using expert knowledge.

For the remaining ellipses, the points outside the ellipses are considered outliers, and the mean latitudes and longitudes of the positions inside the ellipses are calculated to estimate the location of the anchorage sites. A buffer (*port_buffer*) is set to these mean positions to produce a shapefile of polygons for each vessel. A unique shapefile is created for each vessel, but as different vessels can share anchorage sites, it is recommended to find the union between polygons, when overlap occurs, to create a final shapefile with the complete list of anchorage sites in order to speed up the process. Visualization of results is highly recommended at this stage and is included in the R and Python codes.

Post-processing

Depending on the tracking device you are using, huge amounts of data may be generated when a vessel is in port or at anchor. To reduce the amount of data for later analysis, especially before inferring fishing activities, it makes computational sense to remove these observations.

Following the workflow in Fig. 2, and depending on the pathway taken for segmentation (Fig. 2b), there are two

options to remove points close or in port: for trips identified based on daily activities, we suggest removing data points located within a specified distance from the coastline (*cum_dist*) at the beginning and at the end of the trip. The value of *cum_dist* can be set with prior knowledge on how far the fishery usually operates from the coastline and aided by visual exploration of the positional data. This approach is recommended, especially if at least some fishing activities can occur very close to the shore, where applying a spatial buffer around the coastline might delete information on fishing operations. When a list of ports or mooring sites is available (either from a pre-existing list or a list generated with the code available at Supplementary Material—Section 4—potential anchorage sites), positions inside the spatial buffer created around ports (*port_buffer*) can be deleted.

A final step in the workflow in Fig. 2c is to remove unrealistic trips. Here we apply some summary statistics about the trips, and set sensible criteria based on expert knowledge about what constitutes a realistic trip. For example, a minimum distance travelled (*dist_travelled*), a minimum trip duration (*time_travelled*), or a minimum amount of positional data (*n_obs*) can serve as suitable criteria to define plausible trips. A conservative value if *n_obs* is recommended, and the value might be informed by plotting a histogram of the number of observations per trip.

Infer fishing operations

Once the trips conducted by each fishing vessel have been identified, we proceed to infer where and when they are engaging in a fishing operation using the workflow to infer fishing operations (Fig. 4; see Supplementary Material, Section 2—Infer_fishing_operations for corresponding R or Python codes). The European Commission defines a fishing operation as ‘all activities in connection with searching for fish, the shooting, towing, and hauling of active gears, setting, soaking, removing, or resetting of passive gears and the removal of any catch from the gear, keep nets, or from a transport cage to fattening and farming cages’ (European Commission 2011).

Data preparation

Highly resolved geospatial data can occasionally contain temporal gaps (e.g. when the AIS units are turned off when

fishing). Therefore, before interpolating the tracks into a regular temporal frequency for later analysis, it is recommended to divide the points into segments within individual trips based on the distance between consecutive points. This is done to avoid the interpolation of data between two consecutive points with a large spatial gap between them that might lead to erroneous artificial positions classified as part of a fishing operation. Here it is important to note that in some cases, depending on the research question, it might be useful to further look into these gaps, e.g. as an approach to investigate illegal, unregulated, and unreported fishing activities (Welch et al. 2022). Once the points in each trip are grouped, positions are interpolated within groups to a temporal frequency defined by the researcher (t_{interp}), which will depend on the temporal resolution of the data and the known mean duration of a fishing operation in the fishery under scrutiny. For example, in Scottish pots and traps fisheries, the optimal temporal frequency to infer hauling events is 60 s (Mendo et al. 2019a), whereas for bivalve dredges and octopus traps in Portugal, it is <2 min (Rufino et al. 2023). Several approaches can be used to interpolate the data, ranging from linear interpolation to more powerful methods (e.g. Hintzen et al. 2010, Russo et al. 2011). If some of the positions were located close to shore, the interpolation might result in positions on land and these positions must be removed in a similar manner to that described above.

Model based inference

Once the data has been pre-processed, as shown in Fig. 4, the user can decide which method to use to infer when fishing operations are taking place. There are several options available to infer when fishing operations are taking place, such as using a speed threshold, using statistical methods, such as expectation maximization algorithm or hidden Markov models (Mendo et al. 2019b, Rufino et al. 2023), or machine learning algorithms, such as a random forest model (Rodriguez 2023, Rufino et al. 2023). These sources all provide R code to perform these analyses. One limitation worth noting here, is that most of these approaches are tailored to single-gear fishing trips, and multi-gear trips might be more challenging when trying to infer fishing activities adequately. More work is needed to first infer gear used during each trip, and then applying the most suitable model to infer fishing.

Post-processing

After fishing activities have been inferred, a final check (Fig. 4) removes potential spurious fishing locations. This step depends on the duration of the fishing activity and the gear used, and expert knowledge. As an example, isolated records or segments of the trip that are too short in duration to be associated with fishing activities should be reclassified as non-fishing activities. After this step, spatial distribution of fishing activities can be mapped.

Identify fishing trips

Once fishing activities are inferred, only the trips where fishing activities are identified will be categorized as fishing trips (see Supplementary Material, Section 3—Identify_fishing_trips for corresponding R or Python codes).

Table 2. Number of geospatial data points removed in each step during the data preparation section for each case study and percentages shown in brackets.

Case study	Pots—Scotland	Gillnet fishery—Denmark
Initial number of points	253 872	3 046 799
Geographic boundaries	0 (0%)	380 (0%)
Duplicates removal	0 (0%)	0 (0%)
Points on land removal	795 (0.31%)	599 689 (19.6%)
Unrealistic speeds	3201 (1.27%)	487 (0.01%)

Results

The general workflow described in Fig. 1 has been applied to two case studies: the Scottish pots fishery targeting European lobster (*Homarus gammarus*) and brown crab (*Cancer pagurus*), and the Danish gillnet fishery targeting various demersal fish, such as European plaice, Atlantic cod, and lumpsucker (respectively, *Pleuronectes platessa*, *Gadus morhua*, and *Cyclopterus lumpus*), depending on season a geographic location. Details on each of these fisheries and how they operate can be found in Supplementary Material S3. These datasets only comprise a subset of these fisheries' fishing activities, but they can be used to illustrate the different types of issues most commonly encountered with these data, and how each approach described above can be applied. These case studies include different tracking devices used to gather geospatial data (namely, Teltonika trackers and AIS), collecting data at different temporal resolutions.

Identify individual trips

Data preparation

In Table 2, the number of positions remaining after each data preparation step in Fig. 2a is listed. For the Danish gillnet fishery (AIS data), the step 'points on land' removes a significant number of positions. A visual inspection of the points removed confirmed that these points were located within the harbour and inside the coastline, which is to say they are also harbour points. This relates to the precision of both the shapefiles and the geospatial points.

Segmentation and post-processing

In this section, we present the results obtained after segmenting and post-processing the time-series data to define the trips. Table 3 compares the real number of trips conducted against the estimated number of trips resulting when applying the two different methods. Real number of trips in the Scottish fishing fleet were verified by individual visual inspection of the data. In the case of the Danish fishing fleet the real number of trips were verified by going through the same workflow, but using the official harbour polygons to define trips.

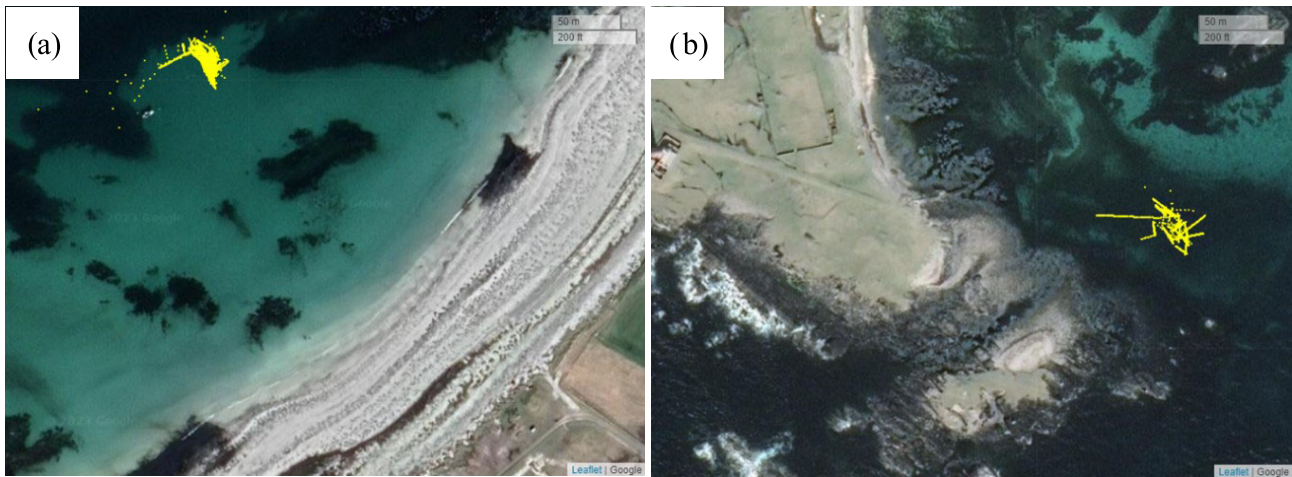
(i) Based on daily activities

Data obtained from the Scottish fleet of pots and traps were used for this example. This is a fleet that mainly operates during the daytime (James et al. 2018). Segmenting the trips at midnight resulted in 193 potential trips. During the post-processing step (Fig. 2c), this number was reduced to 129 trips first by removing the beginning and end parts of each trip with the points are located close to the coastline and then reduced to 122 by removing all trips with less than a threshold

Table 3. Upper rows: comparison of the number of trips resulting from applying the daily activities method and cluster analysis.

	Daily activities Scottish fishing fleet	Cluster analysis	
		Scottish fishing fleet	Danish fishing fleet
Number of vessels	8	8	10
Real number of trips	123	123	761
Estimated number of trips	122	126	765
False positives	0	3	4
False negatives	1	0	0
Number of points after step 3.1.1	249 876	249 876	2 446 234
In-port points removal	2 590 (1.0%)	2 364 (0.9%)	1 898 312 (77.6%)
Spurious trips removal	872 (0.4%)	2 050 (0.8%)	6 658 (0.2%)
Final number of points	246 414	245 462	554 580

Lower rows: number of geospatial data points removed in each step during the segmentation and post-processing section for each case study—percentages shown in brackets.

**Figure 5** Clustering detection of anchorage sites in (a) a beach and (b) an island with a lighthouse.

number of points ($n_{obs} = 50$) and those that were too short ($dist_{travelled} \leq 5 \text{ km}$ and $time_{travelled} \leq 1 \text{ h}$). Given the characteristics of the tracking device in this case study (set to only capture position every hour when no acceleration was detected), most of the trips deleted in the post-processing step, represented data points associated to positional data when in port or anchorage sites.

Of the 122 fishing trips estimated in the procedure, one was a false negative, i.e. two individual trips were wrongly identified as a unique trip. This situation occurred for a vessel that conducted a trip between two different ports in the same day, but only registered this as an individual trip.

(ii) Based on harbours or anchorage sites

In the case of the Scottish pots and traps fishery, the available list of ports did not include all the anchorage sites. In fact, this specific fishery often uses anchoring sites on beaches or private jetties. Therefore, we used the clustering method summarized in Fig. 3 to detect the potential anchorage sites, along with harbour infrastructures in the area. Year of data from eight vessels was used to capture potential anchorage sites.

For the Scottish fishing fleet, interpolation between low-velocity points was needed to increase the accuracy of the k -means method when clustering. Interestingly, results showed unknown anchoring sites in unexpected places. Figure 5a shows a clustering of positions identifying a mooring area near a beach, as well as an anchoring site near an island with a lighthouse (Fig. 5b) where some fishers frequently spend the

night, which was validated with interviews during field trips. Eleven anchorage sites were identified for the 10 vessels used in the case study. Most vessels (six) used between two and three anchorage sites per year, one vessel used a single one, and another, five.

The different post-processing steps (Fig. 2, delete in port observations, minimum number of points per trip, minimum distance travelled, and minimum time spent per trip) reduce the number of estimated trips by 70% in this case study.

The main problem using a list of anchorage sites with a buffer around them is that it can result in extra trips (false positives). As Table 3 shows, this method generated three false positives, which accounts for two vessels fishing nearby one of the anchorage sites. Figure 6a shows an example where one of the vessels left port to go east and passed by the harbour buffer area in the middle of the trip without actually docking. This results in two trips being identified with this method, when in fact only one trip was conducted. Still, the number of false positives represented <5% of the real cases.

From AIS data for 10 Danish gillnetters throughout 1 year, the clustering method was used to identify the harbours, and compared to a verified and updated harbour list. From the position data of the 10 vessels, 15 anchorage sites were identified, all within verified harbours. A 400-m buffer around the port was used, and as can be seen when more vessels use the same harbour, the different polygons were merged to get an equivalent of the complete harbour polygon (Fig. 7). The four false positives resulted from movement out and back into the

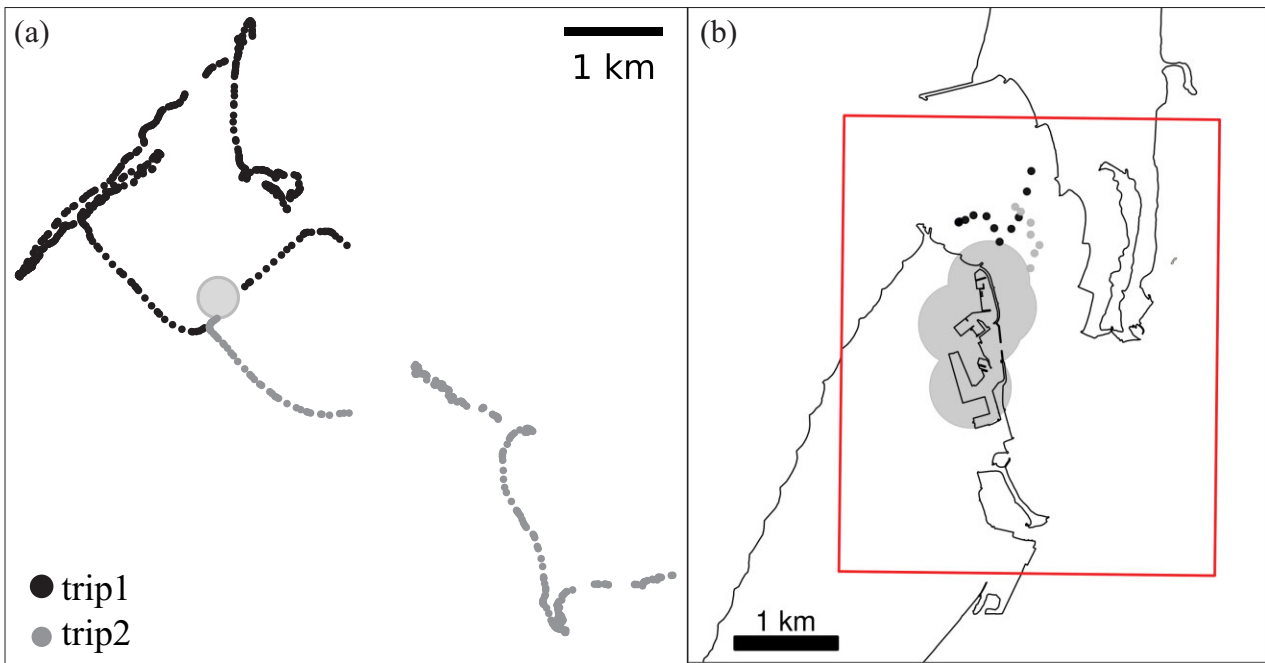


Figure 6 Example of a false positive when using the clustering method to identify anchorage sites (grey circles). Coloured dots represent the segmentation of a real single trip into two different trips. (a): Scottish case study, land not shown due to confidentiality issues; (b): Danish case study, red square is the original shapefile for the port.

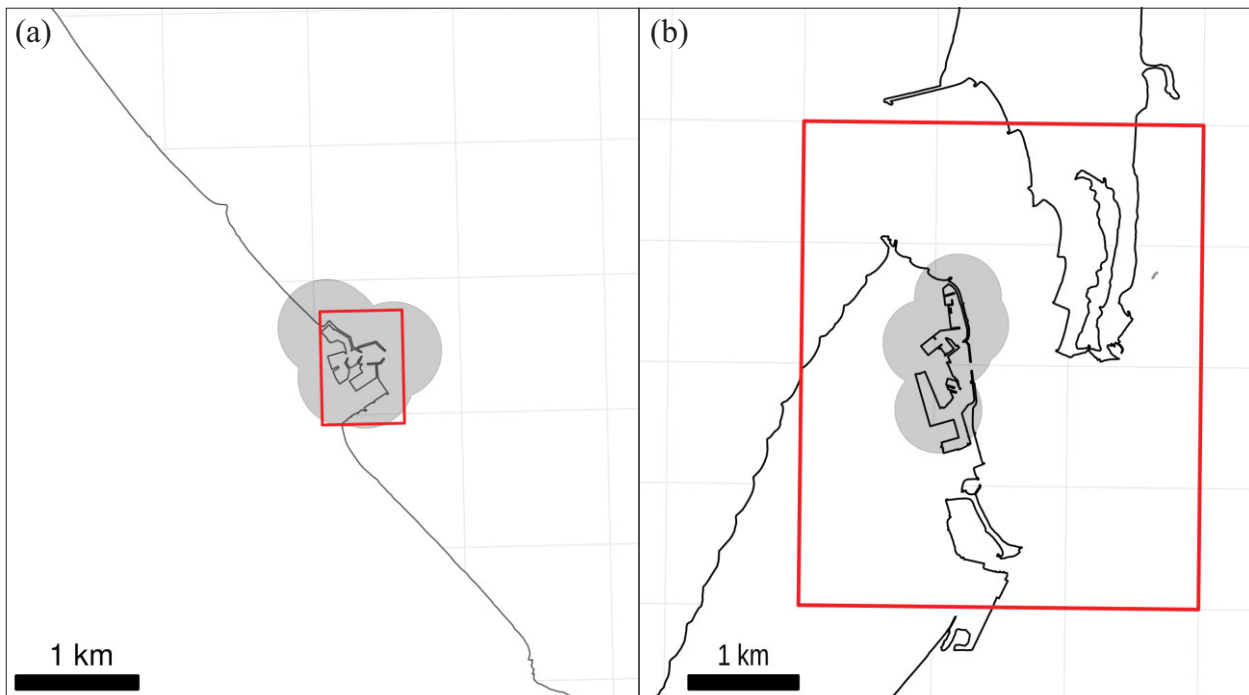


Figure 7 Polygons showing inferred anchorage sites from cluster analysis (grey circles) and official verified harbour polygons (red polygons).

inferred harbour polygon resulting in two unique trips (Fig. 6b).

Infer fishing operations

Once the trips are identified, the next step is to discriminate whether these trips are including at least one fishing operation. To do so, speed thresholds, statistical models, or machine

learning algorithms can be used to identify the points corresponding to a fishing activity within each trip. Data must be prepared to minimize potential errors in inferring fishing activities (Fig. 4).

For the Scottish pot fishery, the data were prepared to minimize errors, before applying the preferred model. Thus, trips were segmented when there were gaps larger than a predefined value (*dist_interp*) to avoid unrealistic points being incorrectly

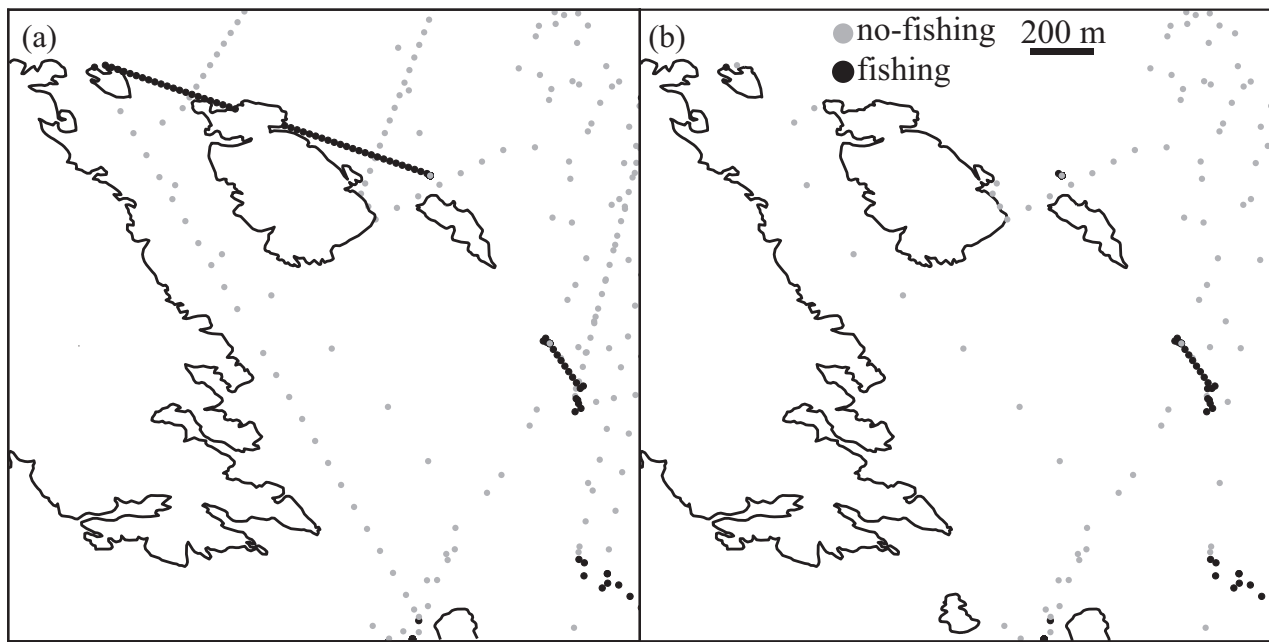


Figure 8 Results obtained to infer fishing operations using a speed threshold. (a) Interpolation between points with no constraints on the distances between interpolated points; (b) Interpolation constrained to be within segments of points previously ‘cut’.

classified by models, as seen in Fig. 8. This segmentation is particularly useful when a loss of data is evident. For the Scottish pot fishery, it was decided that trips would be segmented if the distance between consecutive observations was >500 m. This distance was considered to capture a definite loss of data. After the segmentation, data were interpolated to regular 1-min intervals (Mendo et al. 2019a). Resulting interpolated points falling on land were removed afterwards.

A speed filter was applied to infer fishing operations with a final check to remove positions erroneously classified as a fishing operation. For example, if the previous two positions and subsequent two positions were classified as non-fishing, then the one point in between would be corrected to non-fishing, and vice versa. This was considered a cautious filter, as the operation of hauling pots generally lasts longer than 1 min. A comparison between the results of the method using segmentations vs. without segmentation is shown in Fig. 8.

For Danish gillnetters, once a trip has been found and all the points in port have been deleted, a speed filter is used to determine fishing activity; for this particular fishery, the thresholds used are from 0 to 4 knots (from expert knowledge).

Identify fishing trips

In this section, only trips where a fishing operation was identified are retained as a fishing trip.

Discussion

Highly resolved geo-spatial data are proving useful to assess fine-scale distribution of fishing activities and fishing effort (McCauley et al. 2016, Le Guyader et al. 2017, Behivoke et al. 2021). This information is essential both to evaluate the potential impacts of fishing activities on particular habitats (Russo et al. 2016, Ferrà et al. 2018) but also to appropriately inform marine spatial planning, at scales relevant to fisher’s

activities (James et al. 2018, Metcalfe et al. 2018, Stelzenmüller et al. 2022). While many methods have been used to analyse these data, to date, these methods are mainly used on an ad hoc basis, and do not provide enough detail on the pre-processing steps carried out before the modelling approaches, which limits the ability of researchers to replicate analyses. The present work introduces a workflow to standardize the analysis of highly resolved geo-spatial data, specifically focusing on SSF. The development of this framework is particularly timely, as the number of small-scale, inshore fishing fleets being systematically tracked are increasing, and countries in Europe are beginning or are already tracking vessels as part of fleet-wide or nation-wide vessel tracking programmes for SSF (Burgos et al. 2013, MMO 2022). Importantly, the establishment of an EU-wide monitoring programme for vessels under 12 m in length has been provisionally agreed by representatives of the EU Council, Parliament, and Commission (European Parliament 2023).

Our current framework builds on and expands on seminal work done by Hintzen et al. (2012), and is the outcome of the knowledge and experience gained by the authors over several years working with different kinds of tracking devices (AIS, EM, geositional trackers, and VMS) that present different types of issues (e.g. spurious vessel locations, data gaps) in a variety of fishing fleets (e.g. with different fishing behaviours or levels of information). The framework presented here, therefore, allows the user to choose the path that better aligns with the peculiarities in the fishery being analysed. For example, if fishers only fish during daylight hours, then a segmentation of the data into trips is recommended using the hourly method. If trips can be overnight, then a list of ports can be used to segment data into trips, or if the list of ports is unknown or incomplete, a cluster analysis is proposed. Although the workflow presented can be adapted to different fisheries, there are common points that need to be followed regardless of the fishery (e.g. cleaning data, segmentation, and

interpolation). The precision of the coastline shapefile will be key in the computational speed of the whole process and all the data needs to be prepared and post-processed based on expert knowledge.

The development of the procedure using cluster analysis to identify anchorage areas can be very useful for SSF, particularly in areas where the vessels can use alternative launch sites, other than ports or jetties. For example, in Scotland, 3 out of 11 anchorage sites detected using this procedure lacked had no physical infrastructure, and were not listed in the official list of ports in Scotland (Scottish Government 2023). The clustering method proved to be reliable although it is important to get expert input to validate the anchorage sites resulting from this analysis, e.g. using interactive maps (e.g. Google Earth). Some inferred sites might have to be manually removed after this validation procedure. With this method, we saw improvements in defining trips, as seen, e.g. with the vessel that visited the lighthouse (Fig. 5b), as this anchorage site was not included in the official list of ports. When implementing the clustering method on the AIS data from all the gillnetters in this study, there was a very good agreement with the Danish official port list checked and used in the Danish ministries. Although the centroids of all the inferred anchorage sites were within the established harbours, the size of the buffer played a big role in how well the given harbour was represented. In some cases, the buffer of 400 m seemed too small, but in other cases, this buffer spilled out of harbour polygon. It is possible that if the clustering method were to be run for more vessels, the estimated harbour polygons would resemble more the established harbour polygons, which could prove to be a better approach than increasing the buffer size, as a polygon that stretches too far out could result in potential false positives.

Using a buffer area around ports (either from a known list of ports or resulting from cluster analysis) resulted in a greater loss of geospatial data than the daily activity method (which removes points at the beginning and end of a trip). Buffers around ports do not take into consideration the time sequence of events, and can remove occasional points that might occur during a trip, but near the port (e.g. when fishers fish very close to land). The size of the buffer should therefore be carefully assessed in each case.

Given the large amounts of high-resolution data generated with vessel tracking systems, we decided to make the code available in two commonly used programming languages: R and Python. Computationally, R and Python performed similarly; however, Python tended to be quicker when performing intersections between polygons (e.g. coastline) and data points. While RStudio allows for Python code to be run within R, the time required to load data rendered this impractical. R performed faster when transforming between coordinate reference systems. Overall, there was no difference between the two platforms with respect to computational speed. Although Python is gaining much interest within the scientific community, R is currently the most common programming language used by researchers in fisheries and marine research institutes. Both Python and R are freely available and have a growing list of libraries and packages to analyse and map spatial data. The tidyverse package was used in R as it provides a suite of libraries for cleaning, transforming, visualizing, and modelling data (Wickham *et al.* 2019). For both languages, there are many different approaches that produce the same result, with different computation time, thus both scripts can be improved in the future.

With an increasing focus on small-scale fisheries impacts on ecosystems and a general increase in the level of control and information required to conform to EU regulations (European Commission 2009), the workflow presented in this paper to process and interpret geospatial data, provides a robust method to estimate fishing effort in SSF fleets using the available geospatial data (although see Mendo *et al.* 2023 for a detailed method to estimate effort for passive gears). Having the possibility to dynamically map where the fishing fleet is anchored permits the segmentation of the fishing activity of the fleet into individual (fishing) trips, which in turn allows for deriving important control parameters such as days at sea, fishing days, and fishing hours. Not only does this sort of information help assessing fisheries impact at a fine scale, it can also provide evidence for fishers that a fishing ground may conflict with other existing or planned marine activities and spatial constraints by documenting both the spatial and temporal extent of the areas essential to fishing in terms of target species and income. We hope that this workflow is refined and improved in the future by the scientific and stakeholder communities.

Acknowledgements

The authors would like to acknowledge discussions had with colleagues during two workshops on Geospatial Data for Small-Scale Fisheries (WKSSFGE0 and WKSSFGE02), organized by the International Council for Exploration of the Sea (ICES, Working Group on Spatial Fisheries Data). Also, all fishers that volunteered to carry vessel tracking devices and share with us their time and knowledge. Also, thank you to the two anonymous reviewers for providing encouraging comments and great suggestions. We would like to thank the comments and suggestions from two anonymous reviewers which greatly improved the manuscript.

Supplementary data

Supplementary data is available at the *ICES Journal of Marine Science* online.

Conflict of interest: The authors declare no conflict of interest.

Author contributions

All authors contributed to designing the study. T.M., A.M.C., and J.S. analysed the data. T.M. and A.M.C. wrote the first draft of the paper, and all authors contributed to reviewing, and editing of subsequent drafts.

Funding

TM and MJ acknowledge financial support provided by the “Conserving Atlantic Biodiversity by Supporting Innovative Small-scale Fisheries Co-management” (CABFISHMAN) Project. This project is co-financed by the Interreg Atlantic Area Programme through the European Regional Development Fund. Project N^o: EAPA_134/2018”. AMC acknowledges the Spanish Ministry of Science and Innovation and the Serra Hunter programme from the Generalitat de Catalunya.

Data availability

The data underlying this article cannot be shared publicly to maintain vessel anonymity.

References

- Behivoke F, Etienne M-P, Guitton J *et al.* Estimating fishing effort in small-scale fisheries using GPS tracking data and random forests. *Ecol Indic* 2021;123:107321. <https://doi.org/10.1016/j.ecolind.2020.107321>
- Bennett NJ, Cisneros-Montemayor AM, Blythe J *et al.* Towards a sustainable and equitable blue economy. *Nat Sustain* 2019;2:991–3. <https://doi.org/10.1038/s41893-019-0404-1>
- Burgos C, Gil J, del Olmo LA. The Spanish blackspot seabream (*Pagellus bogaraveo*) fishery in the Strait of Gibraltar: spatial distribution and fishing effort derived from a small-scale GPRS/GSM based fisheries vessel monitoring system. *Aquat Living Resour* 2013;26:399–407. <https://doi.org/10.1051/alr/2013068>
- Draper NR, Smith H. *Applied Regression Analysis*. Ontario: Wiley, 1998.
- European Commission. Council regulation (EC) No 1224/2009 of 20 November 2009 establishing a Community control system for ensuring compliance with the rules of the common fisheries policy, amending Regulations (EC) No 847/96, (EC) No 2371/2002, (EC) No 811/2004, (EC) No 768/2005, (EC) No 2115/2005, (EC) No 2166/2005, (EC) No 388/2006, (EC) No 509/2007, (EC) No 676/2007, (EC) No 1098/2007, (EC) No 1300/2008, (EC) No 1342/2008 and repealing Regulations (EEC) No 2847/93, (EC) No 1627/94 and (EC) No 1966/2006. Brussels: European Commission, 2009.
- European Commission. Regulation of the European Parliament and of the council amending Council Regulation (EC) No 1224/2009, and amending Council Regulations (EC) No 768/2005, (EC) No 1967/2006, (EC) No 1005/2008, and Regulation (EU) No 2016/1139 of the European Parliament and of the Council as regards fisheries control. Brussels: European Commission, 2018.
- European Commission. *Commission Implementing Regulation (EU) No 404/2011 of 8 April 2011 laying down detailed rules for the implementation of Council Regulation (EC) No 1224/2009 establishing a Community control system for ensuring compliance with the rules of the Common Fisheries Policy*. Brussels: European Commission, 2011. http://data.europa.eu/eli/reg_impl/2011/404/oj/eng
- European Parliament. *Small-Scale Fisheries Situation in the EU and Future Perspectives*. Dublin: European Parliament, 2023.
- Ferrà C, Tassetti AN, Grati F *et al.* Mapping change in bottom trawling activity in the Mediterranean Sea through AIS data. *Mar Policy* 2018;94:275–81. <https://doi.org/10.1016/j.marpol.2017.12.013>
- Forgy EW. Cluster analysis of multivariate data: efficiency versus inter-pretability of classifications. *Biometrics* 1965;21:768–80.
- Gerritsen H, Lordan C. Integrating vessel monitoring systems (VMS) data with daily catch data from logbooks to explore the spatial distribution of catch and effort at high resolution. *ICES J Mar Sci* 2011;68:245–52. <https://doi.org/10.1093/icesjms/fsq137>
- Hintzen NT, Bastardie F, Beare D *et al.* VMStools: open-source software for the processing, analysis and visualisation of fisheries log-book and VMS data. *Fish Res* 2012;115–116: 31–43. <https://doi.org/10.1016/j.fishres.2011.11.007>
- Hintzen NT, Piet GJ, Brunel T. Improved estimation of trawling tracks using cubic Hermite spline interpolation of position registration data. *Fish Res* 2010;101:108–15. <https://doi.org/10.1016/j.fishres.2009.09.014>
- Scottish Government. *Ports and Harbours Around Scotland*. Edinburgh, UK: National Marine Plan Interactive, 2023. <https://marinescotland.atkinsgeospatial.com/nmpi/default.aspx?layers=23>
- James M, Mendo T, Jones EL *et al.* AIS data to inform small scale fisheries management and marine spatial planning. *Mar Policy* 2018;91:113–21. <https://doi.org/10.1016/j.marpol.2018.02.012>
- Katara I, Silva A. Mismatch between VMS data temporal resolution and fishing activity time scales. *Fish Res* 2017;188:1–5. <https://doi.org/10.1016/j.fishres.2016.11.023>
- Lee J, South AB, Jennings S. Developing reliable, repeatable, and accessible methods to provide high-resolution estimates of fishing-effort distributions from vessel monitoring system (VMS) data. *ICES J Mar Sci* 2010;67:1260–71. <https://doi.org/10.1093/icesjms/fsq010>
- Le Guyader D, Ray C, Gourmelon F, Brosset D. Defining high-resolution dredge fishing grounds with Automatic Identification System (AIS) data. *Aquat Liv Res* 2017;30:39. <https://doi.org/10.1051/alr/2017038>
- McCauley DJ, Woods P, Sullivan B *et al.* Ending hide and seek at sea. *Science* 2016;351:1148–50. <https://doi.org/10.1126/science.aad5686>
- Mendo T, Glemarec G, Mendo J *et al.* Estimating fishing effort from highly resolved geospatial data: focusing on passive gears. *Ecol Indic* 2023;154:110822. <https://doi.org/10.1016/j.ecolind.2023.110822>
- Mendo T, Smout S, Russo T *et al.* Effect of temporal and spatial resolution on identification of fishing activities in small-scale fisheries using pots and traps. *ICES J Mar Sci* 2019a;76:1601–9.
- Mendo T, Smout S, Photopoulou T *et al.* Identifying fishing grounds from vessel tracks: model-based inference for small scale fisheries. *R Soc Open Sci* 2019b;6:191161. <https://doi.org/10.1098/rsos.191161>
- Metcalf K, Bréheret N, Chauvet E *et al.* Using satellite AIS to improve our understanding of shipping and fill gaps in ocean observation data to support marine spatial planning. *J Appl Ecol* 2018;55:1834–45. <https://doi.org/10.1111/1365-2664.13139>
- MMO. *Inshore Vessel Monitoring (I-VMS) for under-12 m Fishing Vessels Registered in England*. Newcastle upon Tyne, UK: Marine Management Organisation, 2022. <https://www.gov.uk/guidance/inshore-vessel-monitoring-i-vms-for-under-12m-fishing-vessels-registered-in-england>
- Muench A, DePiper GS, Demarest C. On the precision of predicting fishing location using data from the vessel monitoring system (VMS). *Can J Fish Aquat Sci* 2018;75:1036–47. [https://doi.org/10.1139/cjfas-2016-0446\(?PMU?\)](https://doi.org/10.1139/cjfas-2016-0446(?PMU?))
- Mujal-Colilles A, Mendo T, Swift R *et al.* Trends in maritime technology and engineering. In: Soares CG, Santos TA (Eds) *Proceedings of the 6th International Conference on Maritime Technology and Engineering* Lisbon: CRC Press, 2022.
- Natale F, Carvalho N, Paulrud A. Defining small-scale fisheries in the EU on the basis of their operational range of activity the Swedish fleet as a case study. *Fish Res* 2015;164:286–92. <https://doi.org/10.1016/j.fishres.2014.12.013>
- Navarrete Forero G, Miñarro S, Mildenerberger TK *et al.* Participatory boat tracking reveals spatial fishing patterns in an Indonesian Artisanal Fishery. *Front Marine Sci* 2017;4:409.
- Parnell PE, Dayton PK, Fisher RA *et al.* Spatial patterns of fishing effort off San Diego: implications for zonal management and ecosystem function. *Ecol Appl* 2010;20:2203–22. <https://doi.org/10.1890/09-1543.1>
- Python Software Foundation. *Python Language Reference, version 3.10*. Python Software Foundation, 2023. <http://www.python.org>.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, AUS: R Foundation for Statistical Computing, 2022. <https://www.R-project.org/>
- Rodriguez J. *iapesca, A R-package for Manipulating and Interpreting High Resolution Geospatial Data from Fishing Vessels*. Plouzane, FR: IFREMER, 2023.
- Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;20:53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Rufino MM, Mendo T, Samarão J *et al.* Estimating fishing effort in small-scale fisheries using high-resolution spatio-temporal tracking data (an implementation framework illustrated with case studies from Portugal). *Ecol Indic* 2023;154:110628. <https://doi.org/10.1016/j.ecolind.2023.110628>

- Russo T, D'Andrea L, Parisi A *et al.* VMSbase: an R-package for VMS and logbook data management and analysis in fisheries ecology. *PLoS One* 2014;9:e100195. <https://doi.org/10.1371/journal.pone.0100195>
- Russo T, D'Andrea L, Parisi A *et al.* Assessing the fishing footprint using data integrated from different tracking devices: issues and opportunities. *Ecol Indic* 2016;69:818–27. <https://doi.org/10.1016/j.ecolind.2016.04.043>
- Russo T, Parisi A, Cataudella S. New insights in interpolating fishing tracks from VMS data for different métiers. *Fish Res* 2011;108:184–94. <https://doi.org/10.1016/j.fishres.2010.12.020>
- Stelzenmüller V, Letschert J, Gimpel A *et al.* From plate to plug: the impact of offshore renewables on European fisheries and the role of marine spatial planning. *Renew Sustain Energy Rev* 2022;158:112108. <https://doi.org/10.1016/j.rser.2022.112108>
- Stelzenmüller V, Rogers SI, Mills CM. Spatio-temporal patterns of fishing pressure on UK marine landscapes, and their implications for spatial planning and management. *ICES J Mar Sci* 2008;65:1081–91. <https://doi.org/10.1093/icesjms/fsn073>
- Trouillet B. Aligning with dominant interests: the role played by geotechnologies in the place given to fisheries in marine spatial planning. *Geoforum* 2019;107:54–65. <https://doi.org/10.1016/j.geoforum.2019.10.012>
- Urbano F, Cagnacci F. *Spatial Database for GPS Wildlife Tracking Data: A Practical Guide to Creating a Data Management System with PostgreSQL/PostGIS and R*, Berlin, DE: Springer International Publishing, 2014.
- Welch H, Clavelle T, White TD *et al.* Hot spots of unseen fishing vessels. *Sci Adv* 2022;8:eabq2109. <https://doi.org/10.1126/sciadv.abq2109>
- Wickham H, Averick M, Bryan J *et al.* Welcome to the Tidyverse. *J Open Source Software* 2019;4:1686. <https://doi.org/10.21105/joss.01686>
- Wilen JE. Spatial management of fisheries. *Marine Res Econ* 2004;19:7–19. <https://doi.org/10.1086/mre.19.1.42629416>

Handling Editor: Pamela Woods