ICES Journal of Marine Science

ICES CIEM International Council for the Exploration of the Sea Conseil International pour l'Exploration de la Mer

ICES Journal of Marine Science (2017), doi:10.1093/icesjms/fsw244

Improving detection of short-duration fishing behaviour in vessel tracks by feature engineering of training data

Shay O'Farrell[1,*], James N. Sanchirico[1,2], Iliana Chollett[3], Marcy Cockrell[4], Steven A. Murawski[4], Jordan T. Watson[5], Alan Haynie[6], Andrew Strelcheck[7], and Larry Perruso[8]

[1]*Department of Environmental Science and Policy, University of California Davis, One Shields Avenue, Davis, CA 95616, USA*
[2]*Resources for the Future, Washington, DC 20036, USA*
[3]*Smithsonian Marine Station, Smithsonian Institution, Fort Pierce, FL 34949, USA*
[4]*College of Marine Science, University of South Florida, 140 Seventh Avenue South, St. Petersburg, FL 33701, USA*
[5]*National Marine Fisheries Service, Alaska Fisheries Science Center, 17109 Pt. Lena Loop Rd., Juneau, AK 99801, USA*
[6]*National Marine Fisheries Service, Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, WA 98115, USA*
[7]*National Marine Fisheries Service, Southeast Regional Office, 263 13th Avenue South, St. Petersburg, FL 33701, USA*
[8]*National Marine Fisheries Service, Southeast Fisheries Science Center, 75 Virginia Beach Dr, Miami, FL 33149, USA*

**Corresponding author: tel: 530 752 3026; fax: 530 752 3350; e-mail: shay.ofarrell.ac@gmail.com.*

O'Farrell, S., Sanchirico, J. N., Chollett, I., Cockrell, M., Murawski, S. A., Watson, J. T., Haynie, A., Strelcheck, A., and Perruso, L. Improving detection of short-duration fishing behaviour in vessel tracks by feature engineering of training data. – ICES Journal of Marine Science, doi:10.1093/icesjms/fsw244.

Received 3 October 2016; revised 2 December 2016; accepted 7 December 2016.

Big data, such as vessel monitoring system (VMS) data, can provide valuable information on fishing behaviours. However, conventional methods of detecting behaviours in movement data are challenged when behaviours are briefer than signal resolution. We investigate options for improving detection accuracy for short-set fisheries using 581 648 position records from 181 vessels in the Gulf of Mexico bandit-reel fishery. We first investigate the effects of increasing VMS temporal resolution and find that detection accuracy improves with fishing-set duration. We then assess whether a *feature engineering* approach—in our case, changing the way pings are labelled when training a classifier—could improve detection accuracy. From a dataset of 12 184 observed sets, we find that the conventional point-labelling method results in only 49% of pings being correctly labelled as 'fishing', whereas a novel window-labelling method results in 88% of records being labelled as 'fishing'. When the labelled data are used to train classifiers, point labelling attains true-positive/balanced-accuracy rates of only 37%/66%, whereas window labelling achieves 68%/83%. Finally, we map fishing distribution using the two methods, and show that point labelling underestimates the extent of fishing grounds by ∼33%, highlighting the benefits of window labelling in particular, and feature engineering approaches in general.

Keywords: bandit reel, electric reel, fishing effort, Gulf of Mexico, machine learning, pattern recognition, random forest, signal purity, statistical learning, supervised classification, track segmentation, vessel monitoring system, VMS.

Introduction

In recent years, two scientific developments have occurred with the combined potential to transform fisheries management science and policy. First, high-resolution vessel tracking technologies are providing unprecedented opportunities to develop deeper insights into fishing vessel movements both in national waters and on the high seas (McCauley *et al.*, 2016). A number of tracking systems exist, including Automatic Identification System (AIS) and Vessel Monitoring System (VMS), the latter being legally mandated in a wide range of fisheries from Alaska (NOAA, 2016) to Antarctica (CCAMLR, 2016). Second, a wide range of powerful computational tools are now being applied to identify fishing behaviours in vessel tracking data (Russo *et al.*, 2011b; Joo *et al.*, 2013; Chang and Yuan, 2014; de Souza *et al.*, 2016), and show considerable improvement over traditional but simplistic vessel speed filters that can heavily skew results, e.g. 182% overestimation of fishing in a purse seine fishery (Bertrand *et al.*, 2008; but see de Souza *et al.*, 2016).

Most work in segmenting fishing tracks into discrete behaviours has occurred in fisheries whose set durations (i.e. the periods when the gear is deployed) are longer than the corresponding VMS period, such as the Peruvian purse-seine fishery (Bertrand *et al.*, 2008; Joo *et al.*, 2011, 2013), the Norwegian stern-trawl fishery (Skaar *et al.*, 2011) and the Taiwanese pelagic longline fishery (Chang and Yuan, 2014), which have respective set durations of up to 2, 8 and 24 h. However, fisheries with set durations shorter than the VMS period present a distinct challenge (Lambert *et al.*, 2012), as sets often fall *between* sequential VMS signals and thus supervised classification or segmentation tools trained on these datasets may underrepresent the true extent of fishing.

A case in point is the Gulf of Mexico (GoM) bandit-reel fishery. With reported landings of $23M in 2014 (NMFS, 2015), the fishery has a mean set duration of ∼30 min which is considerably shorter than the mean VMS period of one hour. Bandit-reel fisheries have an advantage over other gear types in that they are able to maintain a higher quality product due to the short-set and are better able to target their fishing. Both of these traits are becoming increasingly valuable due to the trends for sustainably caught seafood around the world. Accurately identifying where and when bandit-reel vessels are fishing is therefore an important task for GoM fisheries analysts, but is impeded by the temporal mismatch between the fishing behaviour and the VMS data. Although a policy change to increase VMS signal resolution in the bandit-reel fishery may improve classification accuracy for analysis of future data, the temporal mismatch problem remains when analysing historical data that contain a wealth of information on GoM fisheries, including how fishing was displaced by large-scale disturbances such as the spatial closure imposed after the BP *Deepwater Horizon* oil spill in 2010.

To tackle the challenge of analysing data from short-set fisheries, we propose and test a novel method for labelling datasets to be used in supervised classification. Within the fields of machine/statistical-learning and big data analytics, *feature engineering* (Turner *et al.*, 1998) is a major focus of research effort, whereby gains in model/classifier accuracies are pursued by manipulating predictor variables (*data features*) rather than tuning algorithms/models (Seide *et al.*, 2011). Feature engineering approaches are now commonly evaluated in performing computational tasks as diverse as detecting sarcasm in Twitter® feeds (Bamman and Smith, 2015) and predicting the likelihood that a research paper will be accepted for presentation at a conference (Qian *et al.*, 2016). However, despite the importance of feature engineering in the computer science literature, the topic has seen surprisingly little uptake in human or wild animal movement analysis.

We assess the value of taking a feature engineering perspective when identifying fishing behaviour in a short-set fishery by modifying how training data are labelled when performing a supervised classification. When using an algorithm to recognize patterns in data, it is first necessary to train the algorithm on a dataset in which the different classes of interest have been labelled. In the case of fishing vessel tracks, when on-board observers have recorded when gears were deployed and recovered, a training set of VMS pings can be created by labelling the pings as 'fishing' or 'not fishing' based on the concurrent observer record. Algorithms can be trained to identify fishing behaviour from these labelled vessel tracks, and then identify similar behaviours in 'unseen' vessel tracks (Joo *et al.*, 2011, 2013; Chang and Yuan, 2014). However, as we show here, algorithm accuracy is sensitive to the manner in which the training data are labelled in short-set fisheries.

One of the most common approaches to labelling involves marking pings as either 'fishing' or 'not fishing' whenever an on-board observer has recorded, respectively, that gears were or were not deployed at that moment in time (Chang and Yuan, 2014). This *point-labelling* method (Figure 1, upper timeline) works well with long-set fisheries such as pelagic longline or otter trawl where multiple pings are usually transmitted during each set. However, for short-set fisheries such as bandit reel, point labelling can reduce classifier performance because vessels are usually engaged in 'fishing behaviour' for a period before and after the gear is deployed. If gears are deployed just after a ping has been transmitted—or are recovered just before a ping has been transmitted—that ping will be labelled as 'not fishing', which may mislead the classifier during training. Our proposed solution, *window labelling*, is to label a VMS record as 'fishing' if gears were deployed at any time during the ∼hourly *ping window* surrounding that ping, thereby capturing fishing behaviour that took place when no synchronous ping occurred.

Our study is presented in three parts. First, we investigate whether bandit-reel classification accuracy improves with signal purity (i.e. as the behaviour duration approaches the signal period). Second, we use our window-labelling approach to determine whether classifier accuracy may be improved by appropriate feature engineering, regardless of signal purity. Finally, we

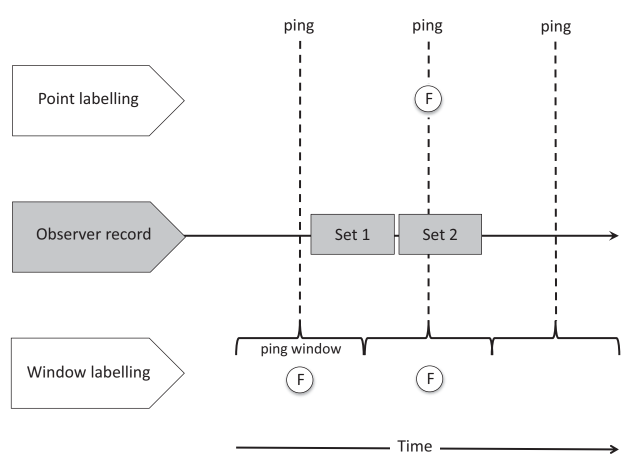

Figure 1. Conceptual diagram illustrating how the point-labelling and window-labelling methods will result in different pings being labelled as 'fishing' or 'not fishing' in a VMS record from a trip on which an observer was aboard the vessel. The middle timeline (grey boxes) shows the observer's record of the times when sequential sets started and ended. VMS ping transmission times are illustrated by vertical dashed lines. The top and bottom timelines, respectively, show how a concurrent VMS record would be labelled using point- and window-labelling, with circled letters *F* indicating which pings would be labelled as 'fishing' by each method. For details, see subsection *Comparing Labelling Methods*.

contrast the outcome of using the point- and window-labelling methods to map bandit-reel fishing distribution in the US GoM. Although we use random forest (Breiman, 2001) in our study, we stress that our method is classifier independent as it simply involves modifying the way in which the training data are labelled, and we hope that it may be useful to researchers working with a wide range of analytical approaches.

Methods

Datasets

Three datasets were used in the analysis, namely Vessel Monitoring System (VMS) data, on-board observer programme data and vessel logbooks. VMS transponders sending hourly (or better) reports have been mandated on all commercial reef-fish fisheries vessels in the GoM since 2007 (eCFR, 2016) but owing to incomplete records in the first year of operation, only VMS data for the years 2008–2012 were used in the present work. The US National Observer programme monitors nearly 50 different US fisheries to satisfy requirements of the Magnuson-Stevens Fishery Conservation and Management Act, the Marine Mammal Protection Act and the Endangered Species Act, among others (NMFS, 2016). In the GoM as elsewhere in the US, observers accompany a sample of commercial fishing trips, recording the gears used on each trip, plus the start- and end-time of each set. Furthermore, all commercial vessels in the GoM are required to maintain a logbook recording gear type used on each trip. The logbook dataset was used at the outset to reduce the 20-million-record VMS dataset and the 31 744-record observer dataset down to trips on which bandit reel was the only gear used. The VMS data were then divided into trips with and without observers. The VMS data from trips with observers were used to create a machine-learning training and validation dataset by cross-referencing the VMS data with the observer data so the start- and end-times of each set recorded by the observers could then be used to label the VMS pings as 'fishing' or 'not fishing'. This labelled training dataset consisted of 43 077 position records for 112 vessels recorded over 275 trips, during which 12 184 individual fishing sets were observed. Finally, the remaining 538 571 VMS records from trips without observers were used to map bandit-reel fishing locations throughout the US GoM, using the trained classifiers to identify fishing behaviour.

Data processing

Rapid loading of the large datasets was enabled by the *R*-package, *data.table* (Dowle *et al*., 2015). The bandit-reel fishery targets mostly red snapper and shallow water groupers, which are found within the boundary of the continental shelf and boats rarely fish beyond the 200-m isobath (Figure 2). To improve processing efficiency, the dataset was therefore reduced by excluding vessel positions that were in water >500 m depth, which was considered to be a conservative threshold. Depth at each vessel position was extracted from the ETOPO1 database (Amante and Eakins, 2009) using the *R*-package, *marmap* (Pante and Simon-Bouhet, 2013). VMS records that had GPS coordinates outside the GoM basin, missing or duplicated timestamps, or impossible speeds (using an arbitrary threshold of 20 ms^{-1}) were omitted from the analysis.

Seven metrics were derived from the VMS data and used as inputs during classification of fishing behaviour. The distance from the preceding GPS position (*leg* distance) and compass heading were calculated using spherical trigonometry functions

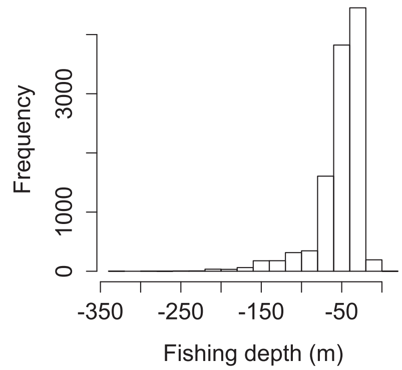

Figure 2. Depths at fishing locations recorded during bandit-reel trips when an observer was aboard the vessel.

(*bearingRhumb* and *distRhumb*) in the *R*-package *geosphere* (Hijmans, 2015). Absolute turning angles—specifically, the magnitude of change in angle regardless of whether clockwise or anticlockwise—for each VMS record were calculated from the headings of the legs to and from that position. Decimal hour of day was calculated from timestamps and expressed to four decimal places. Three metrics of vessel speed were calculated, namely 'pre-position' and 'post-position' speeds, and a three-position average speed. For example, A, B and C are three successive VMS positions. The pre-position speed at B was calculated as distance travelled/time interval from A to B, the post-position speed at B was the distance travelled/time interval from B to C, and the three-position average speed at B was calculated as total distance travelled/total time interval from A to C.

Supervised classification

Supervised classification was conducted using an *R* implementation (Liaw and Wiener, 2002) of *random forest* (Breiman, 2001), which is one of the most popular classifiers in movement pattern recognition (Joo *et al*., 2013). Random forest has the advantage of being robust to overfitting to training data when performing supervised classification, which it achieves by iteratively bootstrapping the labelled training dataset and subsampling the putative predictor variables (aka *features or attributes*), then fitting a classification tree to each subsampled 'bag' (Strobl *et al*., 2009). Individual variables may be re-selected at successive nodes of the same tree, allowing highly nonlinear and complex interactions to be captured. Once an ensemble (forest) of trees has been built in this manner, unlabelled data records are passed to the individual trees. Each tree assesses the record and classifies it according to the tree's own random subset of features, meaning that trees often disagree on the class to which the record belongs. The final classification of the record is declared by taking a 'vote' across the entire forest ensemble, with the most popular class being selected as the final label. About 30% of the labelled records in each bag are not used during the classification itself, but are set aside as 'out-of-bag' (OOB) samples for internally validating the accuracy of the classification. OOB error rate often replaces formal out-of-sample cross-validation (CV) in random forest classifications, and this practice was tested here by comparing OOB error and CV error when analysing signal purity. Two random forest ensembles were constructed to recognize fishing behaviour in VMS data, one using training data labelled with the point-labelling method and the other using training data labelled with the window-labelling method. The final variable sets used in the point- and window-labelling forests were selected using the *R*-package, pRF (Chakravarthy, 2016), which uses permutation tests to determine which variables contributed significantly to the ensemble.

Measuring classifier performance

Biased predictions can occur when the number of cases differs between classes, as with VMS data which will generally contain more 'not fishing' than 'fishing' records. When the class of interest (e.g. fishing) has fewer cases, then metrics of classifier performance will tend to provide optimistic results unless the imbalance is taken into account (Brodersen *et al.*, 2010). Furthermore, true positive rate should not be used in isolation as a performance metric when classifying movement tracks as it rewards correct prediction of a positive case but does not penalize incorrect prediction of a negative case. For instance, consider a dataset with 1000 VMS pings, of which 100 were transmitted during fishing behaviour and the other 900 were transmitted during non-fishing behaviours. If a classifier simply predicted that all 1000 VMS pings represented fishing behaviour, it would achieve a true positive rate of 1.0 (i.e. 100% correct) because it successfully identified all 100 fishing pings. In the example, however, the classifier would receive a false positive rate of 0.9 (i.e. 900/1000 non-fishing pings were incorrectly labelled as fishing) and a true negative rate of 0, as none of the non-fishing pings was successfully identified as non-fishing.

The balanced accuracy metric (Brodersen *et al.*, 2010) simplifies classifier performance as the arithmetic mean of the true positive and true negative rates, or 0.5 in the example. For a two-class problem, a balanced accuracy rate of 0.5 may be interpreted as getting all of one class right and all of the other class wrong. It can be seen that a particular advantage of balanced accuracy is that the metric is insensitive to class imbalance in the data—in our example, 100 records in one class and 900 in the other. Consequently, behaviour classes for which there are large numbers of records, such as for a vessel sitting in port for weeks on end transmitting hourly VMS pings, do not overwhelm behaviour classes for which there are fewer VMS pings, such as for a vessel actively fishing. The present analysis used balanced accuracy alongside true positive rate and false positive rate to assess the performance of the two random forest ensembles.

Assessing the influence of signal purity

In the case of fishing, perfect signal purity occurs when the duration of the set is at least that of the VMS temporal resolution. As the present study used historical VMS data, it was not possible to manipulate the periodicity of the VMS pings to match bandit-reel set duration, so instead the converse was done. Because the duration of fishing sets varied, it was possible to calculate the proportion of each ping window that was spent fishing and then determine the effect of variation in that proportion on the predictive accuracy of the classifier. For example, consider two-hour-long ping windows. First, a window that contained two 25-min fishing sets, meaning that the proportion of that window during which the vessel was fishing was 0.83 (50/60 min). Second, a window that contained only a single six-minute set, giving a fishing proportion of 0.1 (6/60 min). Even though fishing occurred within both ping windows, it would be expected that a classifier would have greater success in correctly identifying that fishing had occurred in the former case, as more of that window was spent engaged in fishing than in other behaviours. For analysing the effect of signal purity on classifier accuracy, ping windows were created around each ping, calculated as the sum of half of the intervals before and after each ping. For example, if consecutive pings were recorded at 9, 10 and 11 a.m., the window around the middle VMS ping would be from 9.30 to 10.30 a.m. The observer data were then used to assign a value to each VMS record representing the proportion of the ping window that was spent fishing, including zero for non-fishing pings. To assess the effects of signal purity on classifier performance, the proportion of each VMS ping window that was spent fishing was calculated, with 'fishing' and 'not fishing' labels being assigned using the conventional point-labelling method. The data were then grouped by signal purity into five levels (0.2, 0.4, 0.6, 0.8 and 1.0) and random forest was sequentially trained and assessed at each level. Both out-of-bag (OOB) error validation and formal out-of-sample cross validation (CV) were performed for 20 replicate runs for each signal purity level.

Comparing labelling methods

Point labelling and window labelling will result in differences in which VMS pings are labelled as 'fishing' or 'not fishing' prior to conducting a supervised classification (Figure 1). In the diagram, a ping was transmitted while Set 2 was underway, so the concurrent VMS ping is marked as 'fishing' on the upper timeline. However, no pings were transmitted when Set 1 was underway. The bottom timeline shows the window-labelling method, where a ping is labelled as 'fishing' whenever a set or part thereof occurred within a time window around that ping, the ping window, illustrated using sequential horizontal braces. Here, the first two pings would be labelled 'fishing'. The improved classifier accuracy that results from window labelling may partially be explained by the fact that vessels are generally engaged in fishing behaviour for a period immediately before and after gears are deployed, such as the vessel remaining stationary while hooks are baited prior to fishing, or while gears are stowed after fishing. In the diagram, the observed vessel would have been displaying movement behaviours characteristic of fishing when the ping was transmitted prior to the start of Set 1, but the point-labelling method would have labelled the ping as 'not fishing' thereby confusing the classifier. In the present analysis, the two different methods were used in turn apply 'fishing' or 'not fishing' labels to the same dataset of VMS pings from trips which had a fisheries observer on board. A random forest ensemble was then built for each training dataset, allowing the effect of the labelling methods on classifier accuracy to be quantified and validated.

Finally, the two random forest ensembles were used to classify unseen VMS pings into 'fishing' or 'not fishing'. Maps of fishing locations were produced by exporting the VMS coordinates labelled as 'fishing' to a Geographic Information System (QGIS, 2016) and calculating the number of pings that fell inside each cell of a 0.1 degree grid within the area of interest (17–31N, 79–98W). The difference in spatial distribution of fishing predictions between the point- and window-labelling methods was then quantified as the number of positive predictions of fishing for window labelling minus the number of positive predictions for point labelling within each grid cell. Because the area of grid cells defined in degrees changes with latitude, the difference in fished area predicted by the two methods was determined by projecting the grid in Web Mercator and summing the areas of the individual cells for each method.

Results

The accuracy of the classifier in correctly identifying fishing behaviour increases with signal purity level (Figure 3), with the relationship starting to plateau as perfect signal purity is approached. The number of training records contained within each signal purity subset is shown in Table 1. The accuracies quantified by the OOB (diamond markers, Figure 3) and CV (circle markers, Figure 3) errors were similar.

To compare the difference in the accuracy of classifier trained using point- vs. window-labelling methods, OOB error was used without any additional CV so that all of the samples could be used. Of the 43 077 observed VMS records (i.e. bandit-reel trips for which synchronous VMS and observer datasets existed), point labelling and window labelling, respectively, resulted in 5948 and 10 723 of the VMS records labelled as 'fishing'. A sensitivity analysis showed that classifier accuracy increased with training sample size for both point labelling and window labelling, but the latter proved consistently the more accurate method across all sample sizes (Figure 4). Point labelling attained a maximum balanced accuracy rate (arithmetic mean of true positive and true negative rates) of 0.66 (Figure 4a), a maximum true positive rate of 0.37 (Figure 4b) and a maximum false positive rate of 0.05 when all 5948 training samples were used. At the same levels of sampling, window labelling achieved a balanced accuracy rate of 0.83 (Figure 4a), a true positive rate of 0.68 (Figure 4b) and a false positive rates of 0.07. Accuracy measures were continuing to increase for window labelling by the time all 10 723 labelled records were used in training the classifier.

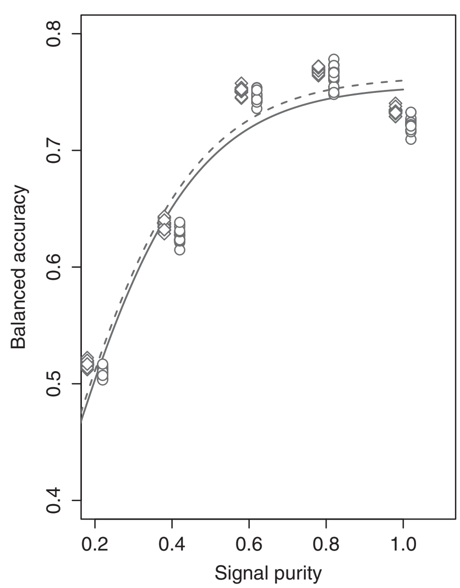

Figure 3. Accuracy in classifying bandit-reel fishing behaviour in VMS data increases with signal purity, expressed here as the proportion of each ~hourly VMS window in which fishing gear was deployed. Balanced accuracy was calculated as the arithmetic mean of true positive rate and true negative rate. Diamonds and dashed line, respectively, show the classifier accuracy calculated from random forest out-of-bag (OOB) samples and a subsequent logistic model fit. Circles and solid line show the same for cross-validated (CV) samples. Markers show 20 replicate classifier runs for each signal purity level for both OOB and CV accuracy calculations. OOB and CV markers have been offset horizontally for clarity.

Point and window labelling were used to map bandit-reel fishing extent and intensity in the GoM. Two random forests were trained to identify fishing behaviour using the 43 077 observed VMS records. The ranked importance of the predictor variables to each forest ensemble is shown in Table 2. Each forest was then used to classify the 538 571 VMS pings from trips for which an observer was not aboard the vessel. The number of pings classified as 'fishing' in each cell was quantified and mapped for point labelling (Figure 5a) and window labelling (Figure 5b).

Finally, the difference between the two maps was quantified on a cell-by-cell basis (Figure 5c) by subtracting the density of point-labelled pings (Figure 5a) from the density of window-labelled pings (Figure 5b). A total fished area of 157 249 km^2 was predicted by point labelling whereas the more accurate window-labelling method predicted a 47% greater area of 230 772 km^2. The difference between the methods was highest in grid cells where the window-labelling method predicted high fishing intensity, indicating that the benefits of window labelling in accurately estimating fishing intensity are most pronounced in heavily fished grounds (Figure 5c).

Discussion

By combing three large datasets from a short-set fishery, we found that the accuracy of a classifier at identifying fishing

Table 1. The number of fishing records contained within VMS subsets according to signal purity, which was used to analyse the effect of increasing signal purity on classifier accuracy.

Signal purity level	$x \leq 0.2$	$0.2 < x \leq 0.4$	$0.4 < x \leq 0.6$	$0.6 < x \leq 0.8$	$0.8 < x \leq 1.0$
Number of records	904	1953	1578	896	617

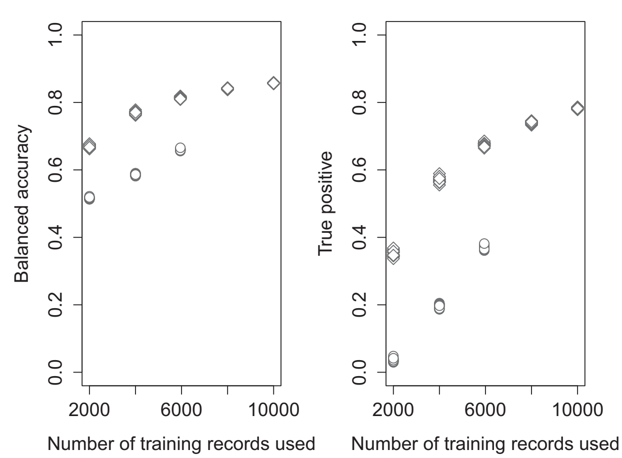

Figure 4. Balanced accuracy (left panel) and true positive rate (right panel) both increase with the number of training samples used for both conventional point labelling (circles) and window labelling (diamonds). Markers show 20 replicate classifier runs for each labelling method and each sampling level.

Table 2. Ranked importance of predictor variables in constructing each forest.

Point-labelled random forest		**Window-labelled random forest**	
Features	**Mean decrease in Gini index**	**Features**	**Mean decrease in Gini index**
Hour of day*	2013	Velocity mean*	3381
Velocity mean*	1899	Hour of day*	3282
Depth*	1746	Depth*	3031
Velocity prior	1534	Leg distance prior*	2487
Leg distance prior	1450	Velocity prior*	2442
Velocity post	1346	Velocity post	1964
Turn angle	1052	Turn angle	1162
Heading	997	Heading	989

Variables that contributed significantly to each forest are indicated with an asterisk.

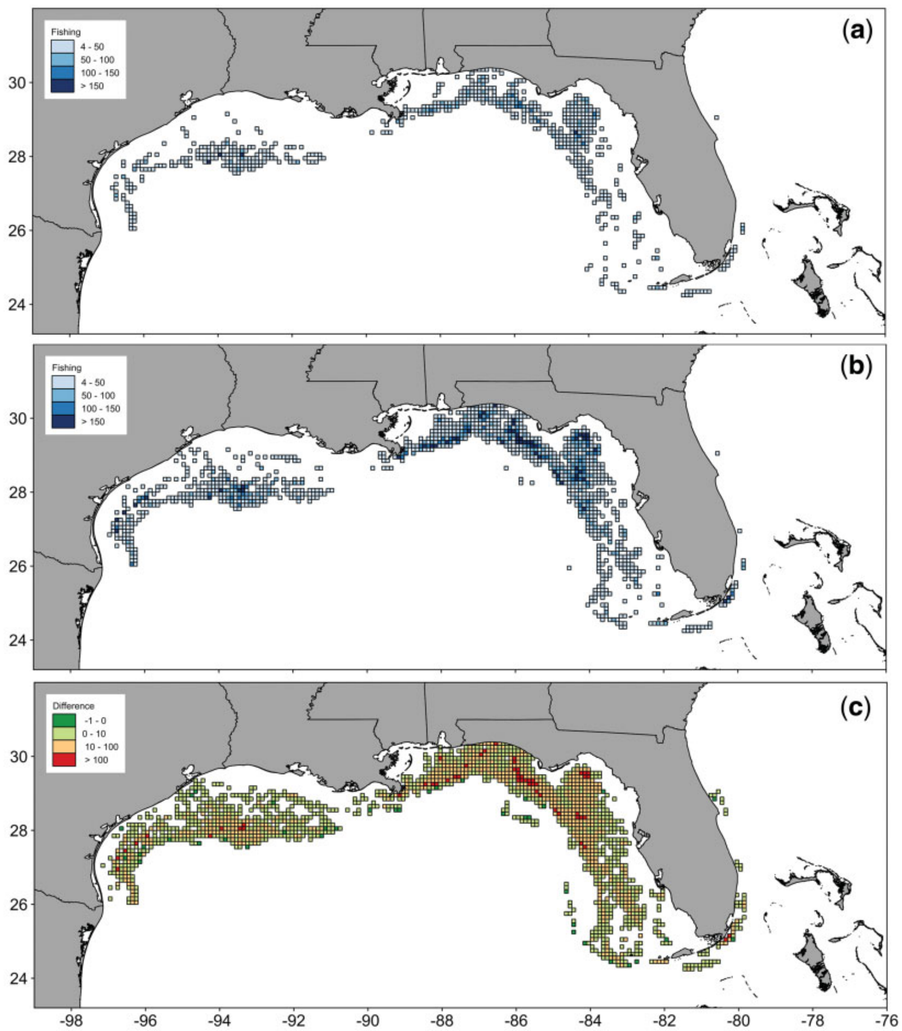

Figure 5. (a and b) Extent and intensity maps of US Gulf of Mexico bandit-reel fishing produced by training a random forest classifier to identify fishing behaviour in VMS data using two methods, point labelling (a) and window labelling (b). Shading shows the count of VMS pings classified as fishing behaviour in each 0.1° grid cell. (c) This shows the difference between the methods in the number of VMS pings identified as fishing behaviour in each cell.

behaviour may be markedly improved either by a policy change to increase the temporal resolution of VMS data gathering or by a change in the way the training data are labelled for classification. Taking both courses of action would, we propose, have the greatest benefit as each has advantages. Increasing the temporal resolution of the VMS data would enhance classifier accuracy in the future, which would in turn improve the identification of fishing locations, increase the spatial resolution at which behaviours can be mapped, and also the estimation of the amount of time spent actively fishing. On the other hand, window labelling may improve accuracy in the present day when analysing either current or historical data whose temporal resolution is lower than the behaviour duration, or when interpolating between missing data points in high resolution data.

Another feature-engineering approach to improving the detection of short-set fishing behaviour could be to statistically amplify the temporal resolution of the data by interpolating between successive VMS pings using approaches such as cubic Hermite splines (Hintzen *et al.*, 2010) or the modified Catmull–Rom algorithm, CRm (Russo *et al.*, 2011a). CRm interpolation accommodates estimates of vessel drift caused by wind and wave action as well as human control and has been shown robust at interpolating between VMS pings across a range of métiers (Russo *et al.*, 2011a) representing active, passive and towed gears. CRm interpolation requires the availability of instantaneous measures of vessel speed and heading, which can be transmitted along with the position and timestamp data. However, as these data are not systematically transmitted by the Gulf of Mexico (GoM) VMS implementation, it was not possible in the present work to investigate the magnitude of further gains in accuracy that could be achieved using the CRm interpolation algorithm.

Although we used random forests as the classifier in our analysis, we hope and intend that window labelling may prove useful for researchers using a wide range of tools in vessel movement analysis as the method involves engineering of the training data labels rather than tuning of a particular model or algorithm. The fundamental problem with conventional point labelling of short-set fishing behaviours is that it can induce unrepresentative labelling of the training data, because the instantaneous nature of signal transmission means that short-duration sets are often 'just missed' and so a VMS ping will be labelled as not-fishing even if the vessel is engaged in fishing activities such as gear setting or catch sorting which are visible to movement data analyses. Improving the training dataset may provide benefits in supervised learning regardless of the learning tool being used, although the magnitude of gains may of course vary among tools.

As well as benefits to window labelling, however, there are also caveats. The most notable limitation of our method is that it can result in a small increase (2% in our analysis) in false positives, as a VMS record will be marked as 'fishing' even if the set finished before the ping was transmitted and the vessel is now steaming. Although our results show that any increase in false positives is more than offset by the decrease in false negatives in terms of balanced accuracy, it does raise concerns about using window labelling for enforcement purposes where fine-scale error may result in a vessel being incorrectly flagged as fishing across a no-take boundary when in fact the vessel is merely steaming after recovering its gear. The limitation is presently hypothetical in many locations as VMS data are not universally used for enforcement at this time, but increasing the temporal resolution of the data from short-set fisheries may mitigate this problem before it arises.

Our finding that classifier accuracy increases with the temporal resolution of the GoM bandit reel VMS data contributes to a body of work showing that vessel monitoring data are often transmitted with suboptimal frequency for detecting movement behaviours. In their study of two fisheries from the Isle of Man, Lambert *et al.* (2012) examined how the temporal resolution of VMS data gathering affected indicators of fishing intensity. The study used data from the king scallop dredge fishery and the queen scallop otter trawl fishery, two benthic towed-gear fisheries which are likely to exhibit notably different movement behaviours to the pelagic vertical-line bandit reel fishery in the GoM. By gathering a study-specific high-resolution movement track, the authors were able to determine that a VMS period of 30 min would provide an optimal balance between data accuracy and data-gathering efficiency. Working with the Peruvian anchovy purse-seine fishery, Joo *et al.* (2013) took a different approach to assessing how temporal resolution affects detection of movement behaviours. By simulating data with one-second resolution, the authors found that accuracy of their models approached 100%. Three analyses (Lambert *et al.*, 2012; Joo *et al.*, 2013 and the present work) using three experimental approaches on four markedly different fisheries all found that increased temporal resolution improved the ability to detect movement behaviours in VMS data.

The maps we present in Figure 5 represent the distribution and intensity of fishing behaviour in the GoM. In addition to these spatial parameters, the standard protocol for measuring bandit reel fishing effort requires information on three further parameters: the number of hooks per line, the number of lines deployed and the total soak time (Scott-Denton *et al.*, 2011). Although these data were not available for our analysis, it is commonplace in the VMS literature to use fishing activity as a proxy for effort (Witt and Godley, 2007) and our maps may be considered analogous.

Although window labelling was developed to improve detection accuracy in the GoM bandit reel fishery, we hope that the method may prove useful to researchers working with other fisheries where the behaviours of interest are shorter than the periodicity of the signal. In particular, other vertical line fisheries may be particularly suitable for window labelling. It is also possible that window labelling may by beneficial in detecting longer duration fishing behaviours. During exploratory analysis of VMS data from the GoM benthic longline fishery that has a mean set duration of greater than two hours, we found that window labelling also improved classifier accuracy, although the gains are not as great as with the bandit-reel data. In general, we feel that feature engineering of vessel movement datasets has received substantially less interest than have comparisons among various classifiers and models to detect fishing behaviours. We would encourage other researchers to consider whether modifying their data, whether through window labelling or other methods, could improve detection accuracy irrespective of the analytical tools being used. With the increasing interest in using VMS and AIS data for monitoring fishing both in national waters and on the high seas (McCauley *et al.*, 2016), any improvements in the accuracy of detecting behaviours may be considered to be beneficial.

Looking to the future, a characteristic of the GoM observer dataset is that it only records when gears were deployed and recovered, but does not record when a vessel was engaged in other important behaviours, such as steaming, repositioning between sets, preparing gears for deployment, processing catch after gear recovery, and searching for fish. The omission is not a fault of the dataset as the observer program was not designed to gather data for training classification algorithms, but it does limit what can be done with the otherwise extensive and meticulously gathered dataset. Observer datasets with higher 'behavioural resolution' have been gathered in other fisheries such as the Peruvian purse-seine fishery (Joo *et al.*, 2013). Given the promise of the new movement technologies and methodologies that are presently being developed in fisheries, it would be beneficial for future analyses to increase the behavioural resolution of the GoM observer dataset, and would incur little additional effort or cost.

In terms of further work, both the point- and window-labelling methods were predictably sensitive to the number of training samples used. In the present study, we used only data from 2008 to 2012, and it is clear that maximum accuracy had not yet been reached by the time all of the samples were used. We anticipate that accuracy may modestly increase as further samples are added, and that an ensemble classifier trained on subsequent data from 2013 to 2016 may improve on our results.

Acknowledgements

We would like to thank Elizabeth Scott-Denton of the Southeast Fisheries Science Center, Galveston Laboratory, for supplying Gulf of Mexico observer data, and Kelly Spalding (NMFS VMS Program Manager, NOAA, 1315 East West Highway, Silver Spring, MD 20910) and Carlos Rivero (Southeast Fisheries Science Center, Beaufort Laboratory) for supplying VMS data.

Funding

Funding for the work came from a National Science Foundation Coastal SEES grant awarded to S.M. and J.N.S. (grant #2500-1565-00) and a NMFS sub-award to S.M. and J.N.S. (award #201502599-01) through the Spatial Economics Toolbox for Fisheries (FishSET).

References

Amante, C., and Eakins, B. W. 2009. ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis. US Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, National Geophysical Data Center, Marine Geology and Geophysics Division Colorado.

Bamman, D., and Smith, N. A. 2015. Contextualized sarcasm detection on twitter. International AAAI Conference on Web and Social Media, North America, April 2015. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10538 (last accessed 05 January 2017).

Bertrand, S., Díaz, E., and Lengaigne, M. 2008. Patterns in the spatial distribution of Peruvian anchovy (Engraulis ringens) revealed by spatially explicit fishing data. Progress in Oceanography, 79: 379–389.

Breiman, L. 2001. Random forests. Machine Learning, 45: 5–32.

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. 2010. The balanced accuracy and its posterior distribution. Pages 3121-3124 in Pattern recognition (ICPR), 2010 20th international conference on. IEEE.

CCAMLR. 2016. Commission for the Conservation of Antarctic Marine Living Resources. Vessel Monitoring System (VMS). https://www.ccamlr.org/en/compliance/vessel-monitoring-system-vms (last accessed 1 July 2016).

Chakravarthy, A. 2016. pRF: Permutation Significance for Random Forests. R package version 1.2. http://CRAN.R-project.org/package=pRF.

Chang, S. K., and Yuan, T. L. 2014. Deriving high-resolution spatiotemporal fishing effort of large-scale longline fishery from vessel monitoring system (VMS) data and validated by observer data. Canadian Journal of Fisheries and Aquatic Sciences, 71: 1363–1370.

de Souza, E. N., Boerder, K., Matwin, S., and Worm, B. 2016. Improving fishing pattern detection from satellite AIS using data mining and machine learning. PLoS One, 11: e0158248.

Dowle, M., Srinivasan, A., Short, T., and Lianoglou, S. 2015. data.table: Extension of Data.frame. R package version 1.9.6. https://CRAN.R-project.org/package=data.table.

eCFR. 2016. Electronic Code of Federal Regulations. US Government Publishing Office. http://www.ecfr.gov/cgi-bin/text-idx?SID=86d3e4e21c5c4a3cd94b7f259d8700e1&node=50:12.0.1.1.2&rgn=div5#se50.12.622_128 (last accessed 1 July 2016).

Hijmans, R. J. 2015. geosphere: Spherical Trigonometry. R package version 1.5-1. https://CRAN.R-project.org/package=geosphere.

Hintzen, N. T., Piet, G. J., and Brunel, T. 2010. Improved estimation of trawling tracks using cubic Hermite spline interpolation of position registration data. Fisheries Research, 101: 108–115.

Joo, R., Bertrand, S., Chaigneau, A., and Niquen, M. 2011. Optimization of an artificial neural network for identifying fishing set positions from VMS data: an example from the Peruvian anchovy purse seine fishery. Ecological Modelling, 222: 1048–1059.

Joo, R., Bertrand, S., Tam, J., and Fablet, R. 2013. Hidden Markov models: the best models for forager movements? PLoS One, 8: e71246.

Lambert, G. I., Jennings, S., Hiddink, J. G., Hintzen, N. T., Hinz, H., Kaiser, M. J., and Murray, L. G. 2012. Implications of using alternative methods of vessel monitoring system (VMS) data analysis to describe fishing activities and impacts. ICES Journal of Marine Science, doi:10.1093/icesjms/fss018.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. R News, 2: 18–22.

McCauley, D. J., Woods, P., Sullivan, B., Bergman, B., Jablonicky, C., Roan, A., Hirshfield, M., Boerder, K., and Worm, B. 2016. Ending hide and seek at sea. Science, 351: 1148–1150.

NMFS. 2015. 2014 Gulf of Mexico Red Snapper Individual Fishing Quota Annual Report. National Marine Fisheries Service. http://sero.nmfs.noaa.gov/sustainable_fisheries/ifq/documents/pdfs/annual_reports/2014_rs_annualreport.pdf (last accessed 19 August 2016).

NMFS. 2016. National Observer Program, National Marine Fisheries Service. http://www.st.nmfs.noaa.gov/st4/nop/ (last accessed 11 July 2016).

NOAA. 2016. Alaska Regional Office, National Marine Fisheries Service. https://alaskafisheries.noaa.gov/fisheries/cwm (last accessed 8 July 2016).

Pante, E., and Simon-Bouhet, B. 2013. marmap: a package for importing, plotting and analyzing bathymetric and topographic data in R. PLoS One, 8: e73051. doi:73010.71371/journal.pone.0073051.

QGIS. 2016. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://www.qgis.org/.

Qian, Y., Dong, Y., Ma, Y., Jin, H., and Li, J. 2016. Feature Engineering and Ensemble Modeling for Paper Acceptance Rank Prediction. 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. arXiv:1611.04369.

Russo, T., Parisi, A., and Cataudella, S. 2011a. New insights in interpolating fishing tracks from VMS data for different métiers. Fisheries Research, 108: 184–194.

Russo, T., Parisi, A., Prorgi, M., Boccoli, F., Cignini, I., Tordoni, M., and Cataudella, S. 2011b. When behaviour reveals activity: Assigning fishing effort to métiers based on VMS data using artificial neural networks. Fisheries Research, 111: 53–64.

Scott-Denton, E., Cryer, P. F., Gocke, J. P., Harrelson, M. R., Kinsella, D. L., Pulver, J. R., Smith, R. C., and Williams, J. A. 2011. Descriptions of the US Gulf of Mexico reef fish bottom longline and vertical line fisheries based on observer data. Marine Fisheries Review, 73: 1–26.

Seide, F., Li, G., Chen, X., and Yu, D. 2011. Feature engineering in context-dependent deep neural networks for conversational speech transcription. Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE. pp. 24–29.

Skaar, K. L., Jørgensen, T., Ulvestad, B. K. H., and Engås, A. 2011. Accuracy of VMS data from Norwegian demersal stern trawlers for estimating trawled areas in the Barents Sea. ICES Journal of Marine Science, 68: 1615–1620.

Strobl, C., Malley, J., and Tutz, G. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods, 14: 323.

Turner, C. R., Wolf, A. L., Fuggetta, A., and Lavazza, L. 1998. Feature engineering. Proceedings of the 9th International Workshop on Software Specification and Design. IEEE Computer Society, pp. 162–164.

Witt, M. J., and Godley, B. J. 2007. A step towards seascape scale conservation: using vessel monitoring systems (VMS) to map fishing activity. PLoS One, 2: e1111.

Handling editor: Stephen Eayrs