Vol. 36: 89–98, 2018
https://doi.org/10.3354/esr00894

ENDANGERED SPECIES RESEARCH
Endang Species Res

Published June 13

Contribution to the Special 'Marine pollution and endangered species'

A machine-learning approach to assign species to 'unidentified' entangled whales

James V. Carretta*

Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, 8901 La Jolla Shores Drive, La Jolla, CA 92037, USA

ABSTRACT: Whale entanglements in US west coast fishing gear are largely represented by opportunistic sightings, and some reports lack species identifications due to rough seas, distance from whales, or a lack of cetacean identification expertise. Unidentified entanglements are often ignored in species risk assessments and thus, entanglement risk is underestimated. To address this negative bias, a species identification model was built from random forest (RF) classification trees using 199 identified entanglements ('model data'). Humpback *Megaptera novaeangliae* and gray whales *Eschrichtius robustus* represented 92% of identified entanglements; the remaining 8% were minke whales *Balaenoptera acutorostrata*, fin whales *B. physalus*, blue whales *B. musculus*, and sperm whales *Physeter macrocephalus*. Predictor variables included year, gear type, location, season, sea surface temperature, water depth, and a multivariate El Niño index. Cross-validated species classifications were correct in 78% (155/199) of cases, significantly higher ($p < 0.001$, permutation test) than the 49% correct classification rate expected by chance. The RF model correctly classified 91% of humpback whale cases, 64% of gray whale cases, and 100% of sperm whale cases, but misclassified all minke, blue, and fin whale cases. The cross-validated RF classification-tree species model was used to classify 35 entanglements without species identifications ('novel data') and each case was assigned a probability of belonging to each of 6 model data species. This approach eliminates the negative bias associated with ignoring unidentified entanglements in species risk assessments. Applications to other wildlife studies where some detections are unidentified include fisheries bycatch, line-transect surveys, and large-whale vessel strikes.

KEY WORDS: Fishery entanglement · Humpback whale · Gray whale · Species assignment · Random forest

INTRODUCTION

Entanglement of large whales in fishing gear and marine debris is a source of anthropogenic mortality and serious injury worldwide (Read et al. 2006, Bradford et al. 2009, Cassoff et al. 2011, Meÿer et al. 2011, Groom & Coughran 2012, Knowlton et al. 2012, Moore 2014, van der Hoop et al. 2017). Documented entanglements represent a minimum accounting of impacts, because not all at-sea entanglements are detected; either the whale is never seen or observers fail to recognize that a whale is entangled. Negative reporting biases are not limited to at-sea sightings. Beach-stranded carcasses may go undetected along remote coastlines or detected carcasses may lack visible evidence of entanglement due to decomposition and thus, are not categorized as anthropogenic mortality. Studies of recovery rates of cetacean carcasses suggest that observed levels of anthropogenic mortality and injury grossly underestimate actual levels (Knowlton & Kraus 2001, Kraus et al. 2005, Williams et al. 2011), even for extremely coastal species (Prado et al. 2013, Wells et al. 2015, Carretta et al. 2016a). Compounding the problem of incomplete detection is that not all at-sea sightings of entanglements are identified to species. Approximately 15% of US west coast whale entanglement cases lack species identifications due to rough seas, observer distance to

*Corresponding author: jim.carretta@noaa.gov

Publisher: Inter-Research · www.int-res.com

whales, or a lack of whale identification expertise (Carretta et al. 2016b). Quantitative methods to prorate unidentified cases to species are lacking in US marine mammal stock assessments (Muto et al. 2016, Waring et al. 2016, Carretta et al. 2017); thus the perceived entanglement risk to some species is negatively biased via omission of these cases. To better account for entanglement risk, I developed a species classification model using random forest (RF) classification trees (Breiman 2001a,b, Liaw & Wiener 2002), which are used to classify unidentified sightings of entangled whales to species.

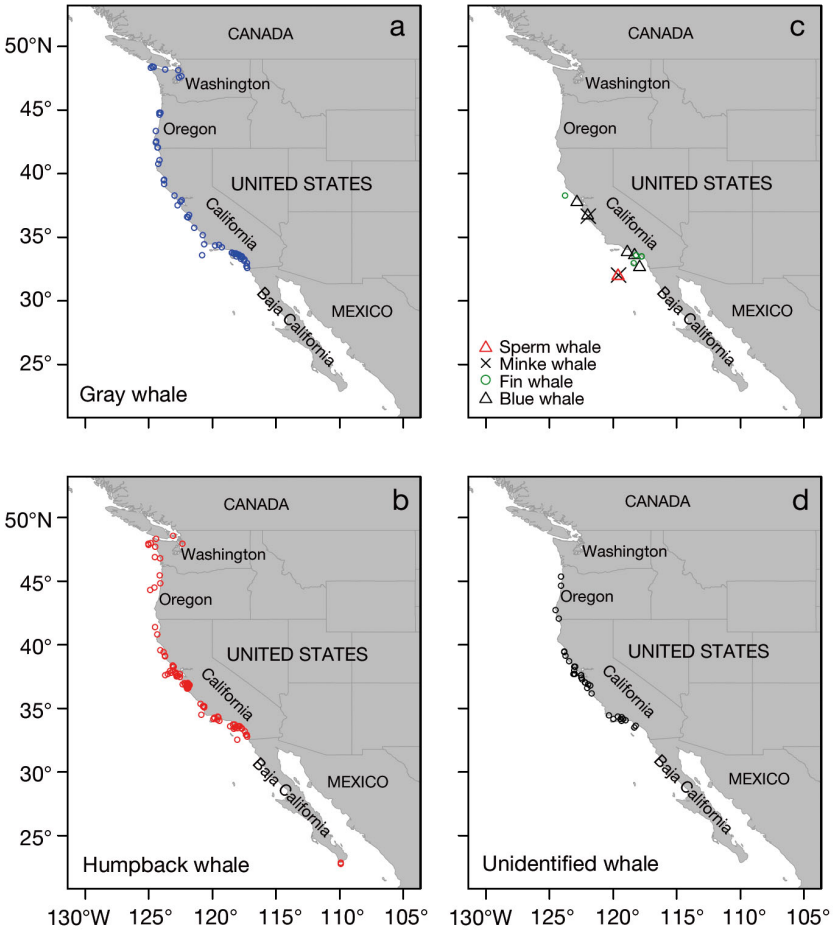

Fig. 1. Locations of (a–c) model data (identified species — a: gray whale, b: humpback whale, c: sperm, minke, fin, and blue whales) and (d) novel data (unidentified) whale entanglements used in this analysis

METHODS

Data and model overview

Data on large-whale entanglements are compiled by the National Oceanic and Atmospheric Administration (NOAA) through regional marine mammal stranding networks and disentanglement teams (Carretta et al. 2016b). Reports and sightings are verified with photos and/or video when possible, but many records are opportunistically reported; thus species identification and the type of fishing gear involved in entanglements are sometimes based on first-hand accounts. Only entanglement records with photo/video documentation or those received from reporting parties considered reliable (i.e. whale-watching companies, researchers, members of the public who sufficiently describe the entanglement and species involved) are included in an entanglement database of known-species identifications (hereafter referred to as 'model data'). Records lacking supporting species identification evidence are categorized as 'unidentified whale' cases, hereafter referred to as 'novel data'). All sighting locations for model data and novel data entanglements are shown in Fig. 1.

The RF machine-learning method, using classification trees (Breiman 2001a,b, Liaw & Wiener 2002), was used to evaluate if known species entanglements (model data) could be accurately classified to species via cross-validation. Once an accurate species ID classifier is created, it is used to classify unidentified large-whale entanglements (novel data) to species. Variables in the RF model included geographic location, season, observation year, sea surface temperature (SST), water depth, an annual multivariate El Niño index, and the type of fishing gear (Table 1). All analyses were performed in the R programming environment, version 3.2.3 (R Development Core Team 2017), using the R package randomForest, version 4.6-12 (Liaw & Wiener 2002). Known-species entanglements (n = 199) documented from 2007 to 2016 (Carretta et al. 2013, 2014, 2015, 2016b) served as model data, from which a classification tree RF was generated. For 35 entanglement cases lacking a species identification (novel data), the variable characteristics associated with these cases were used to classify the species from constructed RF model trees. Details on the RF model parameters and variables used to construct it are summarized below and in Table 1.

Table 1. Variables tested and used in the random forest large-whale entanglement species-identification model. Variables used in the final ID model are in **bold**

Variable name	Description	Type
Interaction.Type	'Unidentified', 'Pot/trap fishery', 'Net fishery'	Categorical
Depth	Water depth (m)	Numeric
LAT	Latitude in decimal degrees	Numeric
LON	Longitude in decimal degrees	Numeric
Day.of.Year	Consecutive day of calendar year, range: 1–365	Integer
MEI.mean	Multivariate El Niño index (mean of 12 bimonthly values for each calendar year)	Numeric
SST	Sea surface temperature (°C)	Numeric
Year	Year of observation	Categorical

Variables used in RF model

'Interaction.Type'

Most documented large-whale entanglements along the US west coast result from fishing gear interactions that include pot/trap fisheries, gillnets, marine debris, and unidentified fisheries (Carretta et al. 2016b). The ability to identify entanglement sources depends upon the level of detail provided by reporting parties and opportunities for whale disentanglement teams to approach the animals. Some entanglement cases are linked to specific fisheries (i.e. 'California Dungeness crab pot'), based on identification of permit tag numbers on buoys associated with the entangled line. Most cases can only be categorized to a generic entanglement category such as 'pot/trap fishery', 'gillnet', or 'unknown fishery interaction' due to the opportunistic nature of reporting and lack of recovered gear (Carretta et al. 2016b). For the purposes of creating an entanglement species model, I treated the variable Interaction.Type as a categorical variable, with values limited to the generic categories 'pot/trap fishery', 'net fishery', and 'unknown fishery interaction'.

Latitude ('LAT'), longitude ('LON'), and water depth ('Depth')

Specific latitude and longitude coordinates of entangled whales were used when available, but such locations were not always recorded because reports and narratives reflect opportunistic sightings (e.g. 'entangled whale seen halfway between Catalina Island and mainland'). In those cases where entanglement narratives lacked latitude and longitude coordinates, there was enough information (e.g. '3 miles [5 km] offshore of Point Loma, San Diego') to infer approximate locations and assign latitude/longitude coordinates. Water depth (in meters) was interpolated for latitude and longitude point data using a geographic information system (GIS) with a world ocean depth raster in ArcGIS software, version 10.4.1. Some depths were assigned a value of zero because they involved entanglements extremely close to shore or beach-stranded animals where GIS water depth interpolations resulted in positive values above sea level.

'SST'

SST data were obtained for each entanglement record from archived data at NOAA's National Data Buoy Center (www.ndbc.noaa.gov/obs.shtml). SST data were obtained from the nearest buoy location to the entanglement and were based on the noon-time temperature for that day.

'Year' and 'MEI.mean'

The calendar year ('Year') of the observed entanglement was included as a categorical RF model variable. In addition to Year, a multivariate El Niño index variable was included to serve as a measure of the broad-scale oceanographic conditions along the US west coast in a given year. The 'MEI.mean' was calculated for each calendar year as the annual mean of 12 bimonthly (2× a month) values obtained from NOAA's Earth System Research Laboratory (NOAA 2017).

Season ('Day.of.Year')

The seasonality of large-whale entanglements varies by species; thus the sequential calendar day of the year (Day.of.Year) was included as a candidate

variable. Day.of.Year was used instead of calendar month, as it represents a finer measure of seasonality. Simultaneous use of both month and Day.of.Year variables is not recommended, as they are highly correlated, which can negatively impact classification accuracy of RF models (Strobl et al. 2008).

RF model construction and cross-validation

The RF model consists of classification trees, since the response is 'Species', a category to be classified. Classification trees are recursive partitioning algorithms. Random subsets of variables (default = $\sqrt{n}$ where n equals the number of variables) are selected at each tree node and the variable that results in the greatest variance reduction of the response is used to split the data into successive daughter nodes. Such variable splits continue until all observations in each terminal node contain the same response variable value or the terminal nodes each contain only a single sample. Each classification tree is built from a bootstrap sample of model data entanglements and those model data omitted from construction of individual RF trees are referred to as 'out-of-bag' (OOB) data. Due to bootstrap sampling with replacement, OOB data represent approximately 1/3 of all data (Efron & Tibshirani 1997). Evaluation of the RF model is based on classification accuracy, based on how often cross-validated OOB model data are correctly classified. The OOB data are introduced to constructed RF trees and species classifications are made for all OOB data, based on variable characteristics of the OOB data. The number of RF trees (n = 500 in this study) is based on the approximate number of RF trees required to return an asymptotic OOB error rate. Cross-validated species classifications for OOB data are summarized as a confusion matrix that includes the number of correctly and incorrectly classified cases by species (see Table 2). Only species for which there were at least 2 documented entanglements were included in the analysis, due to the need for model data cross-validation for each species.

I optimized the RF model by exhaustively searching for the number and combination of variables that maximized OOB correct classification rates for a RF of 500 trees. This strategy was implemented by randomly selecting subsets of all 8 candidate variables, ranging from 2 (the minimum required) to all 8 variables, and recording the OOB correct classification rate for each variable combination. The OOB correct classification rates for the optimized RF model were compared to correct classification rates expected by chance when all model data cases are randomly assigned a species in proportions equal to the observed entanglements (i.e. permutation of the response variable 'Species'). This was done 1000 times to generate a null distribution of correct classification rates. The 1-tailed probability of observing the correct classification rate from OOB model data was calculated as the observed fraction of null distribution correct classification rates greater than or equal to the observed correct classification rate (Fig. 2).

RF offers many tuning parameters for model evaluation. The major ones are: maximum tree depth, number of variables tested at each node, and number of forest trees. These parameters were assessed during model-building and the RF model that was used in this study ultimately included trees grown to full extent and the default number of variables considered for splitting at each node, or $\sqrt{n}$ variables.

Variable importance for the optimized RF model was assessed by permuting variables individually, running a RF model with the permuted variable, and comparing OOB correct classification rates between the RF model run with and without permutation. Negligible declines in classification accuracy with permutation indicate that a given variable is no more important than random noise in predicting species identifications. Conversely, a large decline in classification accuracy indicates that the permuted variable is informative. Variable importance was quantified as a 'permutation cost', which is equal to the number of additional OOB entanglement cases misclassified when a given variable was permuted.

Application of RF model to novel data

The RF model with the lowest OOB classification error rate was applied to 35 novel data entanglement cases lacking a species identification. For each novel data case, a species assignment is generated, based on the consensus predictions of all 500 RF trees (also referred to as the plurality vote; Svetnik et al. 2003). For each novel data case, the number of trees that classify a given species varies from a minimum of zero to the number of trees in the RF. The distribution of species classifications over all 500 RF trees is analogous to a species probability assignment. For example, a RF of 500 trees constructed from model data consisting of 6 species, when applied to a novel data case where the species is unknown, might yield the following classifications: 'Species.1' = 300 trees; 'Species.2' = 100 trees; 'Species.3' = 50 trees; 'Species.4' = 25 trees; 'Species.5' = 25 trees; and 'Species.6' = zero

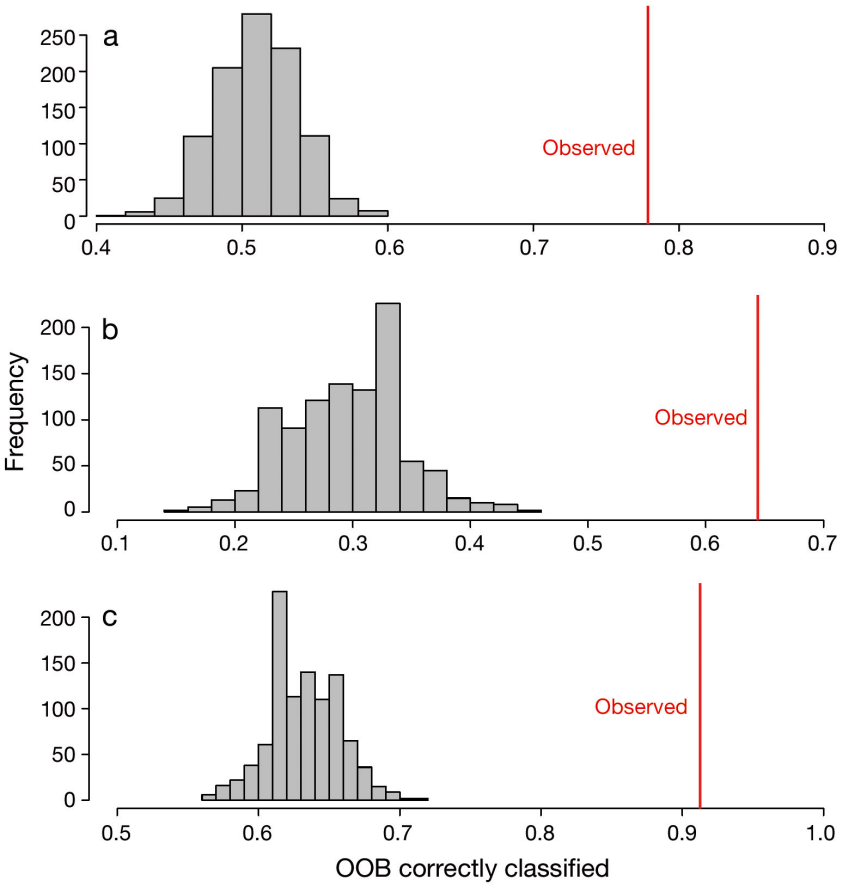

Fig. 2. Expected (null) and observed correct classification rates for cross-validated out-of-bag (OOB) model data. Expected values are based on permuting the response variable 'Species' 1000 times. This is equivalent to random assignment of a species to each model data observation based on observed species proportions. Observed correct classification rates from the random forest model are shown as a vertical red line for (a) all species combined, (b) gray whale, and (c) humpback whale. The probability that observed correct classification rates were less than or equal to null distribution correct classification rates was < 0.001 in all cases

trees. The assigned species in this novel data example is 'Species.1' (300/500 trees = plurality vote) and the probability of assignment to Species 1–6 are 0.60, 0.20, 0.10, 0.05, 0.05, and 0, respectively.

RESULTS

The RF model that minimized model data OOB error rates included 5 of 8 variables evaluated (Day.of.Year + Interaction.Type + LAT + SST + Year) and correctly classified the species in 78% (155/199) of model data cases (Tables 2 & 3). This correct classification rate for all 6 species combined was significantly higher ($p < 0.001$, permutation test) than the rate expected by chance (49%). Humpback whale cases were classified correctly as *Megaptera novaeangliae* 91% of the time, which was significantly higher ($p < 0.001$) than the 63% rate expected by chance. Correct classification (as *Eschrichtius robustus*) of gray whale cases (64%) was significantly higher ($p < 0.001$) than the 29% rate expected by chance. None of the minke (n = 2), blue (n = 5), or fin whale (n = 5) model data cases were correctly classified to species (*Balaenoptera acutorostrata*, *B. musculus*, and *B. physalus*, respectively). Poor classification accuracy for minke, blue, and fin whale cases is not unexpected, given that these species collectively represent only 6% of all model data cases and share many of the same variable attributes as humpback and gray whale records. Both sperm whale cases were correctly classified to species *Physeter macrocephalus* (see 'Discussion'). The 3 most important variables in the RF model were Day.of.Year + Interaction.Type + LAT, based on a comparison of correct classification rates of OOB model data using intact versus permuted versions of each variable (Table 3). Day.of.Year was the most important variable, based on 15 additional misclassified OOB cases when this variable was permuted. The next most important variable was Interaction.Type, with 12 additional misclassified cases, followed by LAT, with 9 additional misclassified cases.

Differences in the documented types of fishing gear entangling humpback and gray whales were evident (Table 4). Entanglements in net fishery gear were relatively rare for humpback whales (7/126 = 5%), compared with gray whales (14/59 = 24%). Entanglements in pot or trap gear were greater for humpback (67/126 = 53%) and gray whales (23/59 = 39%) compared to net gear. The fraction of entanglements where the gear type could not be identified was similar (~40%) for humpbacks (52/126) and gray whales (22/59). Differences in gear types between the 2 species may reflect multiple factors, including the spatial/temporal overlap of each species with different fisheries and possible observation biases in the ability to detect one gear type versus another. For example, monofilament gillnet entanglements are more difficult to detect at a distance than pot/trap gear entanglements; the latter usually include highly visible buoys trailing behind the whale.

Table 2. Random forest confusion matrix and correct classification rates for cross-validated out-of-bag (OOB) large-whale entanglement cases of known species. Rows represent known species and columns represent number of classifications of each species. The overall correct classification rate for OOB entanglement cases was 0.78, or 155 of 199 model data cases. The last column shows expected correct classification rates under the condition of permuting the response variable ('Species'). This is equivalent to a null distribution of OOB correct classification rates where all variables lack predictive value

	Minke	Blue	Fin	Gray	Humpback	Sperm	—Correctly Classified— Observed	Expected
Minke	0	0	0	1	0	1	0	0.01
Blue	0	0	0	0	5	0	0	0.02
Fin	0	0	0	2	3	0	0	0.02
Gray	0	0	0	38	21	0	0.64	0.29
Humpback	0	1	0	10	115	0	0.91	0.63
Sperm	0	0	0	0	0	2	1.00	0.01

Table 3. Variable importance as measured by the decrease in classification accuracy when each variable is permuted. Variables appear in increasing order of importance, where the cost of permutation is the decrease in the number of correctly classified out-of-bag (OOB) cases. Permuting the variable Day.of.Year had the largest cost (8%) to classification accuracy, resulting in 15 fewer correct classifications than a random forest model with all variables intact. BA = minke whale, BM = blue whale, BP = fin whale, ER = gray whale, MN = humpback whale, PM = sperm whale. See Table 1 for description of variables

Permuted variable	OOB% correct	Number of cases correctly classified: All species	BA	BM	BP	ER	MN	PM
None	0.779	155	0	0	0	38	115	2
SST	0.749	149	0	0	0	35	112	2
Year	0.734	146	0	0	0	36	108	2
LAT	0.734	146	0	0	0	35	111	0
Interaction.Type	0.719	143	0	0	0	33	108	2
Day.of.Year	0.704	140	0	0	0	30	108	2

Species classifications for 35 unidentified novel data cases included 24 humpback whales and 11 gray whales (Table 5). These classifications are based on the plurality vote of 500 RF trees. For example, novel data case #1 in Table 5 shows that the overall classification was gray whale, based on 86% of forest trees assigning this species. No novel data cases resulted in an overall classification of minke, blue, fin, or sperm whale, but most novel cases include a small percentage of RF trees assigning these species. Despite the lack of minke, blue, fin, or sperm whale plurality vote classifications, the proportion of trees predicting each species is analogous to a species assignment probability, where higher values imply greater confidence. For example, novel case #20 in Table 5 has the following assignment probabilities for minke, blue, fin, gray, humpback, and sperm whales, respectively: 0.002, 0.002, 0.01, 0.042, 0.944, and 0.00. Thus, the assigned species is humpback whale with a 94% probability. However, all 6 species assignment probabilities can be used to prorate this novel data case. One alternative to accepting species classifications based on the plurality votes is to sum individual species classification probabilities over all 35 novel data cases. This yields fractional species classifications, resulting in 0.218 minke, 0.462 blue, 0.77 fin, 11.97 gray, 21.5 humpback, and 0.078 sperm whale entanglements (Table 5). This approach yields approximately the same number of gray and humpback entanglements as the plurality vote approach, but it does a better job of representing the rare species classes by assigning them some small probability of occurrence, which is otherwise zero with the plurality vote results. The uncertainty of the plurality vote classifications can also be expressed as the range of summed species classifications for each of the 500 individual RF trees. For example, summing the species classifications from tree #1 of the RF model results in the following classifications for the 35 novel data cases: 4 fin whales, 11 gray whales, 20 humpback whales, and zero classifications for the remaining species. Tree #500 yields 1 fin whale, 9 gray whale, 24 humpback whale, and 1 sperm whale classification. Confidence intervals (95%) for all species classifications were calculated by summing species classifications individually for all 500 RF trees and identifying the 2.5th and 97.5th percentiles of the sums over all 35

Table 4. Number of large-whale entanglement cases documented in different gear types

	Net fishery	Pot/trap fishery	Unidentified fishery
Minke	2	0	0
Blue	0	3	2
Fin	0	0	5
Gray	14	23	22
Humpback	7	67	52
Sperm	2	0	0
Unidentified	5	8	21

Table 5. Random forest (RF) species classifications for novel data cases. Columns 2–6 represent variables used in the RF model. Values in species columns are the fraction of RF trees resulting in a given species classification. The overall classification for a novel data case is based on the plurality vote of all RF trees and appears in the last column. BA = minke whale, BM = blue whale, BP = fin whale, ER = gray whale, MN = humpback whale, PM = sperm whale. See Table 1 for description of variables

Case	Day.of. Year	Year	Interaction.Type	SST	LAT	BA	BM	BP	ER	MN	PM	Classification
1	97	2007	Pot/trap fishery	9.6	38.28	0.02	0	0.01	0.86	0.11	0	ER
2	165	2007	Pot/trap fishery	11	37.01	0.004	0	0.002	0.35	0.644	0	MN
3	184	2007	Pot/trap fishery	13.1	39.44	0.002	0	0.072	0.4	0.526	0	MN
4	231	2008	Net fishery	13.9	37.70	0.048	0.002	0.014	0.232	0.702	0.002	MN
5	239	2008	Pot/trap fishery	16.6	37.61	0.008	0	0.012	0.13	0.842	0.008	MN
6	245	2008	Pot/trap fishery	19.5	34.26	0.002	0.012	0.022	0.148	0.81	0.006	MN
7	78	2009	Unidentified fishery interaction	10.6	39.13	0.024	0	0.048	0.78	0.148	0	ER
8	193	2009	Net fishery	17.9	33.46	0.012	0.002	0.114	0.502	0.348	0.022	ER
9	83	2010	Unidentified fishery interaction	10.1	38.70	0.03	0	0.022	0.82	0.128	0	ER
10	122	2010	Unidentified fishery interaction	11.4	34	0.02	0	0.046	0.638	0.292	0.004	ER
11	187	2010	Unidentified fishery interaction	16.6	34.33	0.014	0.002	0.092	0.436	0.434	0.022	ER
12	282	2010	Unidentified fishery interaction	12.3	37.69	0	0	0.026	0.066	0.906	0.002	MN
13	197	2011	Unidentified fishery interaction	8.7	36.80	0	0	0.024	0.502	0.474	0	ER
14	248	2011	Pot/trap fishery	11.5	44.61	0	0.002	0.008	0.268	0.722	0	MN
15	176	2012	Pot/trap fishery	10.8	38.26	0	0	0.046	0.212	0.742	0	MN
16	235	2013	Net fishery	9.8	42.72	0.002	0	0.006	0.796	0.194	0.002	ER
17	63	2014	Pot/trap fishery	10.3	45.33	0.002	0	0	0.694	0.304	0	ER
18	106	2014	Unidentified fishery interaction	8.2	42.04	0.004	0	0	0.848	0.148	0	ER
19	156	2014	Pot/trap fishery	19.6	33.41	0	0.124	0.03	0.154	0.692	0	MN
20	269	2014	Unidentified fishery interaction	18.6	36.99	0.002	0.002	0.01	0.042	0.944	0	MN
21	60	2015	Unidentified fishery interaction	15.5	34.14	0	0	0.014	0.808	0.178	0	ER
22	123	2015	Unidentified fishery interaction	13.7	36.58	0.004	0.002	0	0.148	0.846	0	MN
23	129	2015	Unidentified fishery interaction	9.9	39.42	0	0	0	0.388	0.612	0	MN
24	148	2015	Net fishery	16.7	34.15	0.01	0.008	0.014	0.468	0.5	0	MN
25	237	2015	Unidentified fishery interaction	22.2	33.61	0	0.11	0.084	0.34	0.466	0	MN
26	239	2015	Unidentified fishery interaction	19.3	34.43	0	0.004	0.004	0.174	0.818	0	MN
27	270	2015	Net fishery	23.2	34.04	0	0.01	0.014	0.462	0.514	0	MN
28	311	2015	Unidentified fishery interaction	14.5	37.70	0	0.016	0.002	0.016	0.966	0	MN
29	325	2015	Unidentified fishery interaction	17.9	34.14	0.004	0.002	0.022	0.086	0.882	0.004	MN
30	331	2015	Unidentified fishery interaction	14.4	37.40	0.006	0.008	0	0.006	0.974	0.006	MN
31	119	2016	Unidentified fishery interaction	10.4	37.61	0	0.006	0	0.098	0.896	0	MN
32	134	2016	Unidentified fishery interaction	13.6	36.16	0	0.008	0	0.086	0.906	0	MN
33	136	2016	Unidentified fishery interaction	13.3	37.26	0	0.012	0.002	0.004	0.982	0	MN
34	144	2016	Unidentified fishery interaction	13.1	36.87	0	0	0	0.004	0.996	0	MN
35	272	2016	Unidentified fishery interaction	13.2	37.97	0	0.13	0.01	0.004	0.856	0	MN
Sum of individual species classification probabilities						0.22	0.46	0.77	11.97	21.5	0.078	

novel data cases. The resulting 95% species classification intervals for all 500 RF trees were: 0–2 minke whales, 0–2 blue whales, 0–3 fin whales, 6–17 gray whales, 16–27 humpback whales, and 0–1 sperm whales.

DISCUSSION

Nearly all of the known-species entanglement cases (92%) involved humpback and gray whales, 2 species that tend to utilize the California Current in different seasons and which are documented in net and pot/trap gear at different rates. High rates of correct species classification from the RF model are largely due to differences in seasonal occurrence of gray and humpback whales, proportions of entanglements involving net versus pot/trap gear, and the locations of the observed entanglements (Table 3). These differences are reflected by the identification of Day.of.Year, Interaction.Type, and LAT as the 3 most important variables in terms of their numerical contribution to correct classification rates (Table 3). Low classification accuracy for minke, blue, and fin whale model data cases is expected, given that these cases comprise only 6% of the observations.

The correct classification of both sperm whale model data cases was initially surprising because

they represent only 1% of model data cases and such minor response classes are usually misclassified at a nearly 100% rate. Both sperm whale entanglements occurred in the same gillnet fishing set and thus, cannot be considered independent events because they involved whales from the same social group entangled at the same time and location. All 5 model data variables, Year + Day.of.Year + Interaction.Type + MEI.mean + LAT, are identical for the 2 sperm whale entanglements, so there is a certainty of each sample ending up in the same terminal node of a fully grown classification tree (terminal nodes contain a single sample or response class, the default for classification). Due to bootstrap sampling with replacement of model data during tree construction, 1 sperm whale case has approximately a 2/3 probability of being used for tree construction and the other case has a 1/3 probability to serve as an OOB sample to be cross-validated (Efron & Tibshirani 1997). When the 2 cases are split between tree construction and OOB sample roles, the OOB sample will be assigned to the terminal node occupied by the first sperm whale case, because the variables are identical for each. This represents a special case of overfitting, which could be addressed by excluding the 2 sperm whale entanglements from analysis. However, the value of including these cases is that a RF data model lacking sperm whale entanglements would assign a zero risk of such entanglements in the novel data, which is known *a priori* to be untrue.

Despite poor classification accuracy for a few species with small sample sizes, their inclusion in the RF model is worthwhile because fractional estimates of entanglements can be produced for the novel data, despite the lack of any plurality vote assignments for these minor species. The classification accuracy for humpback and gray whales is, however, encouraging, in terms of prorating unknown species entanglement cases, the majority of which should comprise these 2 species. The importance of accurately assigning unknown cases to species can be considered as a form of risk management. For humpback and gray whale populations along the US west coast, there is a greater penalty for misclassifying a humpback entanglement. This is because humpbacks are less abundant than gray whales (estimated population sizes ~ 2000 and 20000, respectively) and humpbacks have lower allowable anthropogenic injury and mortality thresholds (potential biological removal or PBR; Wade 1998) under the Marine Mammal Protection Act. Current PBR levels for each population are 11 humpbacks versus 624 gray whales (Carretta et al. 2017).

The variable Day.of.Year was identified as the most important predictor variable, based on the greatest permutation cost to correct classification rates, but the context of variable importance is worth discussion. Algorithms such as RFs are designed to simultaneously handle many predictors and automatically deal with interactions between variables (Breiman 2001a,b). However, variable importance in the context of RF usually measures the effect on classification accuracy of permuting a single variable at a time. Some methods used to assess RF variable importance, such as *rfPermute* (Archer 2016), include statistical p-values for each variable. This is a useful tool for considering variables for model inclusion. However, analysts may be tempted to arbitrarily eliminate candidate variables that do not meet default p-value thresholds ($p < 0.05$). Such an approach may unnecessarily exclude multiple non-significant predictors whose collective classification power is superior to a smaller set of significant predictors (Breiman 2001a,b). It is recommended that analysts consider wider inclusion of candidate variables in RF models and examine cross-validated correct classification rates under different suites of variable numbers and combinations.

Species classifications for novel data could also be obtained via simple proration: multiplying observed model data species proportions by the number of novel cases (n = 35). This results in the following number of estimated entanglements for unidentified cases: 0.01 × 35 = 0.35 minke whales, 0.025 × 35 = 0.875 blue whales, 0.025 × 35 = 0.88 fin whales, 0.296 × 35 = 10.4 gray whales, 0.63 × 35 = 22 humpback whales, and 0.01 × 35 = 0.35 sperm whales. The RF model resulted in similar species classifications (0.218 minke, 0.462 blue, 0.77 fin, 11.97 gray, 21.5 humpback, and 0.078 sperm whales). The similarity in species classifications using simple proration and the RF model suggests that the 35 novel data cases may reflect an unbiased sample of the known-species model data observations. However, it is unknown whether or not the model data are representative of all large-whale entanglements. For example, gray whales generally occur closer to shore, compared to other species. This may introduce a positive detection bias for gray whales in the model data, as they may be more likely to be detected and reported from observers on shore or whale-watching vessels. Additionally, recreational vessel traffic is generally concentrated closer to shore, which would amplify this bias. If a positive gray-whale bias exists, the model data may represent an underestimation of other species' entanglements as a fraction of total

entanglements. While the simple proration is easy to implement, it is crude and forfeits potential insights into predictor variables that may be related to entanglement risk. However, if a suitable species identification model cannot be generated using RF or some other method, then at a minimum, unidentified cases should be prorated to fully account for entanglement risks to all species.

The RF species assignment approach described here has applications to other wildlife studies, particularly transect surveys, where a non-trivial fraction of detections may lack species identifications: raptors (Andersen et al. 1985), seabirds (Piatt et al. 2011), large whales (Barlow & Forney 2007), and sea turtles (Seminoff et al. 2014). When unidentified detections are not prorated to species, they are often omitted from analyses and can result in underestimates of animal abundance. Other applications may include species proration of unidentified bycatch in commercial fisheries and species assignments of large-whale vessel strikes.

Acknowledgements. This work was improved by the comments and constructive criticism of Jeff Moore, Karin Forney, and 3 anonymous reviewers. Thanks are also extended to members of the public, and US west coast stranding coordinators and researchers who have reported whale entanglements over the years. Collected data on whale entanglements are summarized and archived by the National Oceanic and Atmospheric Administration (NOAA). Monica DeAngelis, Justin Greenman, Kristin Wilkinson, Dan Lawson, Justin Viezbicke, and Sarah Wilkin made it possible to gather all of the large-whale entanglement data into one useful database.

LITERATURE CITED

Andersen DE, Rongstad OJ, Mytton WR (1985) Line transect analysis of raptor abundance along roads. Wildl Soc Bull 13:533–539

Archer E (2016) *rfPermute*: estimate permutation p-values for random forest importance metrics. R package version 2.1.5. https://cran.r-project.org/web/packages/rfPermute/rfPermute.pdf

Barlow J, Forney KA (2007) Abundance and population density of cetaceans in the California Current ecosystem. Fish Bull 105:509–526

Bradford AL, Weller DW, Ivashchenko YV, Burdin AM, Brownell RL Jr (2009) Anthropogenic scarring of western gray whales (*Eschrichtius robustus*). Mar Mamm Sci 25: 161–175

Breiman L (2001a) Random forests. Mach Learn 45:5–32

Breiman L (2001b) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231

Carretta JV, Wilkin S, Muto MM, Wilkinson K (2013) Sources of human-related injury and mortality for U.S. Pacific West Coast marine mammal stock assessments, 2007–2011. NOAA Tech Memo NOAA-TM-NMFS-SWFSC-514

Carretta JV, Wilkin SM, Muto MM, Wilkinson K, Rusin J (2014) Sources of human-related injury and mortality for U.S. Pacific West Coast marine mammal stock assessments, 2008–2012. NOAA Tech Memo NOAA-TM-NMFS-SWFSC-533

Carretta JV, Muto MM, Wilkin S, Greenman J and others (2015) Sources of human-related injury and mortality for U.S. Pacific West Coast marine mammal stock assessments, 2009–2013. US Department of Commerce, NOAA Tech Memo NOAA-TM-NMFS-SWFSC-548

Carretta JV, Danil K, Chivers SJ, Weller DW and others (2016a) Recovery rates of bottlenose dolphin (*Tursiops truncatus*) carcasses estimated from stranding and survival rate data. Mar Mamm Sci 32:349–362

Carretta JV, Muto MM, Wilkin S, Greenman J and others (2016b) Sources of human-related injury and mortality for U.S. Pacific West Coast marine mammal stock assessments, 2010–2014. US Department of Commerce, NOAA Tech Memo NOAA-TM-NMFS-SWFSC-554

Carretta JV, Forney KA, Oleson EM, Weller DW and others (2017) U.S. Pacific marine mammal stock assessments: 2016. US Department of Commerce, NOAA Tech Memo NOAA-TM-NMFS-SWFSC-577

Cassoff RM, Moore KM, McLellan WA, Barco SG, Rotstein DS, Moore MJ (2011) Lethal entanglement in baleen whales. Dis Aquat Org 96:175–185

Efron B, Tibshirani R (1997) Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 92: 548–560

Groom C, Coughran D (2012) Entanglements of baleen whales off the coast of Western Australia between 1982 and 2010: patterns of occurrence, outcomes and management responses. Pac Conserv Biol 18:203–214

Knowlton AR, Kraus SD (2001) Mortality and serious injury of northern right whales (*Eubalaena glacialis*) in the western North Atlantic Ocean. J Cetacean Res Manag Spec Issue 2:193–208

Knowlton AR, Hamilton PK, Marx MK, Pettis HM, Kraus SD (2012) Monitoring North Atlantic right whale *Eubalaena glacialis* entanglement rates: a 30 yr retrospective. Mar Ecol Prog Ser 466:293–302

Kraus SD, Brown MW, Caswell H, Clark CW and others (2005) North Atlantic right whales in crisis. Science 309: 561–562

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

Meÿer MA, Best PB, Anderson-Reade MD, Cliff G, Dudley SFJ, Kirkman SP (2011) Trends and interventions in large whale entanglement along the South African coast. Afr J Mar Sci 33:429–439

Moore MJ (2014) How we all kill whales. ICES J Mar Sci 71: 760–763

Muto MM, Helker VT, Angliss RP, Allen BA and others (2016) Alaska marine mammal stock assessments, 2015. US Department of Commerce, NOAA Tech Memo NMFS-AFSC-323

NOAA (2017) Earth System Research Laboratory. Multivariate ENSO Index. https://www.esrl.noaa.gov/psd/enso/mei/table.html

Piatt JF, Arimitsu M, Drew G, Madison EN, Bodkin J, Romano MD (2011) Status and trend of the Kittlitz's Murrelet (*Brachyramphus brevirostris*) in Glacier Bay, Alaska. Mar Ornithol 39:65–75

Prado JHF, Secchi ER, Kinas PG (2013) Mark-recapture of the endangered franciscana dolphin (*Pontoporia blainvillei*) killed in gillnet fisheries to estimate past bycatch from time series of stranded carcasses in southern Brazil. Ecol Indic 32:35–41

R Development Core Team (2017) R: a language and environment for statistical computing. https://www.R-project.org

Read AJ, Drinker P, Northridge S (2006) Bycatch of marine mammals in U.S. and global fisheries. Conserv Biol 20: 163–169

Seminoff JA, Eguchi T, Carretta J, Allen CD and others (2014) Loggerhead sea turtle abundance at a foraging hotspot in the eastern Pacific Ocean: implications for at-sea conservation. Endang Species Res 24:207–220

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9:307

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958

van der Hoop J, Corkeron P, Moore M (2017) Entanglement is a costly life-history stage in large whales. Ecol Evol 7: 92–106

Wade PR (1998) Calculating limits to the allowable human-caused mortality of cetaceans and pinnipeds. Mar Mamm Sci 14:1–37

Waring GT, Josephson E, Maze-Foley K, Rosel PE (2016) U.S. Atlantic and Gulf of Mexico marine mammal stock assessments – 2015. NOAA Tech Memo NMFS-NE-238

Wells RS, Allen JB, Lovewell G, Gorzelany J, Delynn RE, Fauquier DA, Barros NB (2015) Carcass-recovery rates for resident bottlenose dolphins in Sarasota Bay, Florida. Mar Mamm Sci 31:355–368

Williams R, Gero S, Bejder L, Calambokidis J and others (2011) Underestimating the damage: interpreting cetacean carcass recoveries in the context of the Deepwater Horizon/BP incident. Conserv Lett 4:228–233

Editorial responsibility: Robert Harcourt, Sydney, New South Wales, Australia

Submitted: December 7, 2017; Accepted: April 17, 2018
Proofs received from author(s): June 4, 2018