The R-Package *SeriousInjury* uses Random Forest (RF) classification trees to assess injury severity of large whale entanglements and vessel strikes. Models are built using the R-Package *rfPermute*.

NOAA assesses anthropogenic injuries and deaths under the Marine Mammal Protection Act through a Serious Injury Policy and Procedural Directive. Injuries are defined as 'Non-Serious' or 'Serious', where the latter is defined as "*any injury that is more likely than not to result in mortality, or any injury that presents a greater than 50 percent chance of death to a marine mammal*".

NOAA is reviewing the Serious Injury policy and procedures for large whales with new data and assessing a new Random Forest method to estimate individual probabilities of a health decline, death, or recovery for entanglements and vessel strikes. The major proposed change is that current serious injury procedures requiring biologists to assess a series of conditions, *e.g.,* "Was entangling gear constricting or loose?", can be automated using a RF model. Model covariates are derived from key words and phrases in injury narratives known to be good predictors of non-serious vs serious injuries. Methods, R functions, and application examples for large whale data are summarized in the R-Package *SeriousInjury* and are summarized in this document.

```
To install the latest SeriousInjury version from GitHub:
# make sure you have devtools installed
if (!require('devtools')) install.packages('devtools')

# install from GitHub
devtools::install_github('JimCarretta/SeriousInjury')

# installing SeriousInjury will also install rfPermute
```

Data & Injury Narratives

The *SeriousInjury* package includes five data frames: WhaleData, data.entangle, data.vessel, data.test.entangle, and data.test.vessel.

'**WhaleData**' is raw data for whale injury cases. Injury descriptions are in the 'Narrative' field and assessed health status in the 'Health.status' field.

'**data.entangle**' and '**data.vessel**' are known-outcome ("DEAD.DECLINE" or "RECOVERED") entanglements and vessel strike cases used to build RF models. Model covariates are generated with the function InjuryCovariates(). These data exclude cases where human intervention to remove entanglements occurred.

'**data.test.entangle**' and '**data.test.vessel**' include cases with 'Health.status' = "UNKNOWN" and are used with the predict() function and the RF objects 'ModelEntanglement' and 'ModelVessel' to assign cases to "DEAD.DECLINE" or "RECOVERED".

Example 'Narrative' from which model covariates are derived. Key words and phrases in the narrative that are coded as presence / absence covariates include 'cyamids', 'fluke', 'peduncle', 'grey skin', 'poor'.

"Entanglement injuries at fluke insertions and peduncle with associated cyamids at injured areas and on head. Grey skin and overal poor appearance."

Covariates

Covariates are derived from injury narratives with the function 'InjuryCovariates()'. Code to define, maintain, and extract covariates from narratives is found in the R-script InjuryCovariates.R. Covariates are defined below.

anchored - evidence a whale was anchored or immobilized by entangling material or gear. Narrative mentions inability to dive or swim, may refer to a heavily-weighted whale with multiple pots/traps impeding normal movement.

calf.juv - narrative includes reference to an injured calf or juvenile or that the injury involves the mother of a dependent calf.

constricting - evidence of a constricting entanglement, including line cutting into whale, wrapped tightly around body or flippers.

decline - narrative includes evidence of a health decline, such as the presence of cyamids, emaciation, discolored skin, deformities caused by a chronic entanglement or severe vessel strike incident.

extensive.severe - a severe injury that can include amputation or necrosis of body parts due to a chronic entanglement or acute vessel strike injury.

fluke.peduncle - includes reference of entanglement or vessel strike injury that involves the tail, flukes, or peduncle.

gear.free - evidence the whale freed itself from entangling material. Typically involves a whale resighted at a later date than the initial entanglement observation.

head - Narrative indicates that the head, mouth, or blowhole was involved in the entanglement or vessel strike injury.

healing - Narrative refers to healing or healed wounds.

laceration.deep - Narrative includes reference to deep laceration resulting from vessel strike or entanglement. May include reference to blubber or muscle layers.

laceration.shallow - Narrative includes reference to shallow or superficial lacerations.

pectoral - Narrative includes involvement of pectoral flipper or 'fins' in entanglement or vessel strike.

swim.dive - Evidence that the whale is swimming, feeding, or diving normally.

trailing - Was the whale trailing gear or other entangling material?

VessSpd - Vessel Speed, coded as a factor with 3 possible states: unknown = VSpdUnk, slow = VSpdSlow, fast = VSpdFast. Based on speed references or inferences from 'Narrative'. Speeds <=10 kts are considered slow, >10 kts are fast.

VessSz - Vessel Size, coded as a factor with 3 possible states: unknown = VSzUnk, small = VSzSmall, large = VSzLarge. Based on size references or inferences from 'Narrative'. Sizes <=65 ft are considered 'small', unless the vessel is much larger than whale. Sizes > 65 ft are considered 'large'.

wraps.multi - Narrative includes reference to whale with multiple wraps of line or gear around body or appendage.

wraps.no - Narrative includes reference to a whale without any wraps of line or gear around body or appendage.

Injury Models

Two models are used in the package SeriousInjury, an entanglement and a vessel strike model. Each is based on n = 1,000 RF classification trees.

Figure 1. Example tree used to classify whale injuries as serious or non-serious. Data are based on known-outcome entanglement and vessel strike cases, where a known-outcome is a documented death, health decline or recovery. Health declines are considered serious injuries and recoveries are considered non-serious.

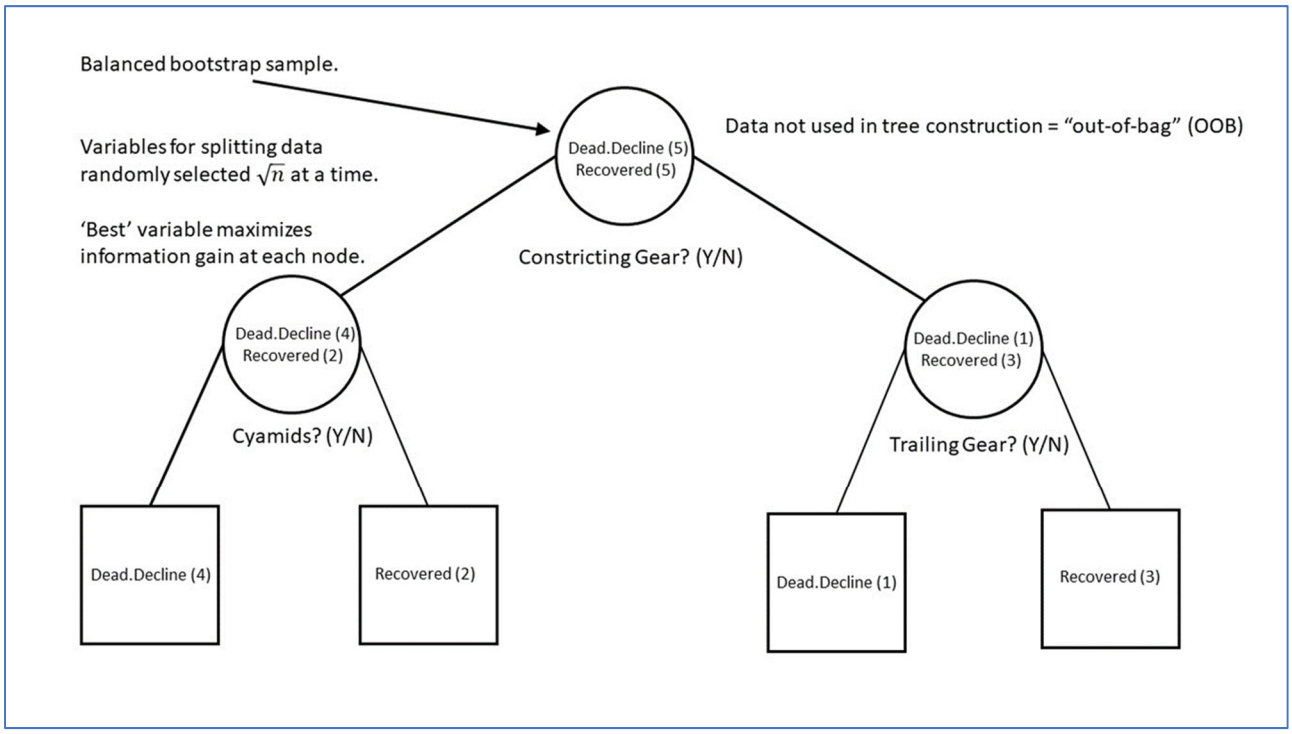
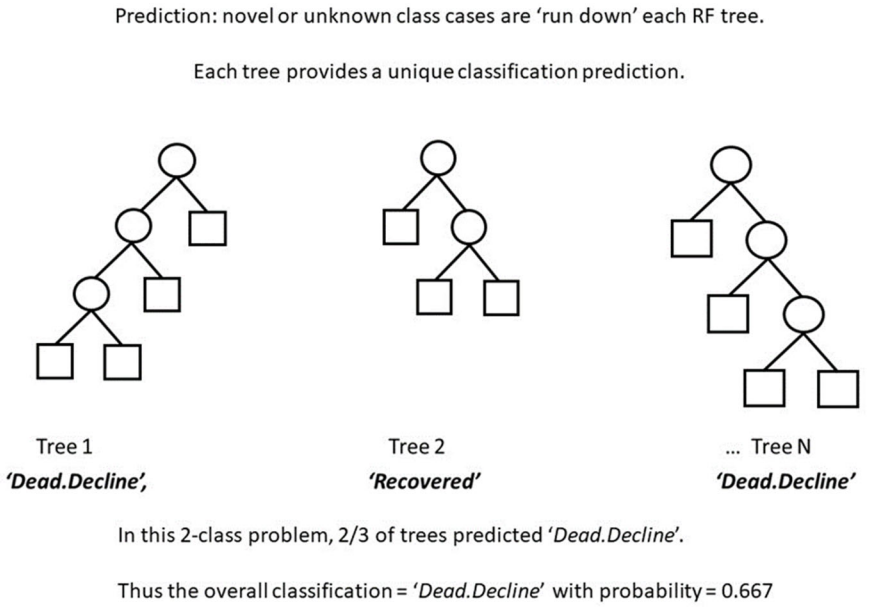


Figure 2. Models consist of multiple bootstrap trees (a random forest) used to classify 'out-of-bag' (OOB) or novel cases. Samples not used in individual tree construction are considered OOB and are used to assess model accuracy through cross-validation. Novel cases represent new data or cases not included in models, for which health status is unknown. The fraction of trees 'voting' for a particular class represents the probability of that case belonging to the class Dead.Decline or Recovered.



The entanglement (ModelEntangle) and vessel strike (ModelVessel) models are objects of class rfPermute. They include known-outcome entanglement and vessel strike injury cases, included as the data frames data.entangle and data.vessel in the R-Package SeriousInjury.

```
# Create randomForest model (using R-Package rfPermute) using known-outcome
entanglement strike data

# set.seed for reproducibility
set.seed(123)

# how many RF trees to build
size.RF = 1000

########### Entanglement Model

# covariates included in ModelEntangle

entangle.covariates = which(names(data.entangle)%in%c("anchored",
"calf.juv", "constricting", "decline", "extensive.severe", "fluke.peduncle",
"gear.free", "head", "healing", "laceration.deep", "laceration.shallow",
"pectoral", "swim.dive", "trailing", "wraps.multi", "wraps.no"))

# balance sample size for each class; we are equally-interested in correctly
#  predicting non-serious and serious injuries

sampsize = balancedSampsize(data.entangle$Health.status)

# RF Entanglement model

ModelEntanglement = rfPermute(data.entangle$Health.status ~ .,
 data.entangle[,entangle.covariates], sampsize=sampsize, ntree=size.RF,
  replace=FALSE, importance=TRUE, proximity=TRUE)

# RF Entanglement model Confusion Matrix
ModelEntanglement

################## Vessel Model

# covariates included in ModelVessel
vessel.covariates = which(names(data.vessel)%in%c("calf.juv", "decline",
"extensive.severe", "fluke.peduncle", "head", "healing",
"laceration.deep","laceration.shallow", "pectoral", "VessSpd", "VessSz"))

# balance sample size for each health class; we are equally-interested in
correctly
#  predicting non-serious and serious injuries

sampsize = balancedSampsize(data.vessel$Health.status)

# RF Vessel Strike model

ModelVessel = rfPermute(data.vessel$Health.status ~ .,
 data.vessel[,c(vessel.covariates)], sampsize=sampsize, ntree=size.RF,
  replace=FALSE, importance=TRUE, proximity=TRUE)
```

```
# RF Vessel Strike model Confusion Matrix
ModelVessel
```

Predictions

Use existing RF models to predict probability of a death, health decline or recovery for cases where the outcome is unknown. Deaths and health declines are considered serious injuries and recoveries are non-serious. Both Dead.Decline and Recovered probabilities are estimated, based on the fraction of RF tree assignments to each class. A binary prediction (either Dead.Decline or Recovered) is also returned, based on the majority class assignment (>50% of trees). In case of ties, which are rare, the model randomly assigns a class.

```
# Code to estimate injury classes for entanglements with unknown outcomes.

# 'data.test.entangle' data frame with required field name 'Narrative' and
appended covariates.

head(data.test.entangle)

# Apply RF model ('ModelEntanglement') to data.test.entangle to generate
binary and probabilistic model predictions

majority.prediction <- predict(ModelEntanglement, data.test.entangle,
type='response')

prob.prediction <- predict(ModelEntanglement, data.test.entangle,
type='prob')

predictions.df <- cbind.data.frame(majority.prediction, prob.prediction,
data.test.entangle)

head(predictions.df)
```