# Species Estimation of Unidentified Bycatch Sea Turtles in the Indian Ocean using RandomForest

Yu Sato[1)], Takahito Masubuchi[1)], Atsuya Yamamoto[1)], Ayumi Shibano[1)], Miyuki Kanaiwa[1)], Kei Okamoto[2)], Daisuke Ochi[2)], and *Minoru Kanaiwa[1)]

## Summary

We attempted to classify unidentified sea turtles that were recorded as bycatch by scientific observers boarded on Japanese longline vessels in the IOTC area with using a random forest model. We constructed two models using only the IOTC area data, and combining the IOTC and the ICCAT data and compared the model performance. The both models showed high accuracy in species estimates.

Key words: *Sea turtle, Bycatch, Longline, Machine learning, Random Forest*

## Introduction

Sea turtle bycatch in the tuna longline fishery is one of the risk factors in the conservation of the sea turtle population. Therefore, the understanding of the amount of bycatch by species is essential as fundamental data, and records by the observer becomes the basic data for this issue. However, species identification cannot be made for all individuals because observer recording is done under the hazardous and severe environment of longline operations. Therefore, if species estimation can be made for unspecified individuals after these data have been aggregated, the understanding of the effects of bycatch will be more accurate.

Okamoto et. al. (2019) used random forest for species estimation of unspecified sea turtles caught by pelagic longline fisheries in the ICCAT areas and showed high performance. If the same approach can be extended to the IOTC area to perform species estimation of unspecified sea turtles, it would be beneficial for ecological risk assessment of those species. On the other hand, observer program in the IOTC area has a shorter history and fewer observations than in the ICCAT area. Therefore, it may be problematic to create a model using the IOTC area data alone, because of the lack of accuracy.

In particular, loggerheads and olive ridley turtles occupy a large proportion of sea

[1] Department of Marine Bioresources, Mie University, 1577 Kurima Machiya cho, Tsu, Mie 514-8507, Japan, kanaiwa@bio.mie-u.ac.jp

[2] Fisheries Resources Institute, National Research and Development Agency, Japan Fisheries Research and Education Agency. 5-7-1 Orido, Shimizu-ku, Shizuoka 424-8633, Japan

turtle bycatch, and it is difficult to identify the species from its morphological characteristics, and most of the individuals that have not been identified species are expected to be these two species.

In this study, we examined the accuracy of the model by comparing the results of two cases: one where only the IOTC area data were used to construct the model and estimate the species of unspecified sea turtles, and the other where the IOTC and ICCAT areas were combined in the model and the species of unspecified sea turtles in the IOTC area were estimated.

## Material and Method

### Operational data

Operational data in the IOTC Conventional Area and the ICCAT area were obtained by scientific observer program for the Japanese longline vessels between May 2010 and Febuary 2017 and January 1997 and October 2016, respectively. The data included operational year, month, latitude, longitude, sea surface temperature, hooks per baskets, observed hooks, the total catch of fish, and catch amount of each fish. In total 4,158 and 11,123 operations were recorded in the IOTC Conventional Area and the ICCAT area, respectively.

Data from the catches of loggerhead turtle (TTL) and olive ridley turtle (LKV) were used to construct the model, and unidentified individuals were used to estimate using the model. 5 TTLs, 50 ORTs and 54 unspecified individuals were fished in the IOTC area and 143 TTLs, 76 ORTs and 152 unspecified individuals were fished in the ICCAT area, respectively.

### Random Forest

To classify the species of unidentified sea turtles, the random forest (Breiman 2001) was applied. Two random forest models (RF models): the Indian Ocean RF model, and the Indian and Atlantic Ocean RF model, were conducted. The cut-off value for species discrimination was set at 1/2 according to the usual method. The number of parameter used in each tree, i.e. mtry, was set at 2 and the number of all tree, i.e. ntree, was set at 30,000.

### The RF model in IOTC area

In this model, only data from the IOTC area were used to estimate the species of unspecified sea turtles. The structure of the Indian Ocean RF model (I RF model) was below;
TTL or LKV~ yr + mon + lat + lon + hpb + obs_hks + total_fish + $\Sigma$ $\text{catch}_i$.
Here, yr, mon, lat, lon, hpb, obs_hks, total_fish and $\text{catch}_i$ represent operational year, month, latitude, longitude, hooks per baskets, number of observed hooks, total catch amount of all

species and the catch amount for each species, respectively. The species names used in the analysis were shown in Table 1.

### The RF model in IOTC and ICCAT area

In this model, we used both IOTC and ICCAT area data to estimate species of unspecified sea turtles. We removed latitude from the explanatory variables in the I RF model because of the different implications of latitude for the Atlantic and Indian Oceans. The structure of the Indian and Atlantic Ocean RF model (IA RF model) was similar with I RF model and the detail was below;

TTL or LKV~ yr + mon + lat + lon + hpb + obs_hks + total_fish + $\Sigma$ $\text{catch}_i$.

In the case of IA RF model, the definition of the species which captured was a little bit different, i.e. White marlin (*Tetrapturus albidus*) was treated as Striped marlin (*T. audax*). The list of the species is shown in Table 1.

### Model performance and variable importance

The accuracies of these RF models were assessed by 2 approaches: confusion matrix, species classification for the sea turtles with identified species.

## Result and discussion

### I RF model

In the confusion matrix, the error rate for TTL was 0.40 and for LKV was 0.02 (Table 2). This model was used to estimate the species of individuals identified the species caught in the IOTC area, and the miss estimation rates for both TTL and LKV were 0.00 (Table 3). The importance of the explanatory variables, in descending order of importance, were lon, BET, lat, hpb, ALB, total_fish and YFT (Fig. 1).

The species estimates for unspecified individuals caught in the IOTC area using this model are shown in Fig. 2, and were estimated as TTL 2 individuals and LKV 52 individuals.

### IA RF model

In the confusion matrix, the error rate for TTL was 0.088 and for LKV was 0.095 (Table 4). This model was used to estimate the species of individuals identified the species caught in the IOTC area, and the miss estimation rates for both TTL and LKV were 0.00 (Table 5). The importance of the explanatory variables, in descending order of importance, were lon, ALB, hpb, ULA, BET, yr and obs_hks (Fig. 3).

The species estimates for unspecified individuals caught in IOTC area using this model are shown in Fig. 4, and were estimated as TTL 2 individuals and LKV 52 individuals.

### The comparison of the results from both models

Both models were shown to provide highly accurate species estimates. There was no difference between the estimated results from both models in the IOTC area. The order of importance of the variables varied in detail, but generally the same variables dominated, with longitude, BET, and ALB in particular being important factors in species estimation.

In general, for the analysis of machine learning, the amount of data has a significant impact on the performance of such estimates using machine learning. In the future, if similar methods can be applied by compiling data from not only Japan but also from other countries, it will be possible to conduct species estimates with higher performance and to evaluate individuals with recorded species.

### References

Breiman, L. (2001). Random forests. Machine Learning 45(1): 5–32.

Okamoto, K., M. Kanaiwa and D. Ochi. (2019) Machine learning approach to estimate species composition of unidentified sea turtles that were recorded on the Japanese longline observer program. IOTC-2019-WPEB15-42.

Table 1 List of species used in the Indian and Atlantic Ocean random forest model.

| English name | Abbreviation | Scientific name |
|---|---|---|
| Albacore | ALB | *Thunnus alalunga* |
| Yellowfin tuna | YFT | *T. albacares* |
| Southern bluefin tuna* | SBT | *T. maccoyii* |
| Bigeye tuna | BET | *T. obesus* |
| Skipjack tuna** | SKJ | *Katsuwonus pelamis* |
| Shortbill spearfish | SSP | *Tetrapturus angustirostris* |
| Striped marlin*** | SPM | *T. audax* |
| Swordfish | SWO | *Xiphias gladius* |
| Unidentified lancetfishes | ULA | *Alepisaurus spp.* |
| Opah | OPA | *Lampris guttatus* |
| Common dolphinfish | DOL | *Coryphaena hippurus* |
| Escolar | ESC | *Lepidocybium flavobrunneum* |
| Oilfish | OIL | *Ruvettus pretiosus* |
| Wahoo | WAH | *Acanthocybium solandri* |
| Shortfin mako | SMA | *Isurus oxyrinchus* |
| Porbeagle | POR | *Lamna nasus* |
| Blue shark | BSH | *Prionace glauca* |
| Pelagic stingray | PES | *Pteroplatytrygon violacea* |

*substituted Pacific bluefin tuna

** used for the Indian Ocean random forest model only

***including White marlin

Table 2 Confusion matrix of the Indian Ocean RF model.

| | | Estimated species | | |
|---|---|---|---|---|
| | | LKV | TTL | Class.error |
| Actual species | LKV | 49 | 1 | 0.020 |
| | TTL | 2 | 3 | 0.400 |

Table 3 Estimated results using the Indian Ocean RF model and actual species.

| | | Estimated | | |
|---|---|---|---|---|
| | | LKV | TTL | Class.error |
| Actual | LKV | 50 | 0 | 0.00 |
| | TTL | 0 | 5 | 0.00 |

Table 4 Confusion matrix of The Indian and Atlantic Ocean RF model.

| | | Estimated species | | |
|---|---|---|---|---|
| | | LKV | TTL | Class.error |
| Actual species | LKV | 114 | 12 | 0.095 |
| | TTL | 13 | 135 | 0.088 |

Table 5 Estimated results using the India and Atlantic Ocean RF model and actual species.

| | | Estimated | | |
|---|---|---|---|---|
| | | LKV | TTL | Class.error |
| Actual | LKV | 50 | 0 | 0.00 |
| | TTL | 0 | 5 | 0.00 |

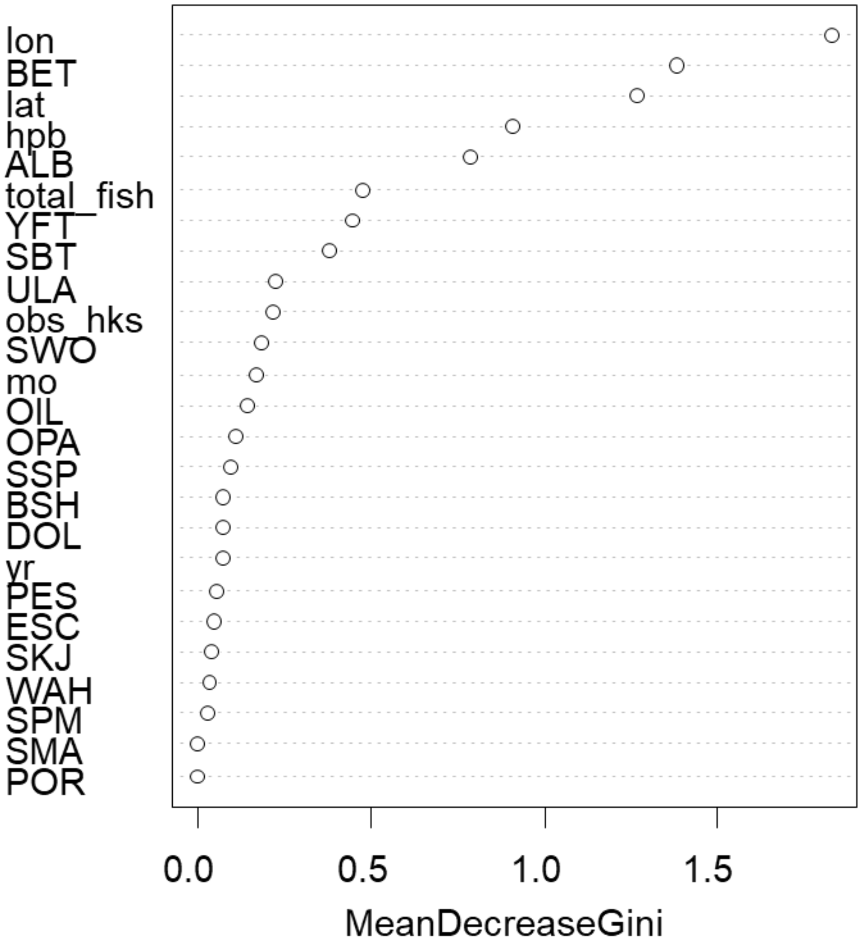

Fig.1 x axis indicate the importance, and y axis indicate each variable

.

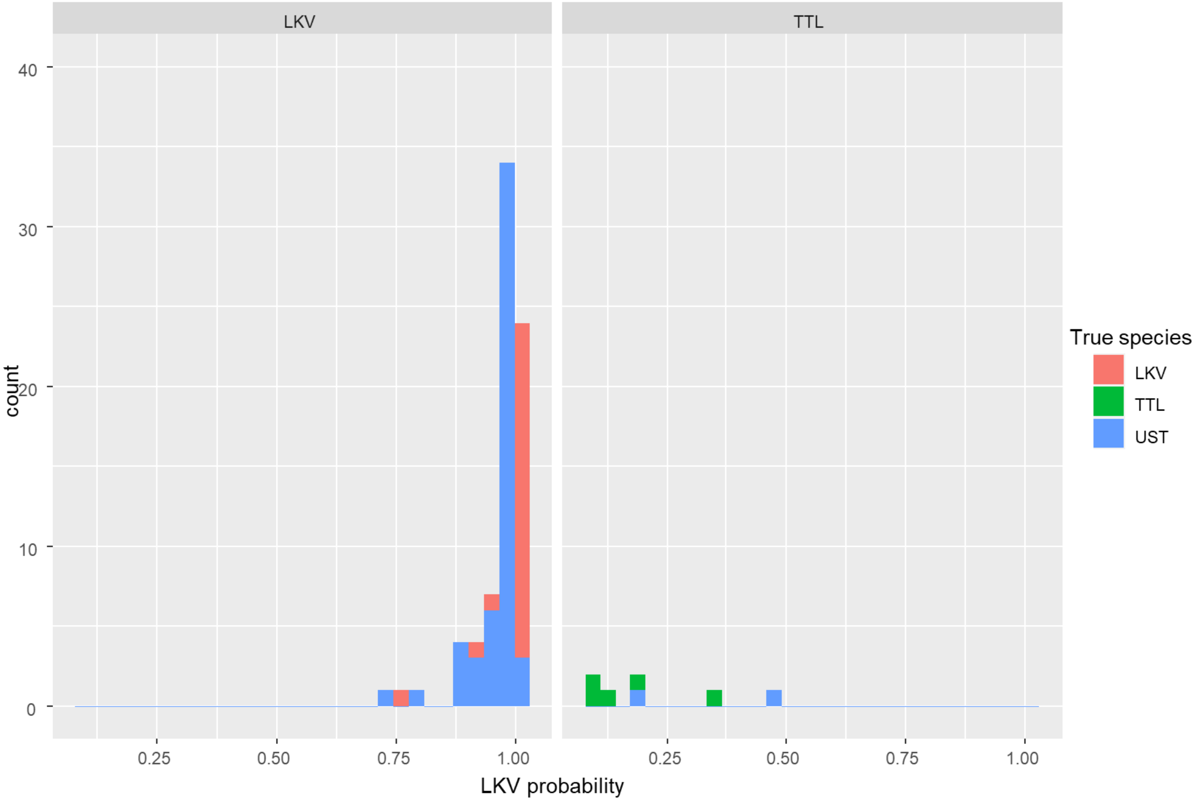

Fig 2 Histogram of estimated probability of LKV.

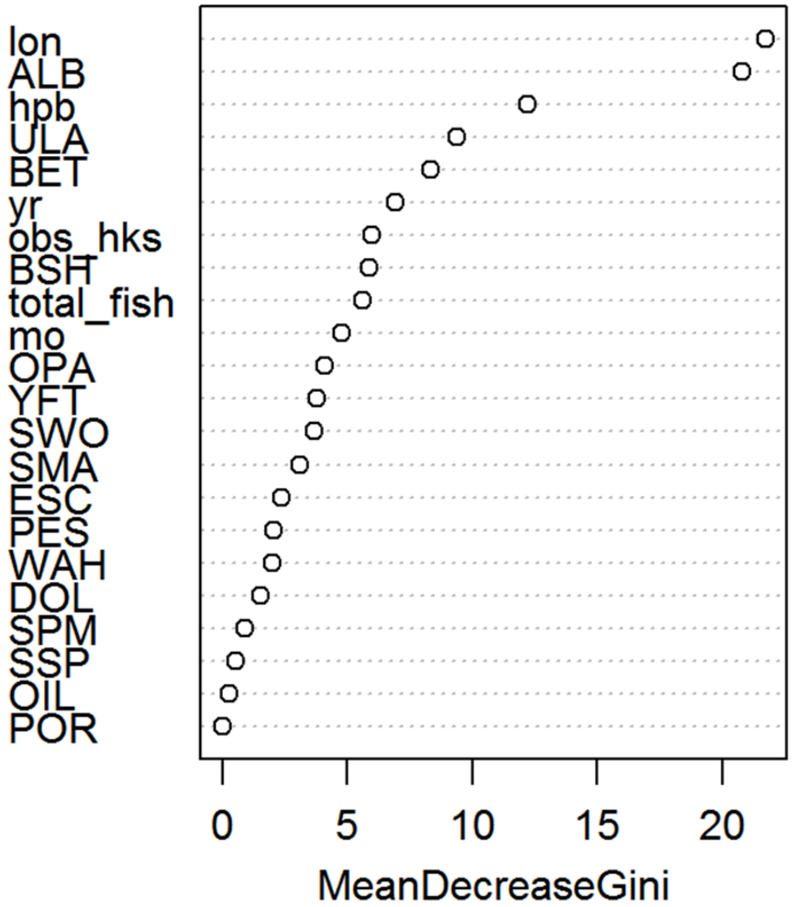

Fig.3 x axis indicate the importance, and y axis indicate each variable

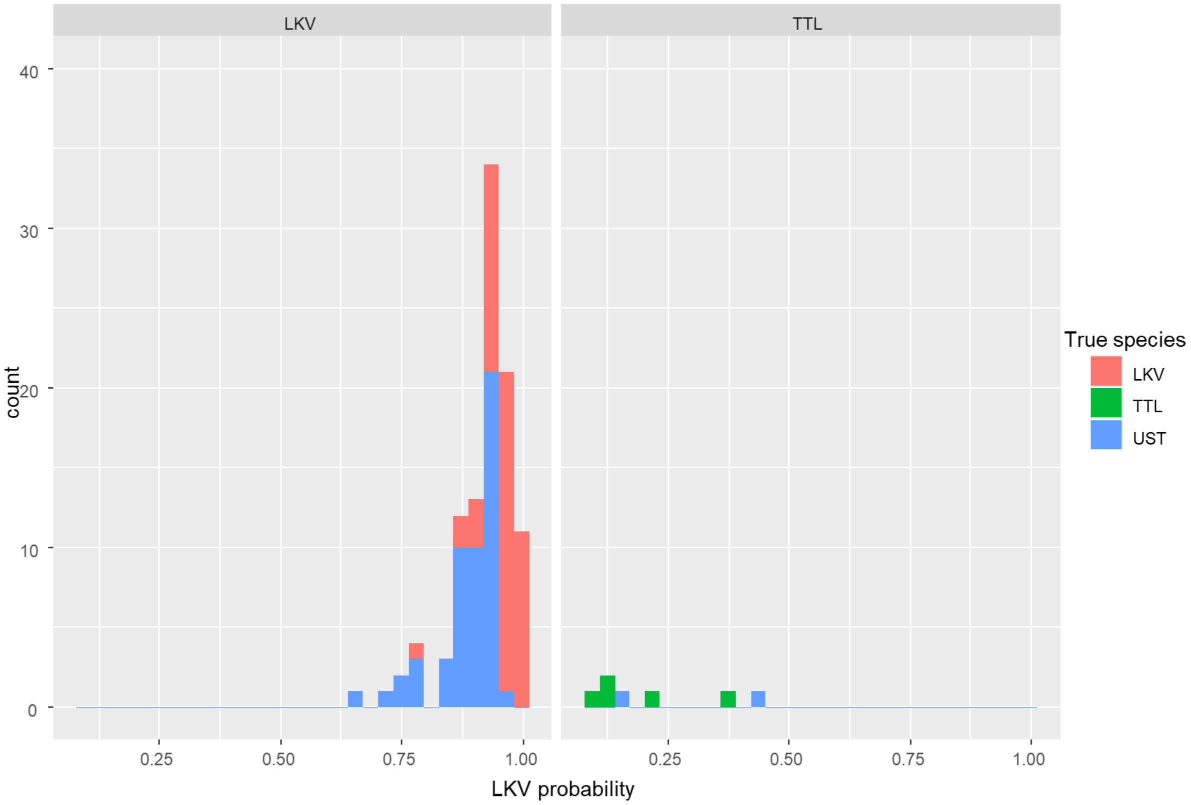

Fig 4 Histogram of estimated probability of LKV.