MACHINE LEARNING APPROACH TO ESTIMATE SPECIES COMPOSITION OF UNIDENTIFIED SEA TURTLES THAT WERE RECORDED ON THE JAPANESE LONGLINE OBSERVER PROGRAM

K. Okamoto[1], M. Kanaiwa[2], and D. Ochi[1]

SUMMARY

Unidentified species is the major source of uncertainties to evaluate the impact of bycatch on sea turtle populations, so we tried to estimate species composition of unidentified sea turtles from operational circumstance via machine learning approach. We used bycatch data from the Japanese scientific observer program, which includes 10,490 operations and catch records of 141 loggerheads, 75 olive ridleys, and 152 unidentified turtles. The random forest, which is a machine learning approaches, was conducted to estimate probability of the species identities (loggerhead or olive ridley). As training datasets, species-identified sea turtle bycatch number including set date, location, sea surface temperature and catch number of target and non-target species such as tunas, billfishes, other teleost fishes, sharks, and sea turtles. As a result, the probabilities of species identity were calculated. When the species was defined as identified (the probability larger than 0.7), the identified 111 turtles were identified as 16 loggerheads and 95 olive ridleys, and 41 could not be identified. We conclude that random forest approach will be helpful to improve the species estimation.

RÉSUMÉ

Les espèces non identifiées constituant la principale source d'incertitudes pour évaluer l'impact des prises accessoires sur les populations de tortues marines, nous avons essayé d'estimer la composition des espèces de tortues marines non identifiées à partir de circonstances opérationnelles via une approche d'apprentissage automatique. Nous avons utilisé les données sur les prises accessoires du programme d'observateurs scientifiques japonais, incluant 10.490 opérations et les registres de capture de 141 tortues caouannes, 75 tortues olivâtres et 152 tortues non identifiées. Le modèle de forêts aléatoires, qui est une approche d'apprentissage automatique, a été appliqué pour estimer la probabilité d'identification des espèces (tortues caouannes ou olivâtres). Les jeux de données utilisés pour la formation incluaient le nombre de prises accessoires de tortues de mer identifiées, la date, le lieu, la température de surface de la mer et le nombre d'espèces cibles et non ciblées capturées telles que les thonidés, les istiophoridés, les autres poissons téléostéens, les requins et les tortues de mer. Ensuite, les probabilités d'identité des espèces ont été calculées. Lorsque l'espèce a été définie comme identifiée (probabilité supérieure à 0,7), les 111 tortues identifiées ont été identifiées comme 16 tortues caouannes et 95 tortues olivâtres, et 41 n'ont pas pu être identifiées. Nous concluons qu'une approche de forêts aléatoires sera utile pour améliorer l'estimation des espèces.

RESUMEN

Las especies no identificadas son la principal fuente de incertidumbres para evaluar el impacto de la captura fortuita en las poblaciones de tortugas marinas, por lo que tratamos de estimar la composición de las especies de tortugas marinas no identificadas a partir de las circunstancias operativas mediante un enfoque de aprendizaje automático. Utilizamos los datos de captura fortuita del programa de observadores científicos japonés, que incluye 10.490 operaciones y registros de captura de 141 tortugas bobas, 75 tortugas golfinas y 152 tortugas no identificadas. El bosque aleatorio, que es un enfoque de aprendizaje automático, se llevó a cabo para estimar la probabilidad de las identidades de las especies (tortuga laúd o tortuga golfina). Los conjuntos de datos de formación incluían: el número de capturas fortuitas de tortugas marinas identificadas por especie, la fecha de la operación, la ubicación, la temperatura de la superficie del mar y el número de capturas de especies objetivo y no objetivo como los atunes, los

[1] National Research Institute of Far Seas Fisheries, Japan Fisheries Research and Education Agency, 5-7-1 Orido, Shimizu 424-8633, Japan, keiokamoto@affrc.go.jp
[2] Department of Marine Bioresources, Mie University, 1577 Kurima Machiya cho, Tsu, Mie 514-8507, Japan

istiofóridos, otros peces teleósteos, los tiburones y las tortugas marinas. A continuación, se calcularon las probabilidades de la identidad de la especie. Cuando se definió la especie como identificada (la probabilidad es mayor de 0,7), de las 111 tortugas identificadas 16 fueron identificadas como tortugas bobas, 95 como tortugas golfinas, y 41 no pudieron ser identificadas. Concluimos que el enfoque de bosque aleatorio será útil para mejorar la estimación de las especies.

KEYWORDS

Catch composition, sea turtle, by catch, longlining, machine learning, pelagic fisheries

1. Introduction

The catch and bycatch information by the Japanese longliners in the ICCAT Convention Area obtained through the scientific observer program are reported to the ICCAT secretariat based on the specified format. The summary of bycatch records for sea turtle have already been reported on the previous SC-ECO meeting (Minami et al. 2013; Okamoto et al. 2017). However, Okamoto et al. (2017) also touched that substantial unidentified turtle bycatch was also recorded, which is caused by poor weather condition or observer safety issues. The bycatch records without species identification are often having the problems when the species-specific bycatch risk assessments would be conducted, thus the species identification is needed in as lower taxonomic lank as possible. In this study, we examined the possibilities of classification of the species-unidentified sea turtle by a machine learning approach with using operational and catch data.

2. Materials and methods

In this study, operational data by the Japanese longliner in the ICCAT convention area collected by observer program including operation year, month, latitude, longitude, sea surface temperature, hooks per baskets and caught amount of each species were used. The dataset includes 10,490 operations and catch records of 141 loggerheads, 75 olive ridleys, and 152 unidentified turtles, whereas green, hawksbill and Kemp's ridley turtles were not observed through all the operations. The random forest (Breiman, 2001) was applied to classify the species of sea turtles, i.e., olive ridley turtle *Lepidochelys olivacea* (LKV) and loggerhead turtle *Caretta caretta* (TTL).

The structure of model was below;

LKV or TTL ~ Year + Month + Lat + Lon + SST + HPB + Obs_Hooks + TotalCatch + ΣCatch for each species.

Here, Year and Month are shown operational year and month, respectively. Lat and Lon are shown operational latitude and longitude, respectively. SST is shown sea surface temperature in operational point. HPB is shown hooks per baskets. Obs_Hooks is shown the number of observed hooks. TotalCatch is shown total catch of all species. Catch for each species are shown the amount of catch for species as shown in **Table 1**.

The followed procedure selected the optimal model setting.

1. The error ratio of identification was used to estimate an optimal number of variables randomly sampled as candidates at each split.
2. The mean decrease of Gini value was used to rank the importance of each variable.
3. Decrease the explanatory variable followed by order of the importance and remove unnecessary variables.
4. Check the convergence by the stability of error ratio and get the optimal number of trees to grow.

3. Results

Throughout the analysis, the probability of species identity for each unidentified turtle bycatch has been calculated. We defined the unidentified turtle identifiable correctly when the probability was larger than 0.7. As a result, we were able to identify 111 turtles (corresponds to 72.5% of all the number of unidentified-sea turtles) which were 16 loggerheads and 95 olive ridleys, and 41 (27.5%) were not identified (**Table 2**). The most important factor to identify them was latitude, followed by sea surface temperature and catch number of albacores. The least number of variables without losing the information were resulted in 32, thus the variables which are catch numbers of two species (Great barracuda and Porbeagle) were removed from the analyses.

4. Discussion

This study indicated that the random forest approach is useful to reclassify the unidentified turtles with using the operational and catch information. In addition, it was considered to be one of the representative tools for the effective usage of catch information obtained by the observer. The fact remains that the ideal strategy to identify the bycaught species is taking pictures on the deck to the extent possible. It will be, however, useful and final solutions to estimate the species. In future analyses, it could be available for the forecast of occurrence if key factors related to the occurrence of each sea turtle species are extracted.

References

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

Minami, H., Matsunaga, H., Inoue, Y., and Ochi, D. (2013). Bycatch distribution and standardized CPUE of sea turtle using data from Japanese scientific observer program of longline fishery in the Atlantic. Collect Vol. Sci. Pap. ICCAT 69(4): 1901–1909.

Okamoto, K., Ochi, D., and Oshima, K. (2017). Review of sea turtle bycatch data in the ICCAT convention area obtained through Japanese scientific observer program. Collect Vol. Sci. Pap. ICCAT 74(7): 3698–3713.

Table 1. List of species used in the analysis.

English name	Scientific name
Albacore	*Thunnus alalunga*
Yellowfin tuna	*Thunnus albacares*
Bigeye tuna	*Thunnus obesus*
Atlantic bluefin tuna	*Thunnus thynnus*
Shortbill spearfish	*Tetrapturus angustirostris*
Longbill spearfish	*Tetrapturus pfluegeri*
Swordfish	*Xiphias gladius*
Atlantic blue marlin	*Makaira nigricans*
Unidentified lancetfishes	*Alepisaurus* spp.
Longnose lancetfish	*Alepisaurus ferox*
Shortnose lancetfish	*Alepisaurus brevirostris*
Opah	*Lampris guttatus*
Great barracuda	*Sphyraena barracuda*
Dolphinfish	*Coryphaena hippurus*
Snake mackerel	*Gempylus serpens*
Escolar	*Lepidocybium flavobrunneum*
Oilfish	*Ruvettus pretiosus*
Wahoo	*Acanthocybium solandri*
Unidentified sunfishes	*Mola* spp.
Crocodile shark	*Pseudocarcharias kamoharai*
Bigeye thresher	*Alopias superciliosus*
Shortfin mako	*Isurus oxyrinchus*
Porbeagle	*Lamna nasus*
Blue shark	*Prionace glauca*
Pelagic stingray	*Pteroplatytrygon violacea*
Leatherback turtle	*Dermochelys coriacea*

Table 2. The number of sea turtles and identified rate before and after the analysis**.**

	Before the analysis	After the analysis			
Latitude	Unidentified species	Loggerhead	Olive ridley	Unidentified species	Identified rate
25N-65N	8	8	0	0	100%
10S-25N	134	8	89	37	72%
35S-10S	10	0	6	4	60%