

Assessing consistency between fisheries logbook and machine learning-derived VMS data for skipjack tuna fishing effort in the Western Sumatra Indian Ocean

Ridwan Nurzеха^{1*}, Jonson Lumban-Gaol¹, Syamsul Bahri Agus¹, and Al Fajar Alam²

¹Department of Marine Science and Technology, Faculty of Fisheries and Marine Science, IPB University, Bogor, West Java, Indonesia

²Ministry of Marine Affairs and Fisheries, Republic of Indonesia, Jakarta, Indonesia

Abstract. This study addresses the need for accurate fisheries data by assessing the consistency between fishing logbook records and fishing efforts derived from Vessel Monitoring System (VMS) data using machine learning in the Western Sumatra Indian Ocean. The objective of this study was to evaluate the reliability of the skipjack tuna (*Katsuwonus pelamis*) fishing effort data. We utilized VMS data from 2014-2023, processed through a *vmstofish* machine learning function, and compared it with logbook data from Pelabuhan Nizam Zachman Jakarta. The *vmstofish* function, utilizing the CatBoost model, demonstrated high effectiveness in detecting fishing effort, achieving a recall of 0.983 and an F1-score of 0.931, proving its validity as an alternative data source. Spatiotemporal analysis revealed a significant increase in perfect match rates between VMS-derived and logbook data from 2019-2023 (86.6%), as the impact of e-logbook implementation, indicating improved logbook data quality in recent years. This research provides a robust method for complementing and evaluating fisheries data, offering a more comprehensive understanding of fishing activities crucial for sustainable management, and contributing to blue economy initiatives.

Keywords: Machine learning, vessel monitoring system, skipjack tuna, Indian Ocean

1 Introduction

The fisheries sector plays a vital role in Indonesia's Blue Economy, serving as the primary protein source and a significant contributor to national income. Skipjack tuna (*Katsuwonus pelamis*) is a high-value commodity, particularly in the western Indian Ocean region of Sumatra. Based on the Badan Survei Statistik, skipjack tuna accounted for the highest contributor to Indonesian capture fisheries, totalling 1,411 tons, and ranked second in terms

* Corresponding author: ridwan.nurzеха@kcp.go.id

of economic value at 2.38 trillion rupiahs after squid [1]. Based on these statistics, it is important to ensure the sustainability of resources.

Fishery-dependent data sources, particularly Vessel Monitoring System (VMS) data and logbook records, have become widely used for calculating fishing efforts. VMS data provides high-resolution spatial and temporal data of fishing vessels through GPS tracking. Such data are abundant and useful for inferring fishing effort distribution; however, they do not directly reflect fishing effort unless complemented by additional metadata or analytical processing (e.g., behavior classification models) [2]. On the other hand, logbook data provide direct evidence of catch occurrence along with associated effort and location [3], making them biologically meaningful for indicating fishing effort. However, logbooks are often limited by inconsistencies in reporting, particularly in terms of spatial information and factors that reduce their accuracy [4].

Understanding the spatiotemporal patterns and comparative capabilities of both datasets is critical for improving skipjack habitat models, particularly in data-limited regions. Despite the growing use of both VMS and logbook data in fishery modelling, few studies have directly compared their data to improve data quality. The challenge with VMS data, however, is that it records all vessel activities (e.g., steaming, hauling, and fishing) [5], not just actual fishing events, necessitating further processing to identify fishing efforts.

Machine learning (ML) technology offers an innovative approach for analyzing large-scale spatiotemporal and environmental data, enabling the identification of patterns that influence fish habitat distribution and fishing activity. This study aims to bridge the gap between traditional logbook data and modern VMS data by developing and evaluating a machine learning model to accurately detect fishing efforts from VMS records.

Therefore, this study aimed to evaluate and compare the VMS and logbook data to evaluate the fishing effort for skipjack tuna in the Western Sumatra Indian Ocean. Specifically, we (1) assessed their spatial and temporal coverage, (2) detected the fishing effort from VMS data using *vmstofish* function, and (3) explored the comparison between them based on the fishing effort result. This research contributes to improving the logbook and VMS data for fisheries management by clarifying how different data types shape the model accuracy and spatial prediction.

2 Methods

2.1 Study area and data

This study focused on the high seas of the Western Sumatra Indian Ocean, specifically within the coordinates of 80° 00' 00" " E to 105° 00' 00" E longitude and 10° 00' 00" S to 5° 00' 00" N (**Fig. 1**). This area was selected based on the operational zones of Indonesian fishing vessels in the high seas, as indicated in the Indonesia National Report to the Scientific Committee of the IOTC [6]. The eastern boundary of the study area is defined by the Indonesian Fisheries Management Area (WPPNRI) 572, ensuring that the analysis focuses on vessels operating exclusively in the high seas as per government regulations.

The primary datasets used in this study include Vessel Monitoring System (VMS) data and daily fishing logbook data from the Nizam Zachman Jakarta Fishing Port. The VMS data, covering the period of 2014-2023, includes details on catch, species, and estimated fishing locations. Vessels using a large pelagic purse seine gear, commonly used for skipjack tuna, were selected for analysis.

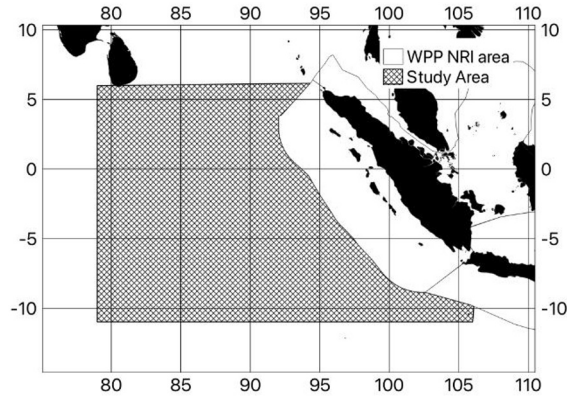


Fig. 1. Study area in Western Sumatra Indian Ocean.

2.2 VMS data processing

The raw VMS data contains records of all vessel movements, which do not directly refer to fishing efforts. To predict fishing efforts from the VMS data, a *vmstofish* function was developed. This function combines data filtering, processing, and a machine learning model. This function processed the VMS data by extracting additional data related to the vessel movement patterns, as shown in **Table 1**.

Table 1. Additional data generated from the raw data.

No	Model	Source
1	<i>Calculated distance</i>	Calculate the distance between data points using Haversine method
2	<i>Calculated heading</i>	Calculate the direction of vessel movement between points
3	<i>Calculated speed</i>	Calculate speed based on calculated distance and the difference in ping_time
4	<i>Heading Change</i>	Information related to changes in vessel heading
5	<i>Circular Pattern Intensity</i>	Calculation for detecting potential circular movements, which are characteristic of purse seine vessels
6	<i>Mean_distance_2h</i>	Average distance in the last 2 hours – used to eliminate anomalous data
7	<i>Mean_heading_2h</i>	Average heading in the last 2 hours – used to detect movement patterns
8	<i>Mean_speed_2h</i>	Used as an indicator of continuous slow vessel movement
9	<i>Nighttime_flag</i>	Filter data to ensure only nighttime data is selected, as it is potentially associated with fishing activity

VMS data from five vessels were visually interpreted to identify potential fishing efforts and filtered for short-duration events below 3 h to minimize false positives. These data were then used as training and test datasets, split at a 70:30 ratio. Several machine learning models have been evaluated for their ability to detect fishing effort, including Random Forest, XGBoost, CatBoost, Balanced Random Forest, EasyEnsemble, and Logistic Regression.

2.3 Logbook and VMS data consistency analysis

To assess the consistency between the logbook and the VMS-derived fishing effort, a spatiotemporal comparison was performed. The comparison utilized a monthly ratio coverage calculation within $0.5^\circ \times 0.5^\circ$ grid cells. The ratio (R) was calculated as the number of VMS-derived data points (C_{VMS}) divided by the number of logbook data points ($C_{Logbook}$) within each grid cell per month and was clamped at 1 if the ratio exceeded 1. The mean ratio across all grid cells per month and year was then computed to provide an overall consistency measure.

$$R = \begin{cases} \frac{C_{VMS}}{C_{Logbook}}, & \text{if } C_{Logbook} \neq 0 \\ 0, & \text{if } C_{Logbook} = 0 \end{cases}$$

$$R = \min(R, 1) \quad (1)$$

$$\text{Meanratio}(Y,M) = \frac{1}{n_{Y,M}} \sum_{i=1}^{n_{Y,M}} R \quad (2)$$

3 Results and discussion

3.1 Overview of fishing activity data

During the 2014-2023 period, logbook data recorded 32,659 fishing efforts from 349 vessels, whereas VMS data accumulated 15,661,196 records from 962 vessels. The substantial difference in data volume is attributed to the VMS recording all vessel movements at sea, whereas logbooks only record actual fishing activities. When compared spatially, the logbook data were mostly found in the southern part of the study area, whereas the VMS data showed a more uniform distribution in the area (**Fig. 2**). This is evident in the broader spatial distribution of the VMS data compared to the logbook data.

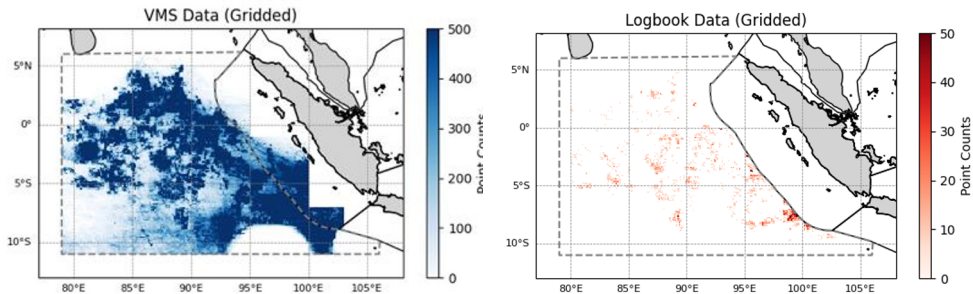


Fig. 2. Spatial distribution comparison between VMS data (left) and logbook data (right).

Based on the logbook data, approximately 49.919 tons of skipjack tuna were caught during 2014-2023, with the highest annual catch reported in 2022 at 19.732 tons (**Fig. 3**). A significant increase in both total catch and the number of operational vessels was observed from 2019 onwards. This increasing trend affected the implementation of the e-logbook system by the Ministry of Marine Affairs and Fisheries in 2018, leading to improved data quality and increased reporting from fishing vessels.

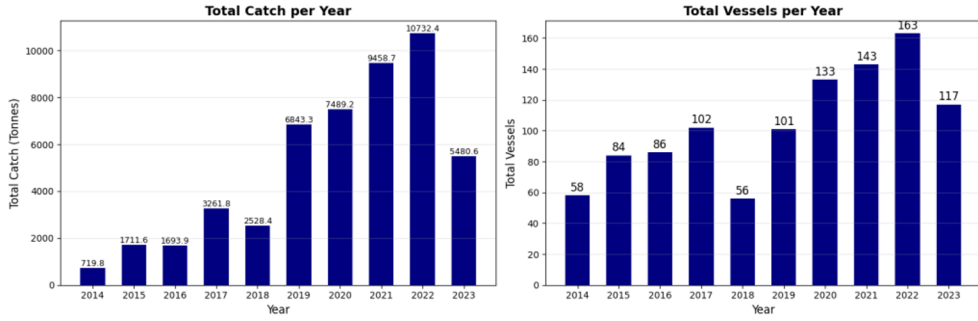


Fig. 3. Yearly trend of total catch and vessels from 2014-2023.

Spatially, fishing operations were initially concentrated in specific areas, particularly near Sunda Strait, during 2014-2016. However, from to 2019-2023, there was a noticeable increase in fishing grounds westward, which correlated with an increase in the number of operated vessels. Consistent fishing grounds were identified within the coordinates of 95° 00' 00" " E to 100° 00' 00" E longitude and 9° 00' 00" S to 4° 00' 00" N. The distance of fishing locations from Nizam Zachman Jakarta Port ranged from 225 to 1671 nautical miles, with an average of 1064 nautical miles and a median of 1101 nautical miles. High-intensity fishing hotspots were observed at distances of approximately 500, 700 nm, 1000-1200 nm, and 1400 nm from the port. Monthly variations in the farthest fishing distances were minimal, although December tended to have the smallest maximum distance, whereas September and October showed the highest.

3.2 Performance of the *vmstofish* function

The sample from the VMS data to be used as the training data consisted of 81.815 data points from five vessels during 2023. In the visual interpretation process, we considered several factors such as speed, heading, and movement patterns. The initial results of the interpretation of the data are shown in **Fig. 4**.

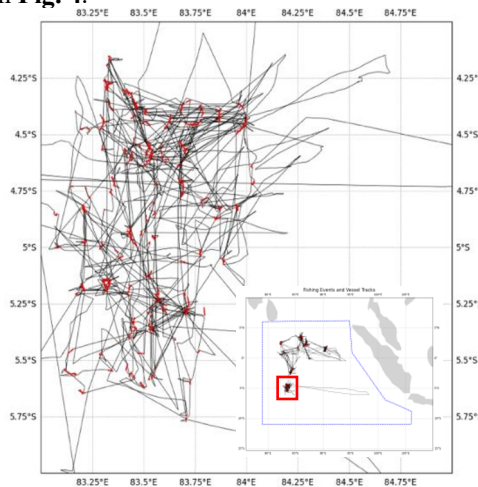


Fig. 4. Result of visual interpretation on detecting fishing effort from VMS data.

The *vmstofish* function resulted in the CatBoost model being the best performer for detecting fishing efforts (**Table 2**). The model demonstrated high accuracy (0.947), recall (0.983), and F1-score (0.931), making it highly effective in detecting actual fishing effort.

The low number of false positives and false negatives in the confusion matrix highlights the model’s balanced predictive capability.

Table 2. Machine learning models performances.

No	Model	Accuracy	Precision	Recall	F1-Score
1	<i>RandomForest</i>	0,947	0,897	0,964	0,929
2	<i>XGBoost</i>	0,946	0,885	0,979	0,929
3	<i>CatBoost</i>	0,947	0,883	0,983	0,931
4	<i>RandomForest</i>	0,947	0,885	0,981	0,931
5	<i>EasyEnsemble</i>	0,940	0,869	0,983	0,922
6	<i>LogisticRegression</i>	0,913	0,817	0,976	0,890

Based on the precision and recall rates, RandomForest, XGBoost, and CatBoost showed similar performances, as described in **Fig. 5**. Logistic Regression resulted in the lowest performance compared to the other models. Based on the confusion matrix in **Fig. 6**, XGBoost produced the lowest false positives, while CatBoost had the lowest false negatives. Overall, CatBoost produced the lowest false detection rate compared with the other models.

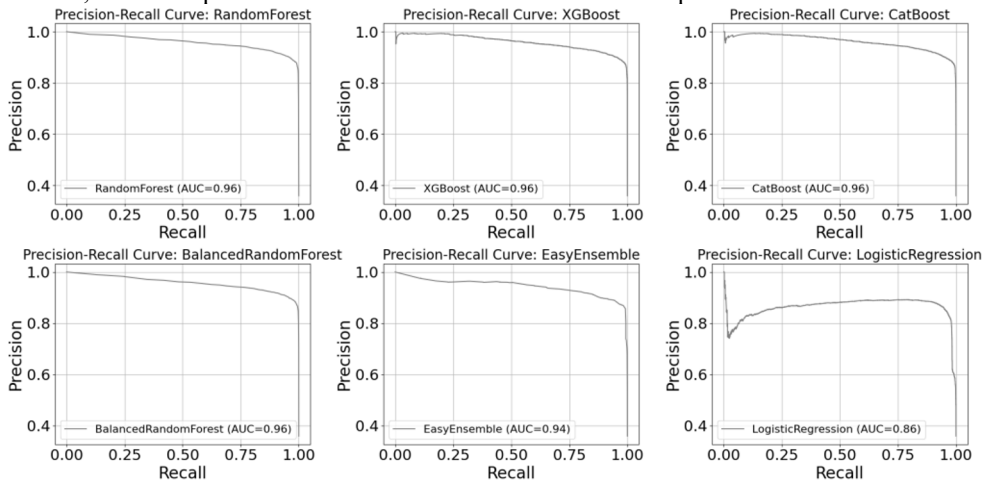


Fig. 5. Precision and recall from each model.

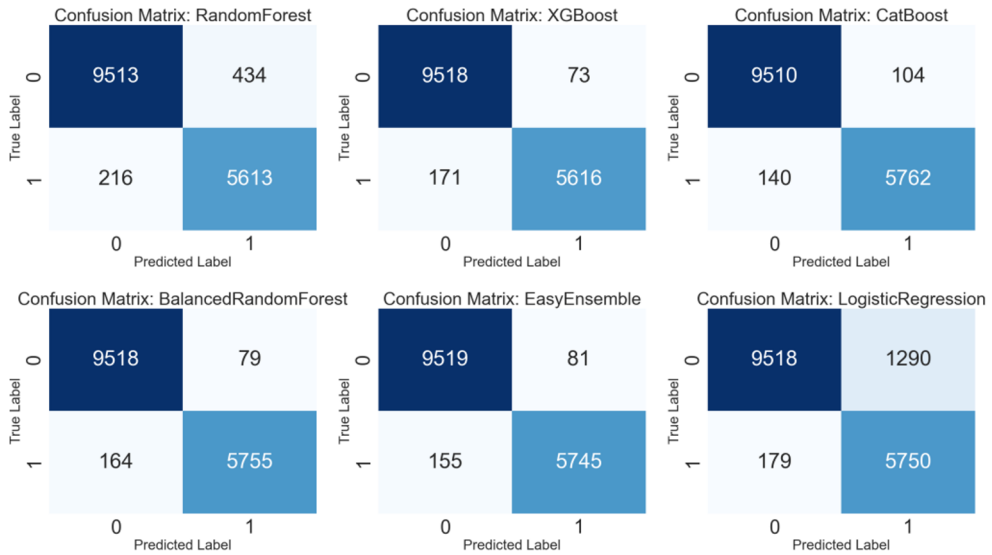


Fig. 6. Confusion matrix from each model.

The model was then applied to the entire VMS raw data set and compared with the logbook data, as shown in **Fig. 7**. Based on the initial analysis, both datasets showed similar seasonal patterns, but with different intensities (**Fig. 8**). The VMS-derived data showed a higher frequency of fishing efforts, leading to several conclusions. First, the model might overestimate fishing efforts compared with logbook data. Second, because the logbook data was a voluntary input from the vessel’s crews, it was possible that the data had unreported fishing efforts, which led to the logbook data being underestimated compared to the possible actual events [7].

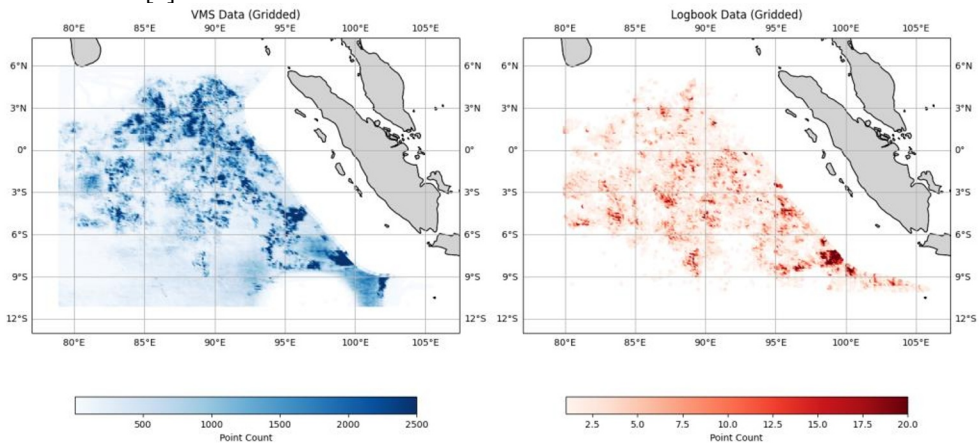


Fig. 7. Comparison between VMS-derived data (left) and logbook data (right).

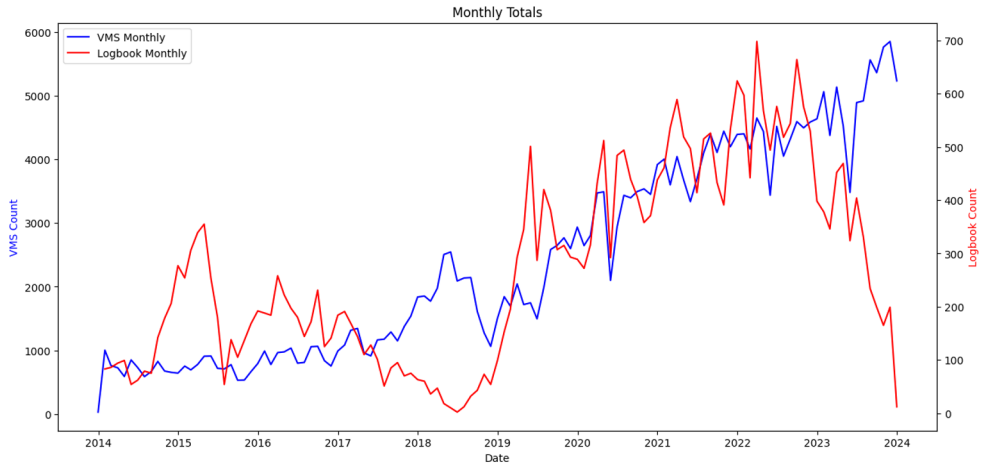


Fig. 8. Comparison between VMS-derived data (left) and logbook data (right).

3.3 Consistency analysis of logbook and VMS-derived data fishing effort

The spatiotemporal consistency analysis between the VMS-derived fishing effort and logbook data yielded a "perfect match rate" of 76% for the entire 2014-2023 period, with a temporal stability deviation of 0.3. A significant improvement in consistency was observed from 2019-2023, where the perfect match rate increased to 86.6% and the temporal deviation decreased to 0.05. This improvement is strongly correlated with the widespread implementation of e-logbooks [8], which led to better quality and more consistent reporting in logbook data from 2019 onwards (**Fig. 9**).

The high perfect match rate and low temporal deviation in recent years demonstrate that fishing effort data generated by the *vmstofish* function can serve as a reliable substitute for logbook data, particularly for periods or regions where logbook reporting may be less robust. Although spatial discrepancies remain, particularly north of the equator, where VMS activity appears higher than logbook entries, the overall consistency indicates that machine learning-processed VMS data can significantly enhance the resolution and reliability of fishing activity records. This enhanced dataset is crucial for accurate stock assessments and fisheries management. The consistently higher perfect match rate and lower temporal deviation from 2019-2023 further support the use of this period's data for training and evaluating habitat models and for future prediction capabilities. On the other hand, the improvement of the matching rate during the implementation of the e-logbook showed that the e-logbook significantly improved the quality of the logbook data overall, with more fishing events recorded.

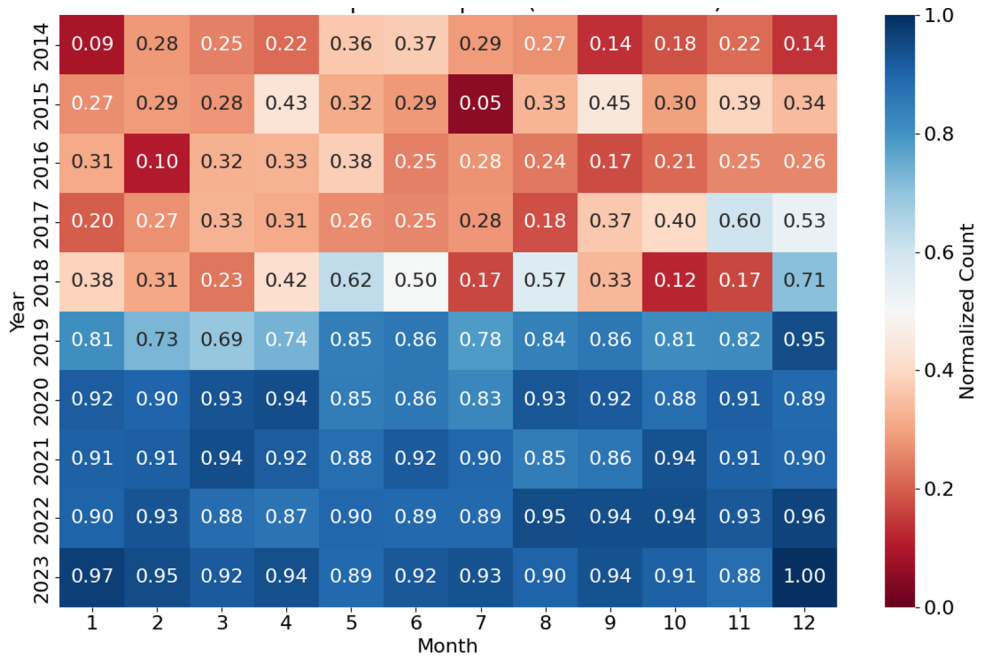


Fig. 9. Spatiotemporal analysis results.

3.4 Models limitation

Despite the strong performance of the *vmstofish* model, several limitations must be addressed. The training dataset was derived from only a limited number of vessels, which may lead to potential bias related to the operational behavior of fishing vessels. As a result, this function may be less accurate when applied to more vessels. Nevertheless, the high recall and F1-score resulting from the function model, along with the high correlation of spatiotemporal consistency between VMS-derived fishing effort and logbook data, were sufficiently representative of purse seine fishing activities in the study area.

3.5 Implications for fisheries management

The consistency between VMS-derived fishing efforts and logbook data, especially with the advancements brought about by e-logbook implementation, has significant implications for sustainable fisheries management in the Western Sumatra Indian Ocean. Accurate and high-resolution fishing effort data are fundamental for robust stock assessment models [9] and are critical for setting sustainable catch limits and designing effective management strategies. The ability of *vmstofish* to identify fishing activities from VMS data provides a powerful tool to complement or even substitute traditional logbook data, particularly in areas or periods of incomplete reporting. This study contributes to a more reliable and complete understanding of fishing pressure on skipjack tuna stocks, thereby supporting the objectives of a sustainable blue economy. The improved data quality allows for more precise spatial and temporal analyses of fishing grounds, enabling authorities to implement targeted conservation measures and adjust fishing regulations more effectively. This study offers a concrete technological advancement for enhancing fisheries monitoring and control, paving the way for data-driven management decisions in the region.

4 Conclusion

This study successfully assessed the high consistency between fishing efforts derived from VMS data using a machine learning model (*vmstofish*) and traditional logbook records for skipjack tuna fisheries in the Western Sumatra Indian Ocean. The CatBoost model within *vmstofish* proved highly effective in identifying fishing events, with strong performance metrics (recall of 0.983 and F1-score of 0.931). Although spatial discrepancies exist, particularly in areas with lower logbook coverage, the overall spatiotemporal match rates, especially after the widespread adoption of e-logbooks in 2019 (86.6% perfect match rate), underscore the reliability of VMS-derived data as a valuable complement or alternative to logbook information. This research significantly contributes to improving the quality and completeness of fisheries data, which is essential for accurate stock assessment, informed management decisions, and ultimately fostering a more sustainable blue economy in the region.

Future research should aim to reduce the potential bias by adding more training data from more vessels and by applying cross-validation approaches to explicitly assess model transferability.

References

1. Badan Pusat Statistik. Statistik Pelabuhan Perikanan (2023)
2. D. Panzeri, T. Russo, E. Arneri, R. Carlucci, G. Cossarini, I. Isajlović, S.K. Šifner, C. Manfredi, F. Masnadi, M. Reale, G. Scarcella, C. Solidoro, M.T. Spedicato, N. Vrgoč, W. Zupa, S. Libralato, Identifying priority areas for spatial management of mixed fisheries using ensemble of multi-species distribution models. *Fish Fisheries*. **25**, 187-204 (2024). <https://doi.org/10.1111/faf.12802>
3. A. Stephens, A. MacCall, A multispecies approach to subsetting logbook data for purposes of estimating CPUE. *Fish. Res.* **70**, 299-310 (2004). <https://doi.org/10.1016/j.fishres.2004.08.009>
4. N.D.P. Gunawardane, M.M. Ariyaratna, U.S. Amarasinghe, M.D.S.T. de Croos, Validating the fishing locations reported in the logbooks using thepositional data of vessel monitoring systems in the multi-day fisheryof Sri Lanka. *Sri lanka J. Aquat. Sci.* **28**, 11 (2023). <https://doi.org/10.4038/sljas.v28i1.7604>
5. K. Mahendra, T. Oktavia, Mapping Fishing Behavior: Machine learning implementation on VMS data, in Proceedings of the ISRITI 2024 conference, IEEE, Yogyakarta, Indonesia, December 11 (2024). <https://doi.org/10.1109/ISRITI64779.2024.10963562>
6. Z. Fahmi, Y. Hikmayani, T. Yunanda, P. Yudianto, B. Setyadji. Indonesia National Report to The Scientific Committee of The Indian Ocean Tuna Commission (2020)
7. T.D. Pratiwi, B. Wiryawan, T.W. Nurani, Implementation of tuna traceability in ocean fishing Port of Nizam Zachman Jakarta. *Mar Fisheries*. **12**, 23-34 (2021). <https://doi.org/10.29244/jmf.v12i1.32827>
8. S. Burhani, S. Amin, S. Hadi, A. Setiawan, Optimalisasi penerapan e-logbook penangkapan ikan di Pelabuhan Perikanan Untia, Makassar. *J. Informasi, Sains, dan Teknologi* **5**, 114-128 (2022). <https://doi.org/10.55606/isaintek.v5i01.107>
9. Y. Chen, Quality of fisheries data and uncertainty in stock assessment. *Sci. Mar.* **67**, 75-87 (2003). <https://doi.org/10.3989/scimar.2003.67s175>