PREDICTION OF POTENTIAL FISHING ZONES FOR SCOMBER JAPONICUS

IN THE EAST PACIFIC USING NEURAL NETWORKS

JESSICA VERA BERMUDEZ

Universitat Oberta de Catalunya

Universitat de Barcelona

MASTER'S DEGREE IN BIOESTATISTICS AND BIOINFORMATICS

ADVISORS: ROMINA REBRIJ
VIVIANA JURADO
CARLES VENTURA ROYO

JANUARY, 2023

AGENDA

- BACKGROUND
- GOALS

- WORK PLAN
- METHODS & RESULTS

- CONCLUSIONS
- MAIN REFERENCES

BACKGROUND

1 OVERFISHING

Fishing is one of the ecosystem services that play a key role around the world.
It is vulnerable to suffering from overexploitation (overfishing), which is a global concern nowadays.

2 SCOMBER JAPONICUS

It is in the top 10 list of the most fished species worldwide (1,360K tons in 2020).
It is the 2nd most caught marine species in Ecuador (160k tons in 2021).
It's found in the Indian and Pacific Oceans, and its diet is based on zooplankton and small fish.

Prediction of potential fishing zones for *S. japonicus*

BACKGROUND

3 UNSUSTAINABILITY

S. japonicus is on the list of biologically sustainable fishery stocks.
The Southeast Pacific ocean had the highest percentage (66.7%) of stocks fished at unsustainable levels.

4 STRATEGIES

FAO pursues "Conserve and sustainably use the oceans, seas and marine resources for sustainable development" (SDG14) and encourages countries to contribute to restoring aquatic habitats.*

(*) Food & Agriculture Organization

Prediction of potential fishing zones for *S. japonicus*

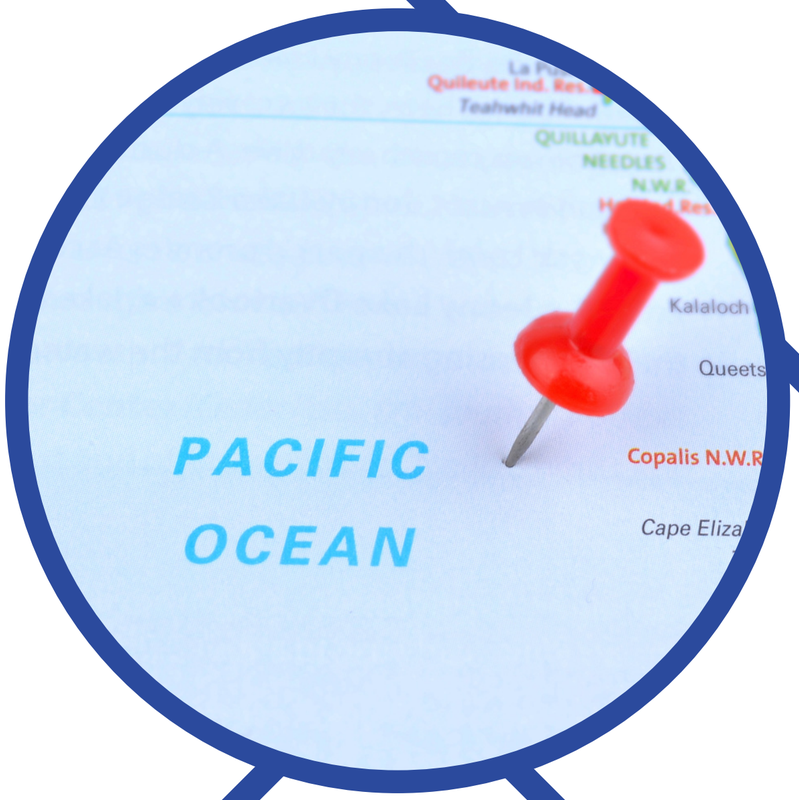
BACKGROUND

5 TOOLS

It is important to understand the distribution of the abundance of the species to manage and monitor the fishery stocks.
This knowledge also contributes to the identification of overfished grounds and potential fishing zones.

6 PROPOSAL

To construct a map of the abundance distribution of the species on the Ecuadorian coast by using neural networks so that potential fishing grounds and zones in need of restoration are identified.

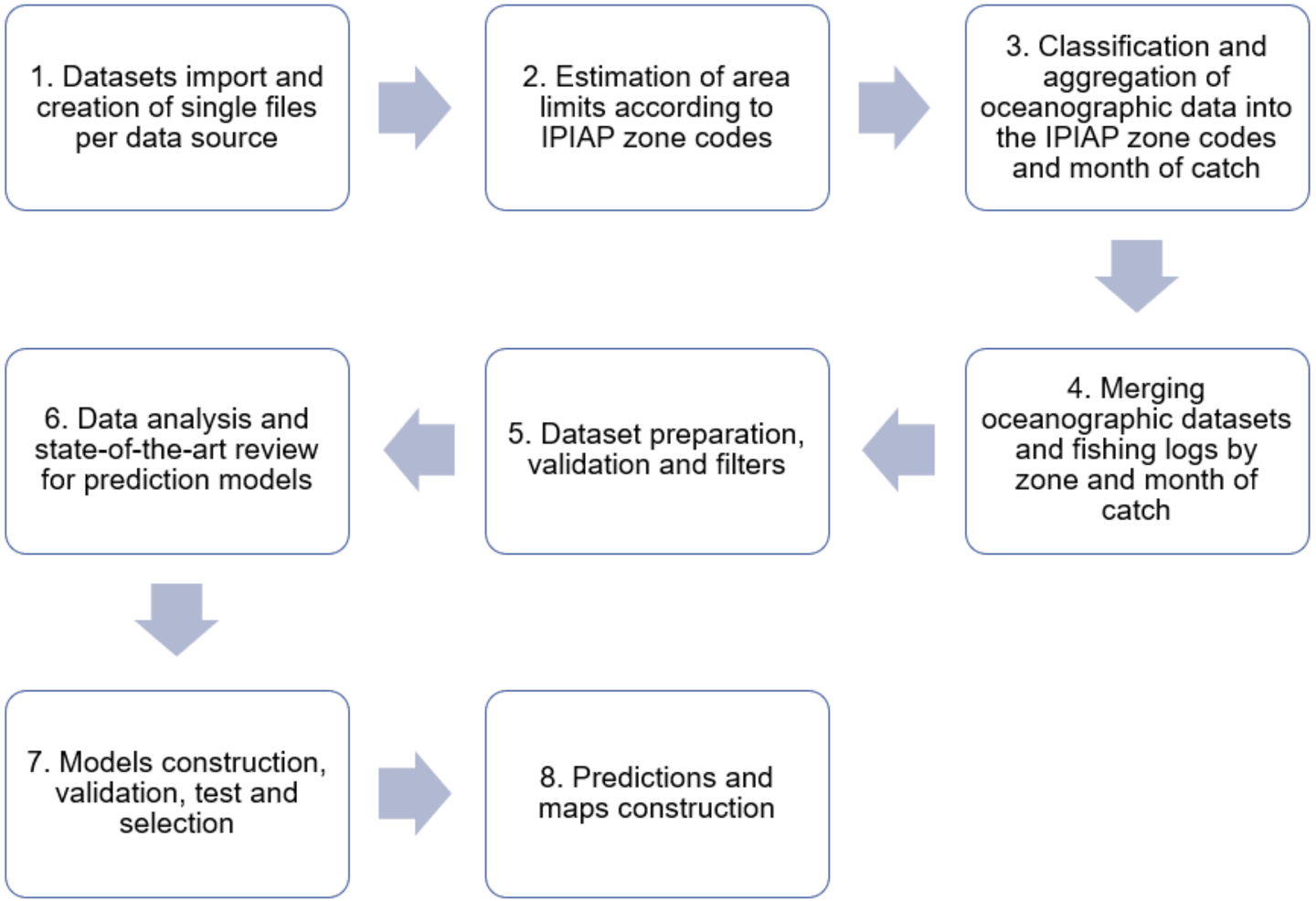


Prediction of potential fishing zones for *S. japonicus*

GOALS

- *To characterize the spatiotemporal distribution of the abundance of Scomber japonicus.*
- *To predict potential fishing zones with the ideal environmental and spatiotemporal conditions for Scomber japonicus by using neural networks.*

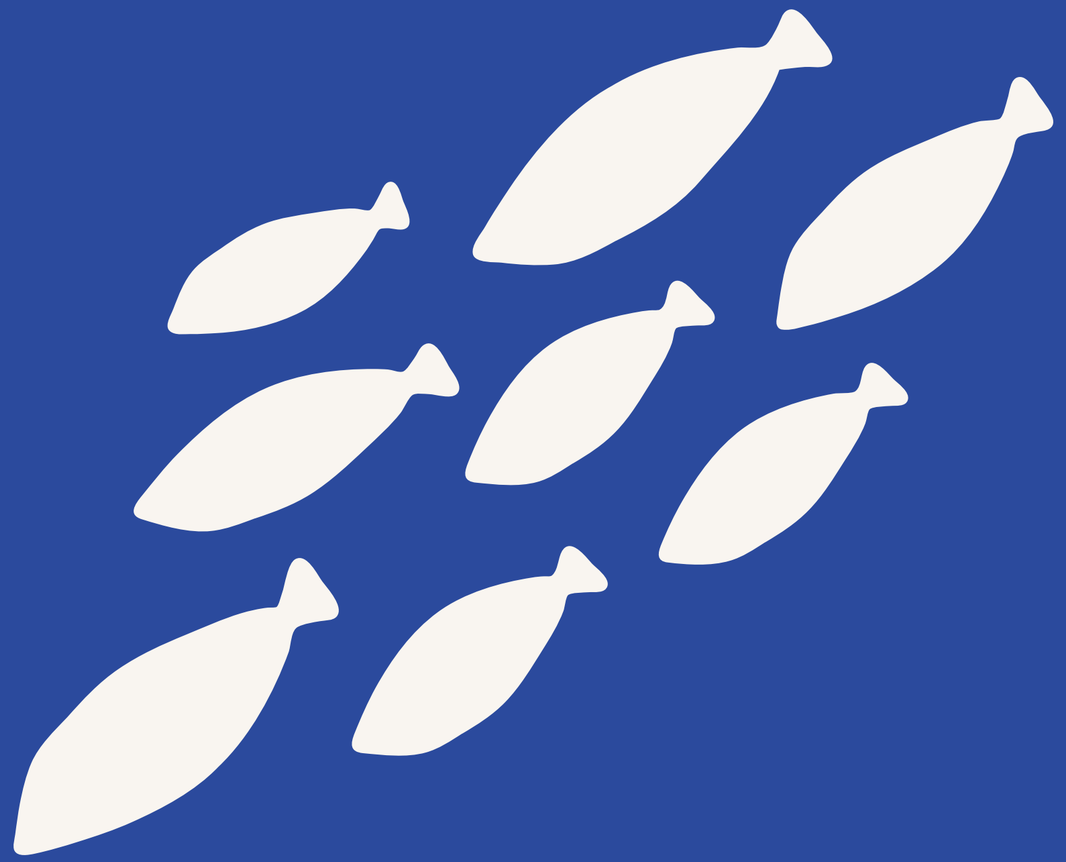
WORK PLAN



(*) IPIAP is the national institution of aquaculture and fishing research in Ecuador

METHODS & RESULTS

WORK PLAN IN ACTION



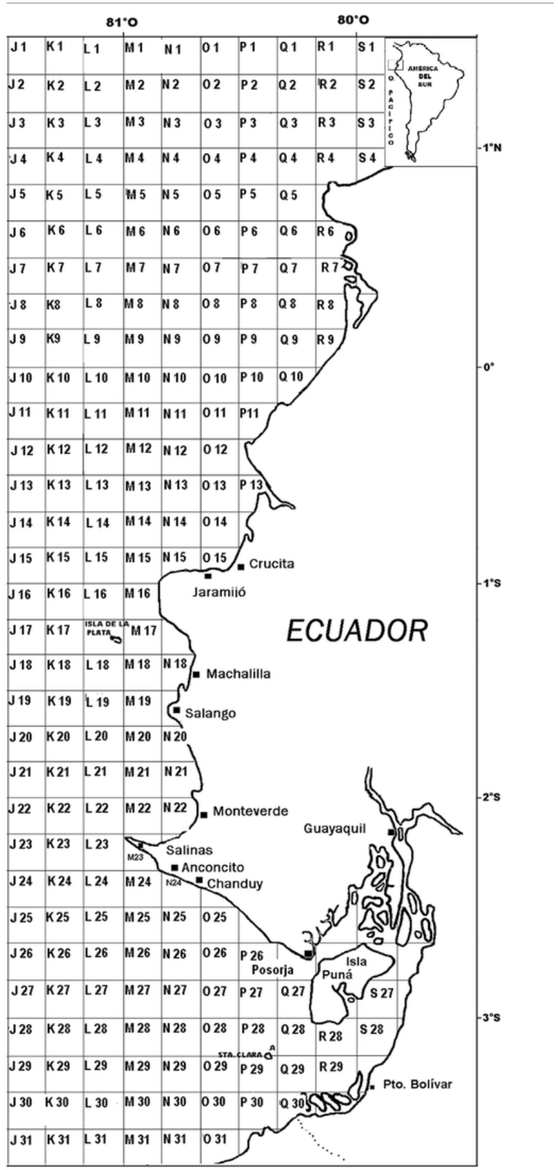
1. Datasets import and single files creation

1 FISHING RECORDS

- *2,146 daily fishing records provided by IPIAP, registered by fishery observers on board.*
- *Records between 2017-01 and 2022-05*
- *219 coded areas of 5X11.5 miles (lat x long), with the central coordinates identified. Only 66 areas had a fishing history.*
- *Location of the Ecuadorian coast: from 01°21' N (lat) and 03°35' S (lat)*
- *The catch was reported in tons.*

2 OCEAN CHLOROPHYLL CONCENTRATION

- *Taken from the sensor MODIS records (NASA Aqua satellite).*
- *Unit of measurement: mgxm-3*
- *Monthly files with. Resolution of 1 km.*



(*) NASA: National Aeronautics and Space Administration

1. Datasets import and single files creation

3 SEA SURFACE TEMPERATURE

- *Taken from the sensor MODIS records (NASA Aqua satellite).*
- *Unit of measurement: degree Celsius*
- *Monthly files.*
- *Resolution: 1 km.*
- *Accuracy: 0.5° C.*

4 WIND MAGNITUDE

- *Taken from NOAA satellite records.*
- *Unit of measurement: m/s*
- *Monthly files.*
- *Height of measurement: 10 MASL.*
- *Accuracy: 1°x1° (long x lat)*

(*) NOAA: National Oceanic and Atmospheric Administration

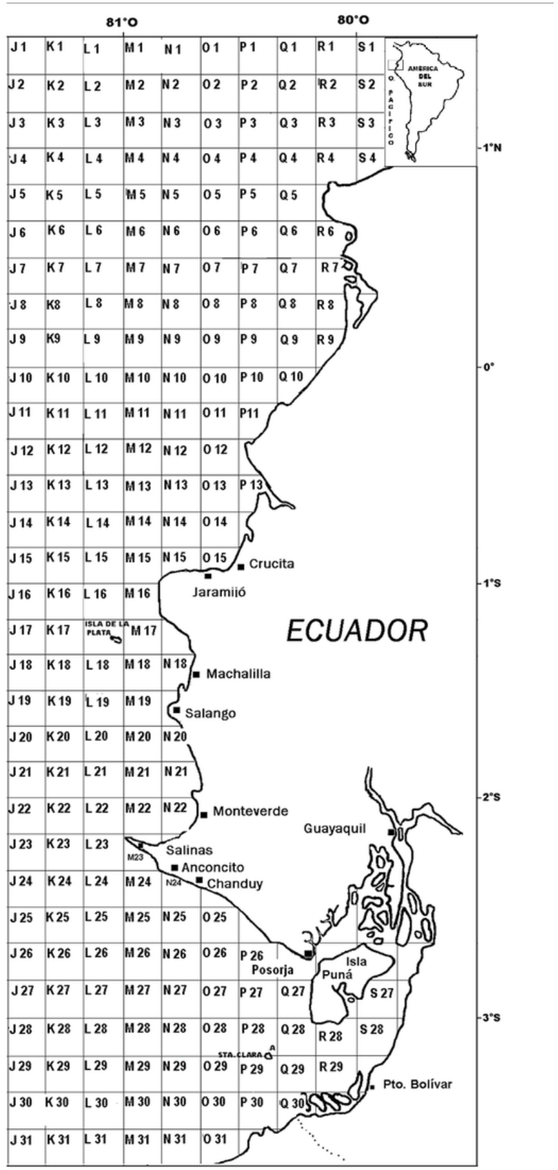
2. Estimation of IPIAP area limits

1

LIMITS ESTIMATION

- *Given the **central coordinates** (Pxi, Pyi), **limits were estimated** for each 5x11.5 (miles) area, by estimating the longitude (Qxij) and latitude (Qyij) of the four **vertices** (j) **of each polygon-area** (i). These estimations were based on the reverse of the Haversine distance.*
- *Assumptions (estimations around the equator):*
  - *Distance between degrees are similar for longitude and latitude.*
  - *Earth radius is 6,378.14 km*

$$Q_{xij} = P_{xi} \pm \frac{\left[\frac{d_x}{2} \times \frac{180}{1000\pi r}\right]}{\cos\left(\frac{P_{yi}\pi}{180}\right)} \qquad Q_{yij} = P_{yi} \pm \left[\frac{d_y}{2} \times \frac{180}{1000\pi r}\right]$$



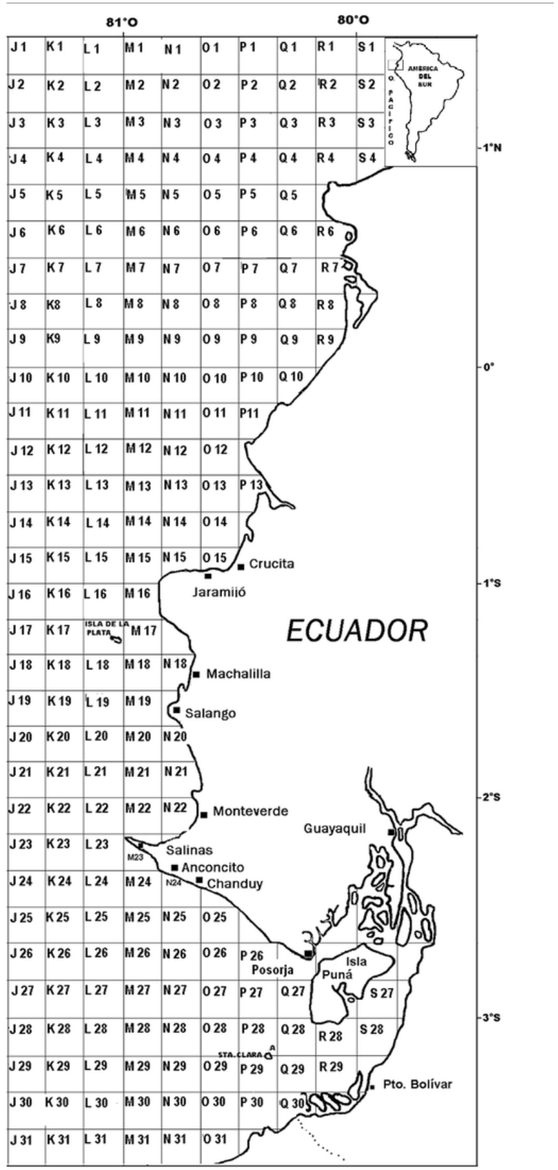
3 + 4. Classification, aggregation & datasets join

1 CORRECTION OF THE LIMITS ESTIMATION

A correction *was needed for the* ***latitude*** *limits so that adjacent areas share the same limit. The distance between the estimated limits that should match was divided by two and added (or subtracted) to them.*

2 DATASETS JOIN

The limits were used to ***cluster the oceanographic points*** *that are within them and aggregate* ***(average) their information*** *in order to get a single value per area and month of each feature.*
The ***left join*** *was used in order to get as many rows as in the fishing records dataset. Match features were the area and month of the catch.*



5. Dataset preparation, validation and filters

1 RESPONSE VARIABLE DEFINITION

- *The catch per unit of effort (**CPUE**) was estimated as the average of the catch per net.*
- *In order to scale and interpret the CPUE value, the relative abundance index was estimated (**IAR**) as CPUE divided by the maximum historical CPUE of the month.*
- *IAR ranges between 0 and 1, where one stands for the maximum relative abundance of the species.*
- *In order to use a symmetric variable as a response in the model, the **square root transformation** was applied to IAR*

$$CPUE_{ymi} = \frac{C_{ymi}}{f_{ymi}} \qquad sqrtIAR_{ymi} = \sqrt{IAR} = \sqrt{\frac{CPUE_{ymi}}{\max\left(CPUE_m\right)}}$$

5. Dataset preparation, validation and filters

2

DATA VALIDATION AND FILTERS

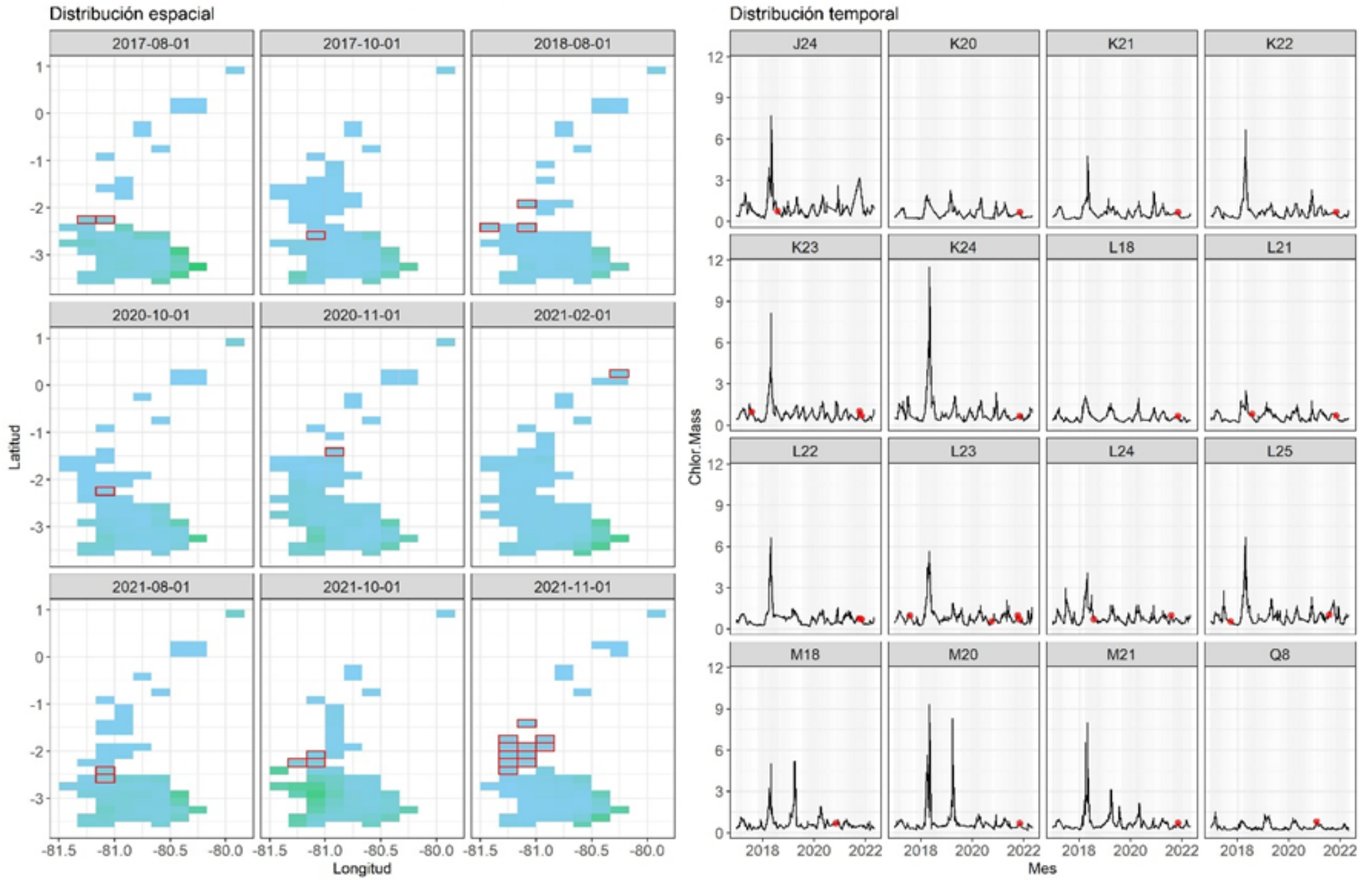
- *Fishing logs during the **closure months** (March and September, according to IPIAP) were removed from the dataset.*
- *36 **outliers in the CPUE variable** were also removed, according to the adjusted boxplot method for asymmetric distributions (whisker length = 3) since CPUE is naturally biased.*

Variable	*Min*	*Max*	*Q1*	*Q2*	*Q3*	*Mean*	*SD*	*Skewness*
CPUE	0.00	210	2.00	5.00	12.00	10.13	15.45	5.18

5. Dataset preparation, validation and filters

3 DATA VALIDATION AND FILTERS

- ***Imputation*** *of 68 missing values from the* ***chlorophyll concentration*** *variable by using the Inverse Distance Weighing interpolation method (IDW). The power parameter was a=2*
- *The final dataset decreased to* ***1,836 examples****.*



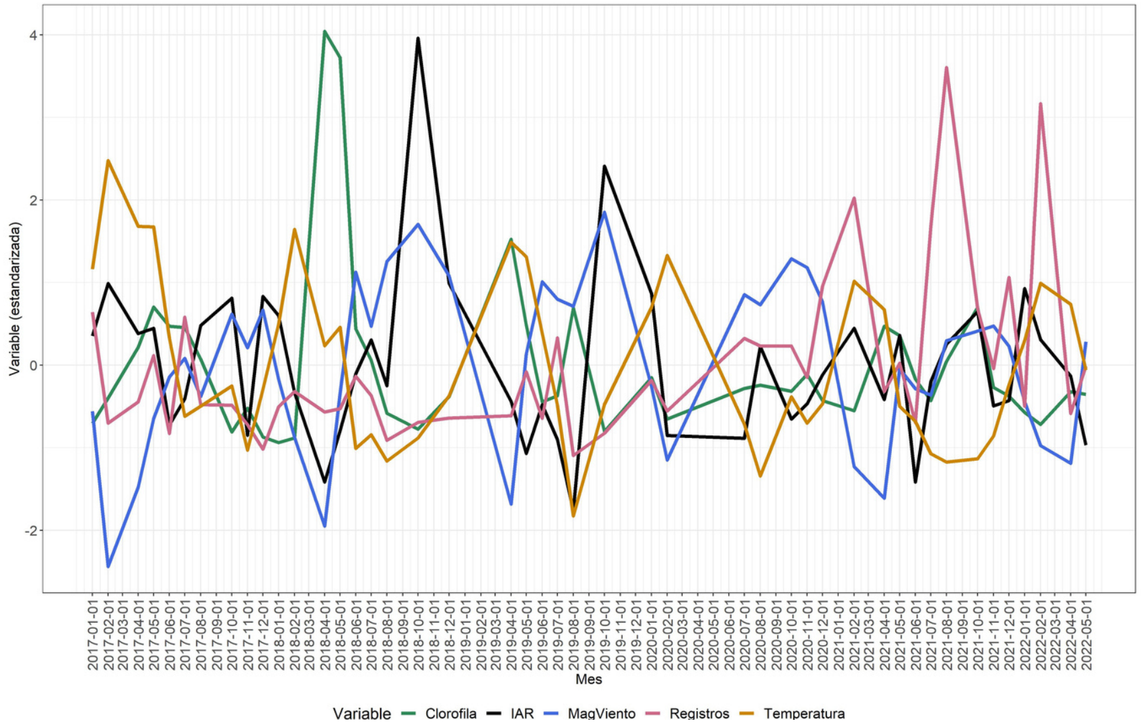
6. Data analysis and variables selection

1 TEMPORAL DIMENSION

- *Most fishing captures were registered in 2021, 2020 and 2017. Whereas 2019 had the lowest number of fishing records. However,* ***2022-02*** *registered more events than other months.*
- *IAR generally increases when temperature decreases (standardized average by month). The highest IAR was registered in* ***2018-10 y 2019-10****.*

2 SPATIAL DIMENSION

- *Most fishing records occurred in the areas located* ***between longitudes N-M, and latitudes between 28-27,*** *regardless of the month. However, the highest IAR was registered in* ***L30, L31 and P30****.*
- *In general, the south of the Ecuadorian coast has been the most frequent zone among the fishers.*



6. Data analysis and variables selection

3 VARIABLES OF INTEREST

- ***Chlorophyll concentration*** *ranged between 0.17 and 7.64 mg/m3, but was highly skewed.*
- ***Sea surface temperature*** *ranged between 18°C and 29.63°C, and the* ***wind magnitude*** *was between 1.73 m/s and 7.15 m/s. Both features were virtually symmetric around their mean values.*
- *CPUE ranged between 0 and 54 tons per net, and the average* ***IAR was 0.18, regardless of the area or month of the catch****.*
- *The correlation between each variable and sqrtIAR was weak.*

Variable	*Min*	*Max*	*Q1*	*Q2*	*Q3*	*Mean*	*SD*	*Skewness*
Chlorophyll	0.17	7.64	0.50	0.70	0.97	0.80	0.51	5.11
Temperature	18	29.13	23.24	24.28	26.12	24.66	1.69	0.25
Wind Magnit.	1.73	7.15	4.27	4.63	4.96	4.59	0.63	-0.39
CPUE	0.00	54.00	2.00	5.00	10.00	8.56	9.41	1.92
IAR	0.00	1.00	0.04	0.11	0.25	0.18	0.19	1.86
sqrtIAR	0.00	1.00	0.21	0.33	0.50	0.37	0.21	0.78

ρ_s / p-value	*Chlorophyll*	*Temperature*	*Wind Magnitude*	*sqrtIAR*
Chlorophyll	1.00	<0.001	<0.001	<0.001
Temperature	-0.37	1.00	<0.001	<0.001
Wind Magnitude	0.26	-0.56	1.00	<0.001
sqrtIAR	-0.09	0.08	-0.11	1.00

(*) ρ_s stands for the Spearman correlation

7. Model

1 SAMPLES

- *Training: 1,088 examples from 2017-1 to 2021-5 (59.3%)*
- *Validation: 373 examples from 2021-6 to 2021-11 (20.3%)*
- *Test : 375 examples from 2021-12 to 2022-5 (20.4%)*

2 MODEL FEATURES

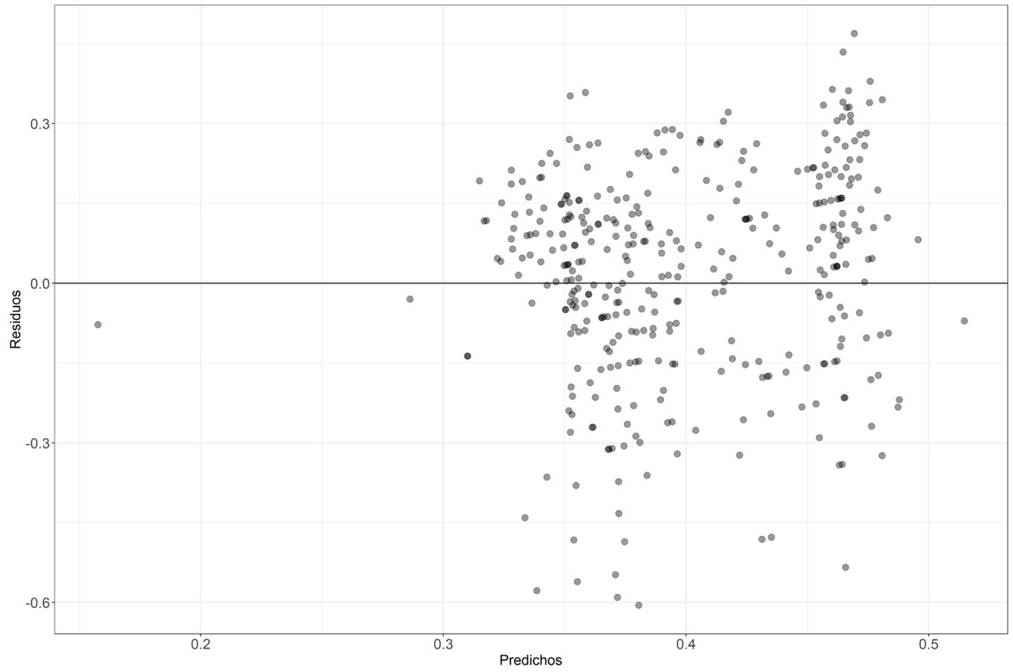
- *Type: MLP*
- *Layers: 4*
- *Input nodes: 6*
- *Nodes of hidden layer 1: 32*
- *Nodes of hidden layer 2: 7*
- *Output nodes: 1*
- *Activation functions: ReLu, Lineal*
- *Total parameters: 463 (224, 231, 8)*
- *Epochs: 50*
- *Batch size: 32*
- *Optimizer: RMSProp*
- *Loss function: MSE*
- *Metric: MAE*

(*) Analysis conducted in R version 4.1.2. Models built on Keras.

3 MODEL PERFORMANCE

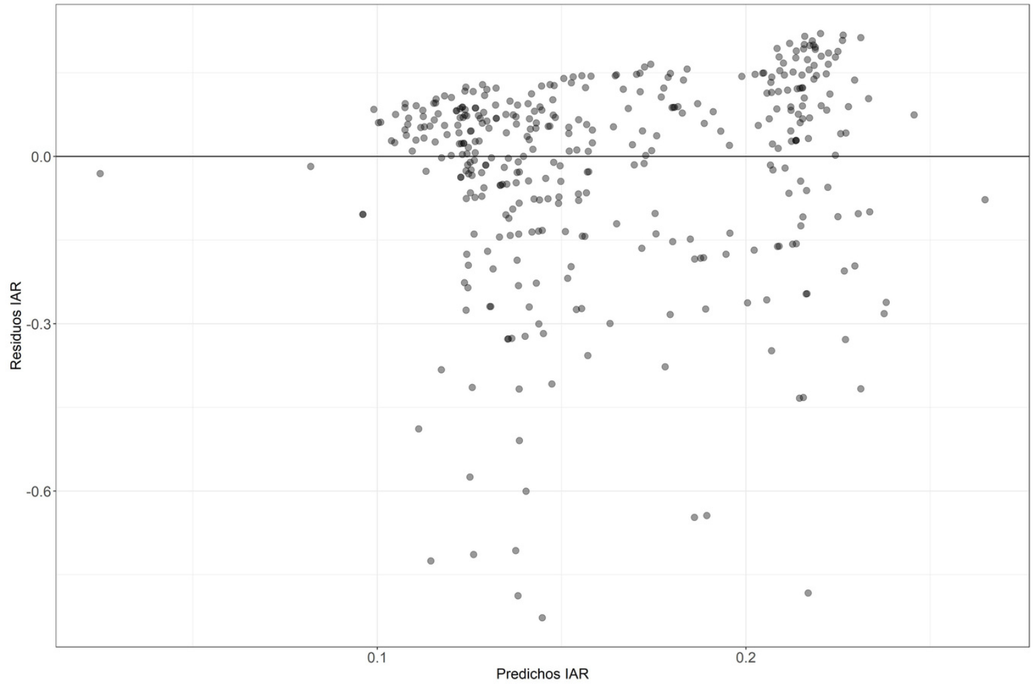
- *The **average RMSE** in the test sample was **0.2** and the **MAE of the predicted IAR** was on average **0.02**.*
- *The percentage of unexplained variability was on average 51%.*
- *The model **underestimated** some values.*

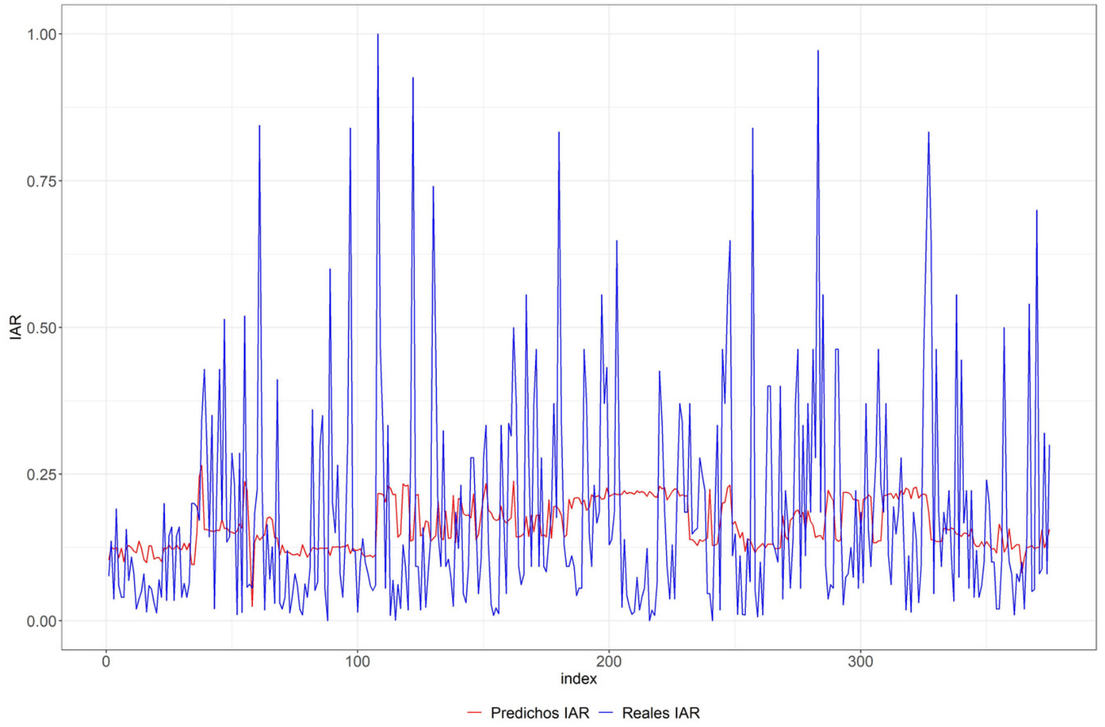
Index	*Sample*	*Iteration*										*Average*
		1	2	3	4	5	6	7	8	9	10	
MSE	Train	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	Val	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	Test	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
MAE (sqrtIAR)	Train	0.17	0.17	0.16	0.16	0.17	0.17	0.16	0.17	0.17	0.17	0.17
	Val	0.16	0.16	0.17	0.16	0.16	0.16	0.16	0.17	0.16	0.16	0.16
	Test	0.15	0.15	0.17	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.16
RMSE	Train	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
	Val	0.21	0.21	0.21	0.21	0.21	0.20	0.21	0.21	0.20	0.20	0.21
	Test	0.19	0.19	0.20	0.19	0.19	0.19	0.19	0.20	0.20	0.20	0.20
MAE (IAR)	Train	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	Val	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	Test	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.02
$RMSE_{test}/\bar{y}_{test}$		0.50	0.51	0.53	0.51	0.51	0.51	0.51	0.52	0.51	0.52	0.51



3 MODEL PERFORMANCE

When transforming the sqrtIAR to IAR residuals, the underestimation in the highest IAR values is more evident since these values were predicted ***over 0.6 units below the actual values****.*

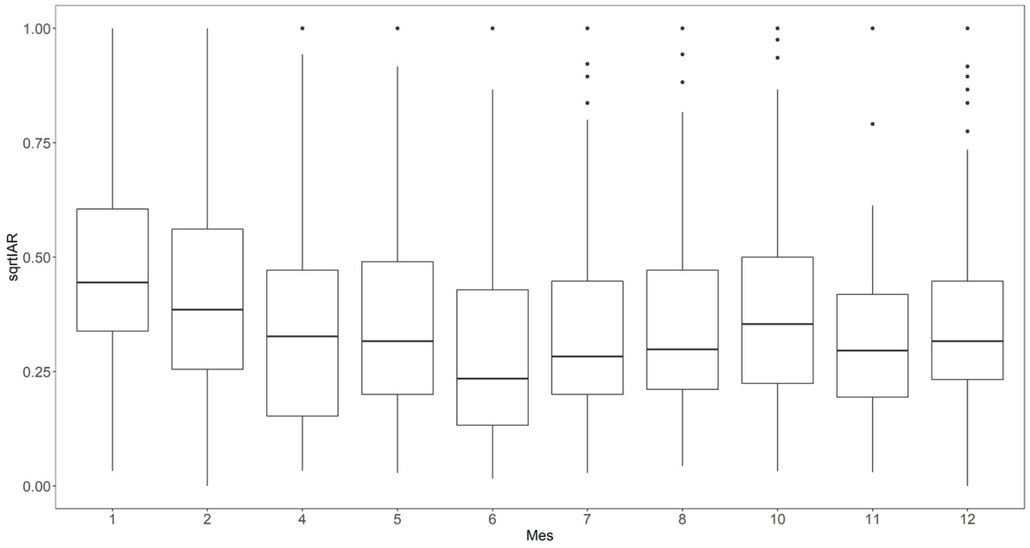




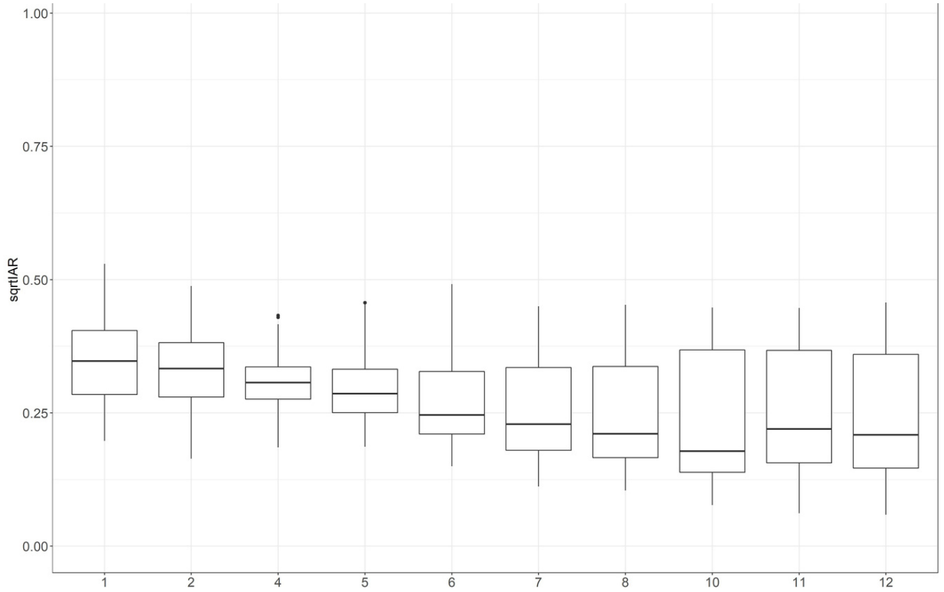
1 PREDICTIONS

The monthly oceanographic information of all 219 areas was used to predict IAR. Later, the monthly distribution of the predicted IAR was compared to the monthly abundance calculated from the fishery records (66 areas). The comparison displays that the model learned about the monthly trend, even though the values were underestimated. This mainly affects the January and February predictions.

Distribution of observed values (66 areas)



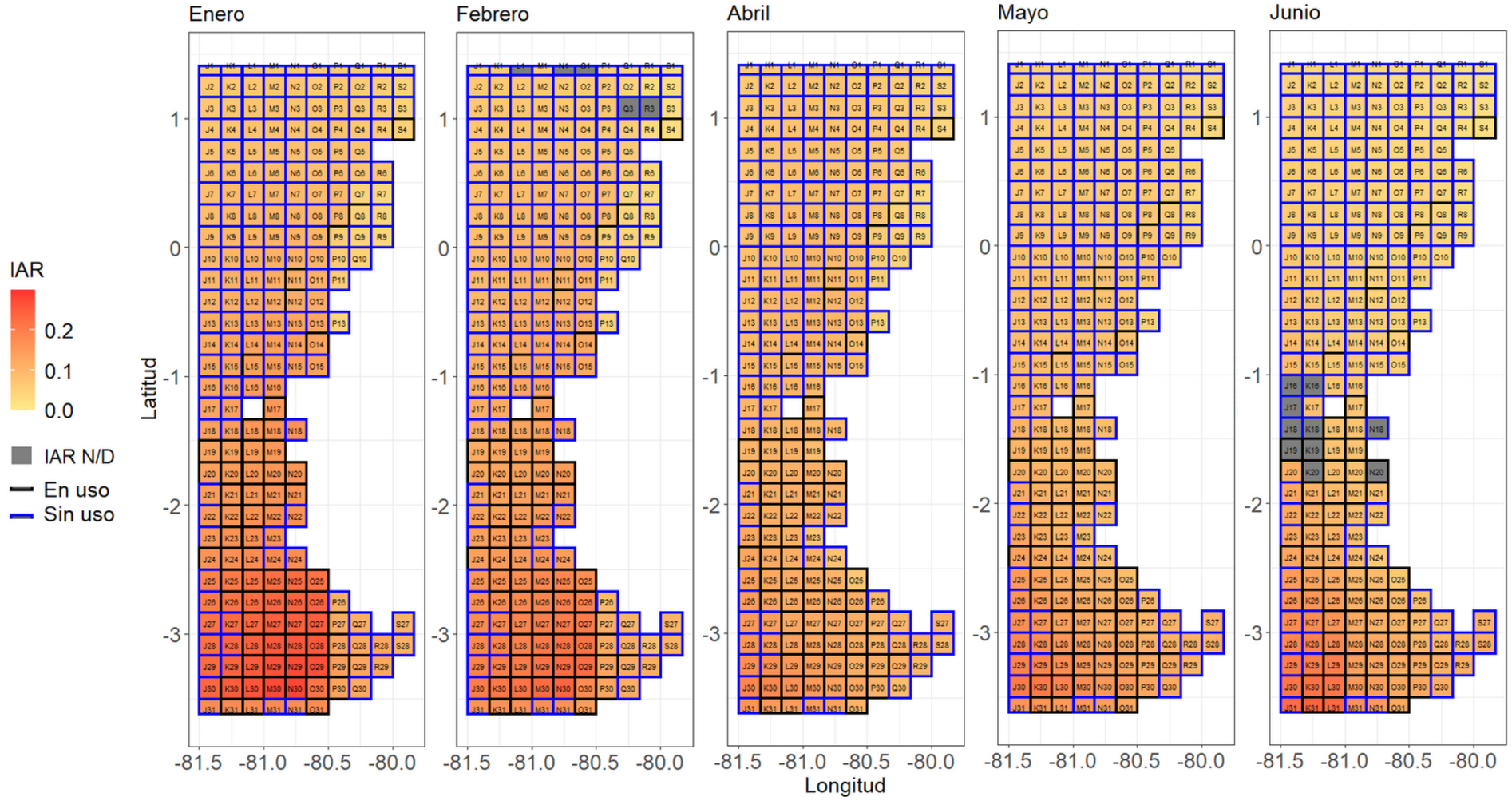
Distribution of predicted values (219 areas)



2 MAPS

Some areas predicted as ***highly abundant*** *have* ***low*** *(black border) or* ***no fishery record*** *(blue border).*

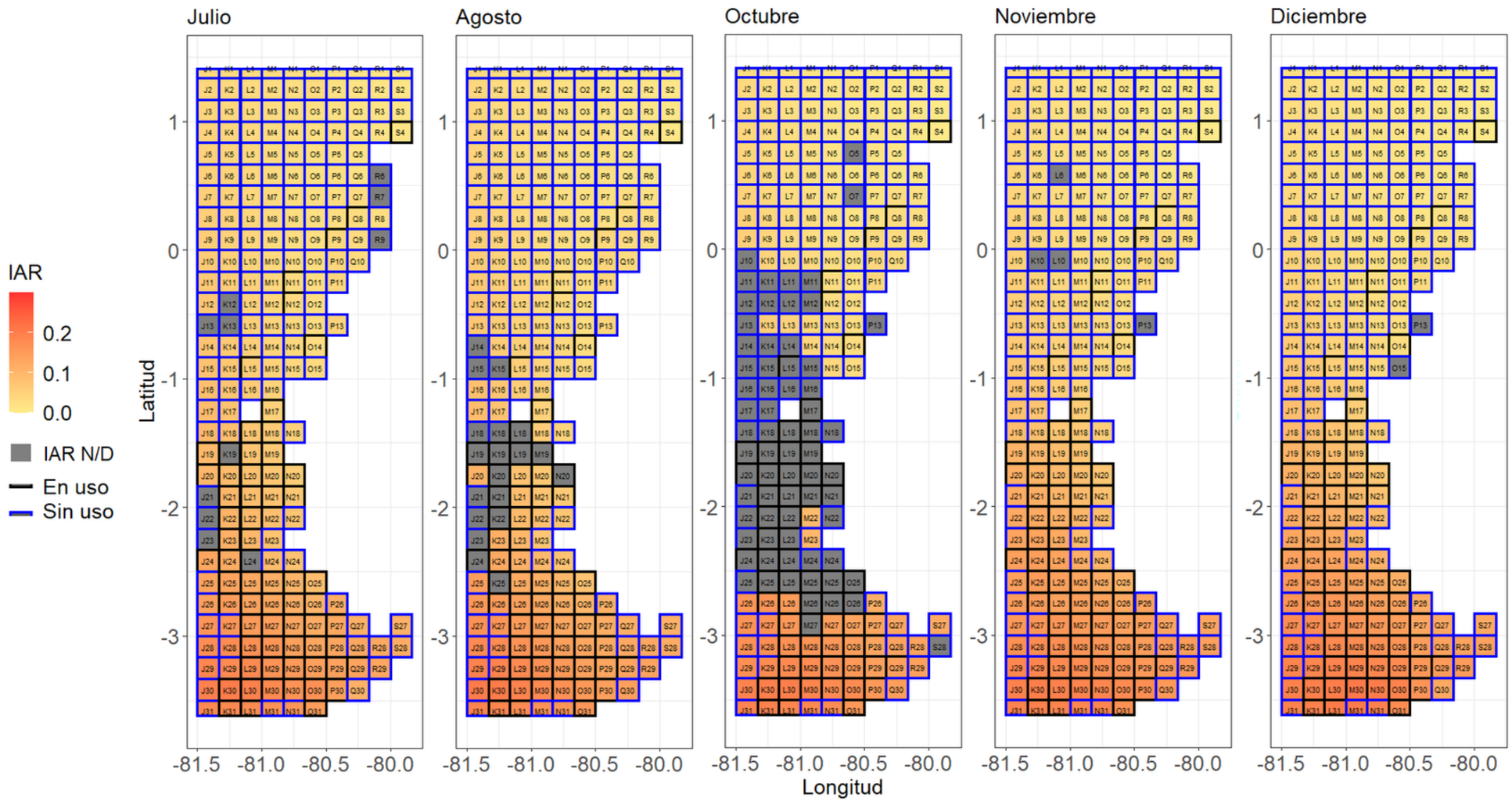
For example: From the ***J25-J31 areas*** *and others from the row 31*



8. Predictions and maps

2 MAPS

*In the **second semester** of the year, IAR is predicted to be **low**, even though the south appears more abundant. Also, some areas are not predicted because of the **missing values in chlorophyll**.*



Prediction of potential fishing zones for *S. japonicus*

CONCLUSIONS

- *The S. japonicus capture has been observed in* ***30% of the areas in Ecuador****, where the most abundant ones since 2017 have been coded as* ***L30 and L31****, and the most frequent areas were located in the* ***N-M and 28-27 zone****, regardless of the month.*
- ***February and January*** *were the most* ***frequent and abundant*** *months in the records.*
- *The areas located in the south were predicted with the highest IAR (Castro Hernández et al. 2000).*
- *The model explains* ***49%*** *of total variability; its main disadvantage is the prediction of higher IAR values. Nonetheless,* ***it kept the spatial and monthly trend****.*
- ***No correlation*** *was reported between the* ***input and the response variable****. However, an* ***indirect relationship*** *between* ***sea surface temperature and IAR*** *was observed across the months (Canales & Jurado 2021).*
- ***Potential fishing zones*** *of the species are between the* ***J25-J31 and all the areas in the 31st row****. In addition, the least frequent areas in the south should also be considered PFZ.*

- ***Less complex models*** *performed better (Armas et al. 2022; Wang et al. 2015). Also,* ***smaller batch sizes*** *and the use of ReLu improved the* ***model performance*** *(Masters & Luschi 2018).*
- *Taking* ***oceanographic measurements at the moment of the capture and designing an adequate sample plan*** *play a key role in prediction accuracy.*
- *The main challenges in this work were at the dataset preparation and modeling stages.*
- *The proposed* ***maps contribute to the SDG14 accomplishment*** *as long as IPIAP uses them as a decision-making tool.*
- *The data quality and sample size were the main limitations of this research, restricting the search of models, e.g., RNN.*
- *Nonetheless, this work becomes an approach toward exploring the abundance distribution of the marine species in Ecuador, which also offers opportunities to improve the sampling process and data analysis in this regard.*

REFERENCES

- *Armas, Elier, Hugo Arancibia, and Sergio Neira. 2022. "Identification and Forecast of Potential Fishing Grounds for Anchovy (Engraulis ringens) in Northern Chile Using Neural Networks Modeling". Fishes 7 (4).*
- *Beysolow II, Taweh. 2017. "Introduction to Deep Learning". In Introduction to Deep Learning Using R: A Step-by-Step Guide to Learning and Implementing Deep Learning Models Using R, edited by Taweh Beysolow II, 1–9. Berkeley, CA: Apress.*
- *Canales, Cristian M, and Viviana Jurado. 2021. "Evaluación del stock de recursos pelágicos pequeños del Ecuador. Año 2021". Technical. Guayaquil, Ecuador: Instituto Público de Investigación de Acuicultura y Pesca.*
- *Castro Hernández, José J., and Ana T. Santana Ortega. 2000. Synopsis of Biological Data on the Chub Mackerel (Scomber Japonicus Houttuyn; 1782). Rome, Italy: FAO Fisheries Synopsis.*
- Chen, Xinjun, Gang Li, Bo Feng, and Siquan Tian. 2009. "Habitat suitability index of Chub mackerel (Scomber japonicus) from July to September in the East China Sea". Journal of Oceanography 65 (February): 93–102.
- Chollet, François, Tomasz Kalinowski, and Joseph J Allaire. 2022. Deep Learning with R. 2nd ed. New York, NY: Manning.
- *FAO*. 2022. The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation. The State of World Fisheries and Aquaculture (SOFIA) 2022. Rome, Italy: FAO.
- Huang, Guang-Bin. 2003. "Learning capability and storage capacity of two-hidden-layer feedforward networks". IEEE Transactions on Neural Networks 14 (2): 274–81.
- Hubert, Wayne A, and Mary C Fabrizio. 2007. "Relative Abundance and Catch per Unit Effort". In Analysis and Interpretation of Freshwater Fisheries Data, 279–325. Bethesda, Maryland: American Fisheries Society.
- Plant, Richard E. 2018. Spatial Data Analysis in Ecology and Agriculture Using R. 2a ed. Boca Raton: CRC Press.
- Wang, Jintao, Wei Yu, Xinjun Chen, Lin Lei, and Yong Chen. 2015. "Detection of potential fishing zones for neon flying squid based on remote-sensing data in the Northwest Pacific Ocean using an artificial neural network". International Journal of Remote Sensing 36 (13): 3317–30.
- Wilcox, Rand R. 2013. "Some Outlier Detection Methods". In Introduction to Robust Estimation and Hypothesis Testing, 3rd ed., 96–100. St. Louis: Elsevier Science & Technology.

URL a la memoria del TFM