Universidad Nacional José Faustino Sánchez Carrión
Escuela de Posgrado

Servicio pesquero predictivo de la densidad de los cardúmenes de anchoveta en el mar peruano

Tesis

Para optar el Grado de Maestro en Gerencia de Servicios Pesqueros

Autor
Oswaldo Miguel Flores Huaman

Asesor
Dr. Oswaldo Francisco Flores Saldaña



Huacho - Perú
2026



Reconocimiento - No Comercial – Sin Derivadas - Sin restricciones adicionales
https://creativecommons.org/licenses/by-nc-nd/4.0/
Reconocimiento: Debe otorgar el crédito correspondiente, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerlo de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante lo respalda a usted o su uso. **No Comercial:** No puede utilizar el material con fines comerciales. **Sin Derivadas:** Si remezcla, transforma o construye sobre el material, no puede distribuir el material modificado. **Sin restricciones adicionales:** No puede aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros de hacer cualquier cosa que permita la licencia.

UNIVERSIDAD NACIONAL
JOSÉ FAUSTINO SÁNCHEZ CARRIÓN

LICENCIADA

(Resolución de Consejo Directivo N° 012-2020-SUNEDU/CD de fecha 27/01/2020

METADATOS

DATOS DEL AUTOR (ES):		
APELLIDOS Y NOMBRES	**DNI**	**FECHA DE SUSTENTACIÓN**
Oswaldo Miguel Flores Huaman	**15738203**	**23 de octubre 2026.**
DATOS DEL ASESOR:		
APELLIDOS Y NOMBRES	**DNI**	**CÓDIGO ORCID**
Oswaldo Francisco Flores Saldaña	**15582719**	**https://orcid.org/0000-0001-7582-7430**
DATOS DE LOS MIEMBROS DE JURADOS – PREGRADO/POSGRADO-MAESTRÍA-DOCTORADO:		
APELLIDOS Y NOMBRES	**DNI**	**CODIGO ORCID**
Fidel Alberto Reyes Ulfe	**15632620**	**https://orcid.org/0000-0002-2737-0357**
Juan Eduardo Navarro Rojas	**15958780**	**https://orcid.org/0000-0002-6204-6157**
Jesús Gustavo Barreto Meza	**15589980**	**https://orcid.org/0000-0002-5790-6757**

OSWALDO MIGUEL FLORES HUAMAN (2026-058365)

SERVICIO PESQUERO PREDICTIVO DE LA DENSIDAD DE LOS CARDÚMENES DE ANCHOVETA EN EL MAR PERUANO

DGI-POSGRADO 2026

Dirección de Gestión de la Investigación-VRI 2026

DIRECCION DE GESTION DE LA INVESTIGACION

Detalles del documento

Identificador de la entrega
trn:oid:::1:3633384226

Fecha de entrega
25 ago 2026, 3:27 p.m. GMT-5

Fecha de descarga
25 ago 2026, 3:41 p.m. GMT-5

Nombre del archivo
sidad_de_los_card_menes_de_anchoveta_en_el_mar_peruano_ok_1.pdf

Tamaño del archivo
4.5 MB

183 páginas

45.812 palabras

272.482 caracteres

17% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Fuentes principales

15% Fuentes de Internet

11% Publicaciones

8% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.° de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

DEDICATORIA

A mis queridos padres:

Norma y Oswaldo

Su amor, sacrificio y fe constante han sido el motor que me ha impulsado a superar cada obstáculo. Esta tesis es un reflejo de su dedicación y de los valores que me transmitieron. Sin ustedes, esta meta no habría sido posible.

A mi amada esposa:

Milagros

Por su paciencia, comprensión y apoyo incondicional, que me permitieron encontrar el equilibrio necesario para dedicarme a este proyecto. Tu amor fue una fuente constante de inspiración para dar siempre lo mejor de mí.

A mis amadas hijas:

Saori y Rubi

Desde el primer instante, ustedes han sido mi mayor fuente de inspiración, fortaleza y amor. Este logro lleva sus nombres grabados; es testimonio del amor infinito que les tengo.

A mi familia y a la familia de mi esposa:

Francisco F., Miguelina T., Rogelio H., Gisla H., Nelly D., Arminda M., Enrique R., Fernando R., Veronica R. y David R.

Por su apoyo incondicional en los momentos difíciles y por celebrar cada uno de mis triunfos. Sin ustedes, esta meta no habría sido posible.

AGRADECIMIENTOS

Expreso mi más sincero y especial agradecimiento al profesor Robert A. Rigby, cocreador de los Modelos Aditivos Generalizados para Ubicación, Escala y Forma (GAMLSS) y profesor de la Escuela de Computación y Ciencias Matemáticas de la Universidad de Greenwich (Londres, Reino Unido), por su valioso asesoramiento, orientación académica y generosa enseñanza. Su apoyo fue fundamental para comprender con mayor profundidad los fundamentos teóricos y aplicados de los modelos GAMLSS; sin su guía, no habría sido posible alcanzar el nivel de avance metodológico logrado en esta tesis.

Asimismo, agradezco a la Dra. Fernanda De Bastiani, coautora del paquete estadístico gamlss, creadora de los diagramas de cubo y de la extensión gamlss.spatial, profesora del Departamento de Estadística de la Universidad Federal de Pernambuco (Recife, Pernambuco, Brasil); al Dr. Nikolaus Umlauf, creador del paquete estadístico Modelos Aditivos Bayesianos para Ubicación, Escala y Forma (BAMLSS), desarrollador principal de gamlss2 y profesor del Departamento de Estadística de la Universidad de Innsbruck (Innsbruck, Austria); al Dr. Abu Hossain, profesor de la Universidad Metropolitana de Londres (Londres, Reino Unido) y Dr. Julian Merder , por su asesoramiento en el uso de los marcos estadísticos avanzados empleados en el desarrollo de esta tesis.

Agradezco al Instituto del Mar del Perú por facilitar el acceso a la información empleada en esta investigación, así como a mis colegas científicos Marceliano Segura Zamudio, Juan Manuel Valles Meza, Jesús Ledesma Rivera, Carlos Paulino Rojas y Rodolfo Cornejo Urbina, por sus sugerencias y aportes que contribuyeron a mejorar el contenido de esta tesis.

ÍNDICE

DEDICATORIA v
AGRADECIMIENTOS vi
ÍNDICE vii
RESUMEN xiii
ABSTRACT xiv
INTRODUCCIÓN xv
CAPÍTULO I. 1
1. Planteamiento del problema 1
1.1 Descripción de la realidad problemática 1
1.2 Formulación del problema 3
1.2.1. Problema general 3
1.2.2. Problemas específicos 3
1.3 Objetivos de la investigación 4
1.3.1. Objetivo general 4
1.3.2. Objetivos específicos 4
1.4 Justificación de la investigación 5
1.5 Delimitaciones del estudio 10
CAPÍTULO II. 12
2. MARCO TEÓRICO 12
2.1 Antecedentes de la investigación 12
2.1.1. Investigaciones internacionales 12
2.1.2. Investigaciones nacionales 16
2.2 Bases teóricas 20
2.2.1. Sistemas de detección acústica 20
2.2.2. Ecoconteo y Ecointegración 21
2.2.3. Calibración de sistemas acústicos 21
2.2.4. Detección acústica 22
2.2.5. Coeficiente del área de dispersión 22
2.2.6. Ecosondas científicas 24
2.2.7. Ecosondas comerciales 24
2.2.8. Servicio pesquero predictivo 24
2.2.9. Densidad acústica de los cardúmenes de anchoveta 25
2.2.10. Agregación extrema de cardúmenes 25
2.2.11. Análisis predictivo 25
2.2.12. Variable de Respuesta 26
2.2.13. Variable Explicativa 26
2.2.14. Inferencia estadística 26
2.2.14.1. Población 26
2.2.14.2. Muestra 27
2.2.14.3. Modelo 28
2.2.14.4. Modelos paramétricos 28
2.2.14.5. Modelos no paramétricos 29
2.2.15. Modelos lineales generalizados (GLM) 30
2.2.16. Modelos aditivos generalizados (GAM) 31
2.2.17. Redes neuronales 32
2.2.18. Bosques aleatorios 34
2.2.19. Aproximación de Laplace integrada anidada 35
2.2.20. Modelo vectorial autorregresivo espacio-temporal (VAST) 36

2.2.21. Modelos aditivos generalizados para locación escala y forma 38
2.2.21.1. GAMLSS como modelo de regresión distribucional 39
2.2.21.2. Función de verosimilitud 40
2.2.21.3. Función de verosimilitud de una variable continua 41
2.2.21.4. Estimación de máxima verosimilitud 42
2.2.21.5. Propiedades estadísticas del estimador de máxima verosimilitud (MLE) en un modelo correcto 42
2.2.21.6. Evaluación del ajuste de los modelos GAMLSS: PIT y z-score 44
2.2.21.7. Selección de modelos 45
2.2.21.7.1. Desviación global 45
2.2.21.7.2. Selección de modelos mediante el Criterio de Información Generalizado de Akaike (GAIC) 46
2.2.21.8. Distribuciones de respuesta 47
2.2.21.9. Funciones de enlace 48
2.2.21.10. Notación de un modelo GAMLSS 49
2.2.21.11. Distribuciones continuas 50
2.2.21.12. Distribuciones continuas en la línea real $\mathbb{R}(-\infty, +\infty)$ 50
2.2.21.13. Distribuciones continuas en la línea real positiva $\mathbb{R} + (\mathbf{0}, +\infty)$ 53
2.2.21.14. Términos del modelo aditivo 57
2.2.21.15. Splines penalizados 58
2.2.22. Modelos aditivos bayesianos para locación escala y forma 59
2.2.22.1. Estructura del modelo BAMLSS 60
2.2.22.2. Respuesta y distribución posterior del modelo BAMLSS. 61
2.3 Bases filosóficas 62
2.4 Definición de términos básicos 63
2.4.1. Servicio 63
2.4.2. Industria pesquera 64
2.4.3. Modelado estadístico 64
2.4.4. Cardumen de peces 65
2.4.5. Descriptores morfológicos 65
2.4.6. Factores espacio-temporales 66
2.5 Hipótesis de la investigación 66
2.5.1. Hipótesis general 66
2.5.2. Hipótesis específicas 66
2.6 Operacionalización de las variables 68
CAPÍTULO III. 69
3. METODOLOGÍA 69
3.1 Diseño metodológico 69
3.2 Población y muestra 69
3.2.1. Población 69
3.2.2. Muestra 69
3.2.3. Recolección de datos de la densidad de los cardúmenes de anchoveta 70
3.2.4. Descriptores morfológicos utilizando algoritmos SHAPES 70
3.2.5. Variables espaciales y temporales 74
3.3 Técnicas para procesamiento de la información 74
3.3.1. Análisis exploratorio 74
3.3.2. División de los datos en conjuntos de entrenamiento, validación y prueba 76
3.3.3. Selección del modelo 76
3.3.4. Evaluación del modelo ajustado 82

3.3.5. Representación y estimación de los efectos de las variables explicativas sobre el s_A de los cardúmenes de anchoveta 82
3.3.6. Representación de la función de densidad de probabilidad (pdf) y estimación de la excedencia de umbrales ecológicamente relevantes del s_A 83
3.3.7. Predicción del parámetro de localización de la densidad de los cardúmenes de anchoveta 83
3.3.8. Evaluación del desempeño predictivo de los modelos 85
CAPÍTULO IV 87
4. RESULTADOS 87
4.1 Análisis de resultados 87
4.1.1. Análisis exploratorio de los datos 87
4.1.2. Regresión clásica centrada en la media 89
4.1.3. Modelo espacio-temporal autorregresivo vectorial (VAST) 93
4.1.4. Enfoques estadísticos distribucionales 93
4.1.5. Efectos de las variables explicativas sobre la densidad típica $Q_{0.50}$ y las densidades más altas $Q_{0.95}$ de los cardúmenes de anchoveta 104
4.1.6. Función de densidad de probabilidad y excedencia de umbrales como índices biológico-pesqueros de la anchoveta 109
4.1.7. Predicción y desempeño predictivo de los modelos GAMLSS, BAMLSS, NN-GAMLSS y RF-GAMLSS para el s_A de los cardúmenes de anchoveta 113
4.2 Contrastación de hipótesis 117
4.2.1. Hipótesis general 117
4.2.2. Hipótesis específica 1 117
4.2.3. Hipótesis específica 2 118
4.2.4. Hipótesis específica 3 118
4.2.5. Hipótesis específica 4 118
CAPÍTULO V 119
5. DISCUSIÓN 119
5.1. Discusión de resultados 119
5.1.1. Inconsistencia de la validación basada en comparar la variable observada Y con el parámetro de localización 119
5.1.2. Inconsistencia de pretransformar Y, la constante arbitraria +1 y el enlace logarítmico 120
5.1.3. Distribución apropiada para la densidad de los cardúmenes de anchoveta 122
5.1.4. Comparabilidad de desempeño predictivo entre modelos de regresión distribucional e incertidumbre predictiva 125
5.1.5. Diagnóstico residual engañoso en los diagramas Q-Q 127
5.1.6. Insuficiencia del Campo Latente y mallas espaciales 127
5.1.7. Limitaciones de la regresión clásica centrada en la media frente a la regresión distribucional 129
5.1.8. La superioridad de un modelo simple frente a uno complejo 133
5.1.9. Ausencia de datos de ecosondas comerciales calibradas en el entrenamiento y validación de los modelos 135
5.1.10. Ventajas de la regresión distribucional en el estudio de la ecología de la anchoveta 137
5.1.11. Implicaciones para las evaluaciones acústicas y la gestión pesquera 138
CAPÍTULO VI 140
6. CONCLUSIONES Y RECOMENDACIONES 140
6.1. Conclusiones 140
6.2. Recomendaciones 141

CAPÍTULO VII. 142
7. REFERENCIAS BIBLIOGRÁFICAS 142
ANEXOS 161

ÍNDICE DE TABLAS

Tabla 1..... 68
Tabla 2..... 87
Tabla 3..... 98
Tabla 4..... 100
Tabla 5..... 101
Tabla 6..... 102
Tabla 7..... 108
Tabla 8..... 112
Tabla 9..... 114
Tabla 10..... 115
Tabla 11..... 117
Tabla 12..... 161

ÍNDICE DE FIGURAS

Figura 1 50
Figura 2 52
Figura 3 53
Figura 4 54
Figura 5 56
Figura 6 57
Figura 7 71
Figura 8 74
Figura 9 88
Figura 10 89
Figura 11 90
Figura 12 91
Figura 13 92
Figura 14 92
Figura 15 93
Figura 16 99
Figura 17 100
Figura 18 102
Figura 19 103
Figura 20 104
Figura 21 106
Figura 22 107
Figura 23 108
Figura 24 111

RESUMEN

El objetivo de la tesis fue evaluar la validez residual y la capacidad predictiva de los modelos clásicos mediante GLM, GAM, INLA y VAST con una distribución Gamma, y de la regresión distribucional (RD) mediante GAMLSS, BAMLSS, NN-GAMLSS y RF-GAMLSS, con las distribuciones logSEP3 y logSEP4, como base metodológica para un servicio pesquero predictivo de la densidad acústica (s_A) de los cardúmenes de anchoveta basado en descriptores morfológicos y factores espacio-temporales de ecosondas comerciales no calibradas. Se analizó la información obtenida por IMARPE entre 2010 y 2017. La validación residual incluyó diagramas Q-Q, de gusano y de cubo. El desempeño predictivo se evaluó mediante RMSE, MAE y CRPS con intervalos bootstrap al 95 %. Los modelos clásicos presentaron patrones residuales sistemáticos, indicativos de modelos incorrectos; la incorporación de un campo latente espacial en INLA y VAST no corrigió la mala especificación. La RD no exhibió patrones sistemáticos que evidencien modelos incorrectos, excepto RF-GAMLSS en invierno y primavera. Las covariables longitud, latitud, distancia a la costa y altura media del cardumen ejercieron los mayores efectos sobre la densidad típica ($Q_{0.5}$) y las densidades extremas ($Q_{0.95}$), con un marcado patrón estacional. El desempeño predictivo fue casi indistinguible entre los modelos distribucionales, a excepción de NN-GAMLSS, que fue el peor en invierno. GAMLSS y BAMLSS presentaron un desempeño predictivo consistente en todas las estaciones y proporcionan una base metodológica sólida para un servicio pesquero predictivo que estime densidad acústica, índices ecológicos y probabilidades de excedencia con datos de ecosondas comerciales sin calibración previa.

Palabras clave: densidad acústica; regresión clásica; regresión distribucional; cardúmenes de anchoveta; servicio pesquero predictivo.

ABSTRACT

The aim of this thesis was to assess the residual validity and predictive capacity of classical models fitted with GLM, GAM, INLA, and VAST under a Gamma distribution, and of distributional regression (DR) through GAMLSS, BAMLSS, NN-GAMLSS, and RF-GAMLSS under the logSEP3 and logSEP4 distributions, as a methodological basis for a predictive fisheries service for the acoustic density (s_A) of anchovy schools based on morphological descriptors and spatiotemporal factors from uncalibrated commercial echosounders. Data collected by IMARPE between 2010 and 2017 were analyzed. Residual validation included Q-Q, worm, and bucket plots. Predictive performance was assessed with RMSE, MAE, and CRPS with 95% bootstrap intervals. The classical models showed systematic residual patterns indicative of misspecified models; adding a spatial latent field in INLA and VAST did not correct the misspecification. DR showed no systematic patterns indicative of misspecified models, except for RF-GAMLSS in winter and spring. The covariates longitude, latitude, distance to the coast, and mean school height exerted the largest effects on typical density ($Q_{0.5}$) and extreme densities ($Q_{0.95}$), with a marked seasonal pattern. Predictive performance was nearly indistinguishable among the distributional models, except for NN-GAMLSS, which was the poorest in winter. GAMLSS and BAMLSS showed consistent predictive performance across all seasons and provide a sound methodological basis for a predictive fisheries service that estimates acoustic density, ecological indices, and exceedance probabilities from uncalibrated commercial echosounder data.

Acoustic density; classical regression; distributional regression; anchoveta schools; predictive fishing service.

INTRODUCCIÓN

La anchoveta peruana (*Engraulis ringens*) es la especie dominante en el ecosistema marino de la Corriente de Humboldt. Desde un enfoque ecológico y económico, se considera el eslabón clave de la cadena trófica y la materia prima para la producción de harina y aceite de pescado de una de las pesquerías más importantes a nivel global.

Esta especie desempeña un papel central en la transferencia de energía en el ecosistema, ya que conecta los niveles tróficos más bajos (el plancton) con los niveles más altos (los depredadores). Se alimenta de plancton, tanto de fitoplancton (microalgas) como de zooplancton (crustáceos pequeños, como eufáusidos y copépodos). Al hacerlo, ayuda a regular las poblaciones de plancton e influye en la base de la pirámide trófica. Estudios han confirmado que el zooplancton, especialmente los eufáusidos, constituye la parte principal de su dieta, mientras que el fitoplancton es un componente secundario (Espinoza y Bertrand, 2008). Es la principal presa de una gran variedad de depredadores, entre los que se incluyen aves guaneras, mamíferos marinos (delfines y lobos marinos) y peces más grandes. La abundancia de la anchoveta es fundamental para el sustento de estas poblaciones y una disminución de sus existencias puede tener efectos en cadena en todo el ecosistema (Pauly y Tsukayama, 1987).

La abundancia y distribución de la anchoveta están fuertemente influenciadas por las condiciones ambientales del sistema de la Corriente de Humboldt, que se caracteriza por su alta variabilidad. La productividad del mar peruano se debe al afloramiento de aguas profundas ricas en nutrientes, que fomentan el crecimiento de fitoplancton y zooplancton, de los que se alimenta la anchoveta (Pauly y Tsukayama, 1987; Chávez et al., 2003).

Los eventos climáticos, como El Niño, provocan cambios en la distribución y concentración de los cardúmenes de anchoveta, lo que afecta a la disponibilidad de esta especie. La pesquería de la anchoveta es la más grande del mundo dedicada a una sola

especie, por lo que su gestión sostenible es crucial. Para ello, la industria y la administración pesquera necesitan información precisa sobre la distribución, la densidad y la biomasa de la anchoveta para establecer las cuotas de pesca y planificar las estrategias de búsqueda para la gestión de la flota pesquera. A fin de garantizar que una extracción se realice de manera responsable, el Instituto del Mar del Perú (IMARPE) se dedica a evaluar la abundancia o biomasa, así como la dinámica y el comportamiento en el espacio y tiempo.

La evaluación de su abundancia y distribución se basa principalmente en técnicas hidroacústicas, que permiten estimar la densidad de los cardúmenes mediante el coeficiente del área de dispersión (s_A o NASC). Esta variable representa la densidad relativa en la columna de agua y constituye la base para la estimación de biomasa y la toma de decisiones en la gestión pesquera.

Sin embargo, la estructura estadística de la densidad acústica presenta propiedades complejas, tales como heterogeneidad de varianza, asimetría extrema y colas pesadas que indican la ocurrencia de eventos extremos. Estas características no corresponden a desviaciones aleatorias, sino a propiedades intrínsecas del proceso ecológico. A pesar de ello, gran parte de los estudios previos han abordado el modelamiento del s_A mediante transformaciones logarítmicas (por ejemplo, $log(s_A + 1)$) y modelos centrados en la media, como los modelos lineales generalizados (GLM) y los modelos aditivos generalizados (GAM). Este enfoque asume, de manera implícita, que la transformación logarítmica permite aproximar la variable a una distribución adecuada para la inferencia, sin evaluar formalmente si tras la transformación se cumple con los supuestos distribucionales del modelo.

Como consecuencia, estos enfoques pueden generar modelos aparentemente bien ajustados en términos de métricas tradicionales, pero con deficiencias en la representación de la variabilidad, la asimetría y el comportamiento de los valores extremos, lo que limita

su capacidad para describir adecuadamente la dinámica de agregación de la anchoveta y para generar predicciones consistentes en la escala original.

Adicionalmente, desde un punto de vista operativo, la disponibilidad de información hidroacústica cuantitativa se encuentra restringida a un número limitado de embarcaciones científicas calibradas. En contraste, la flota pesquera dispone de un gran número de embarcaciones equipadas con ecosondas para la pesca comercial, entre las que se incluyen sistemas acústicos de la marca KONGSBERG SIMRAD, de los modelos ES60 y ES70, de transmisión de onda continua (CW), y el modelo ES80, de última generación, capaz de operar en un amplio rango de frecuencias, desde 18 kHz hasta 333 kHz (incluso 500 kHz), y en el modo estándar de onda continua (CW) utilizada por sus antecesoras. Estos sistemas pueden generar grandes volúmenes de información acústica cuyo aprovechamiento cuantitativo plantea importantes desafíos técnicos y metodológicos.

En este contexto, se plantea la necesidad de desarrollar enfoques de modelamiento estadístico que permitan representar explícitamente la estructura de la densidad acústica en su escala original, incorporando su naturaleza altamente asimétrica y agregada, y que, al mismo tiempo, permitan aprovechar de manera indirecta la información generada por ecosondas comerciales no calibradas. Este enfoque constituye la base para el desarrollo de un servicio pesquero predictivo orientado a mejorar la estimación de la densidad de los cardúmenes y fortalecer la gestión sostenible del recurso.

CAPÍTULO I.

1. Planteamiento del problema

1.1 Descripción de la realidad problemática

La anchoveta peruana (*Engraulis ringens*) constituye la especie dominante del ecosistema asociado a la Corriente de Humboldt y el principal sostén de una de las pesquerías industriales más importantes a nivel mundial (Pauly & Tsukayama, 1987; Ñiquen & Bouchon, 2004). Desde el punto de vista ecológico, desempeña un papel clave en la transferencia de energía dentro de la red trófica marina, mientras que desde el punto de vista económico representa la base de la producción global de harina y aceite de pescado (Chavez et al., 2008).

Para garantizar una gestión sostenible del recurso, el Instituto del Mar del Perú (IMARPE) realiza evaluaciones periódicas de abundancia y biomasa mediante métodos hidroacústicos, los cuales permiten estimar la densidad acústica de los cardúmenes a partir del área de retrodispersión (MacLennan et al., 2002; Simmonds & MacLennan, 2005). La precisión de estas estimaciones depende críticamente de la correcta calibración de las ecosondas científicas (Demer et al., 2015), procedimiento que asegura la confiabilidad cuantitativa de los registros acústicos.

Sin embargo, aunque la flota pesquera industrial dispone de ecosondas comerciales capaces de generar información de alta resolución, la calibración individual de estos equipos resulta compleja y costosa, y no siempre se realiza bajo estándares científicos (Demer et al., 2015). En consecuencia, una proporción considerable de los registros acústicos generados por la flota no puede utilizarse directamente para estimaciones cuantitativas robustas de densidad de cardúmenes.

Adicionalmente, los datos acústicos asociados a la densidad de cardúmenes presentan características estadísticas complejas, tales como asimetría marcada,

heterogeneidad de varianza y presencia de colas pesadas, fenómenos frecuentes en sistemas ecológicos altamente agregados (Clark, 2007; Freckleton, 2009). Los enfoques estadísticos clásicos centrados exclusivamente en la media, aun cuando incorporan transformaciones logarítmicas, pueden resultar insuficientes para representar adecuadamente la estructura distribucional subyacente (Burnham & Anderson, 2002; Halsey, 2019).

En este contexto, diversos estudios orientados a comprender la influencia de la variabilidad oceanográfica sobre los recursos pesqueros han recurrido a enfoques estadísticos clásicos, tales como modelos lineales generalizados (GLM), modelos aditivos generalizados (GAM) e inferencia bayesiana, asumiendo que la transformación logarítmica de la variable respuesta, $log(s_A + 1)$, conduce a una distribución aproximadamente gaussiana o perteneciente a la familia exponencial, como la gamma o la Tweedie (Bertrand et al., 2004; Gutiérrez et al., 2008; Yuffra, 2017; Quiroz et al., 2014; Castillo, 2017; Castillo et al., 2018; Cornejo et al., 2022).

Sin embargo, estos enfoques han asumido, sin una evaluación formal, que la transformación $log(s_A + 1)$ es adecuada para representar la estructura estadística subyacente de la densidad acústica, abordando el exceso de ceros de manera indirecta y sin verificar si dicha transformación preserva propiedades fundamentales del proceso observado, tales como la presencia de colas pesadas, la heterocedasticidad estructural, la proporción de ceros o la forma funcional que relaciona la variable respuesta con las variables explicativas. Esta omisión puede dar lugar a modelos mal especificados, así como a interpretaciones ecológicas y predicciones potencialmente sesgadas de los procesos de agregación y disponibilidad espacial de la anchoveta.

En comparaciones entre modelos ajustados sobre la transformación logarítmica de la densidad acústica de anchoveta, $log(s_A + 1)$, y modelos ajustados en la escala original, las métricas relativas de ajuste [como el coeficiente de determinación (R^2), la desviación

explicada o la validación cruzada generalizada (GCV)] así como las métricas de error dependientes de la unidad (p. ej., MAE y BIAS), no son comparables entre escalas. En consecuencia, la validación y comparación de modelos debe sustentarse en el diagnóstico de residuos y en métricas basadas en la verosimilitud, como el GAIC, incorporando la corrección jacobiana cuando se comparan modelos formulados en distintas escalas (Flores et al., 2026, p. 5).

Por tanto, se evidencia la necesidad de desarrollar enfoques de modelamiento estadístico capaces de representar explícitamente la estructura distribucional de la densidad acústica (s_A) y de integrar información morfológica y espacio-temporal de los cardúmenes, con el fin de establecer bases metodológicas sólidas para un servicio pesquero predictivo que contribuya al aprovechamiento eficiente de la información generada por la flota comercial y al fortalecimiento de la gestión sostenible del recurso.

1.2 Formulación del problema

1.2.1. Problema general

¿La regresión clásica centrada en la media y transformaciones logarítmicas presenta limitaciones frente a la regresión distribucional en escala original, en cuanto a representar adecuadamente la naturaleza heterogénea, altamente asimétrica y no gaussiana de los datos de la densidad acústica s_A de los cardúmenes de anchoveta, lo que limita su aplicabilidad como base de un servicio pesquero predictivo orientado al uso de ecosondas comerciales no calibradas?

1.2.2. Problemas específicos

¿Qué enfoques de modelamiento estadístico, regresión clásica centrada en la media con transformaciones logarítmicas y regresión distribucional en la escala original, resultan adecuados para predecir la densidad acústica s_A de los cardúmenes de anchoveta en el mar peruano?

¿Qué variables explicativas presentan efectos sobre la densidad típica ($Q_{0.5}$) y las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta en el mar peruano?

¿Cómo diagnosticar la especificación residual y el desempeño predictivo de los modelos de regresión clásicos centrados en la media con transformaciones logarítmicas y la regresión distribucional en la escala original?

¿En qué medida los modelos desarrollados pueden servir como base metodológica para un servicio pesquero de predicción de la densidad acústica (s_A) de los cardúmenes de anchoveta para ecosondas comerciales no calibradas?

1.3 Objetivos de la investigación

1.3.1. Objetivo general

Modelar la densidad acústica (s_A) de los cardúmenes de anchoveta mediante regresión clásica centrada en la media con transformaciones logarítmicas y regresión distribucional en escala original, evaluando su adecuación estructural, validez residual y capacidad predictiva para sustentar un servicio pesquero con ecosondas no calibradas.

1.3.2. Objetivos específicos

Aplicar la regresión clásica centrada en la media con transformaciones logarítmicas y regresión distribucional en la escala original, con el fin de evaluar si representan adecuadamente la estructura de los datos, en términos de variabilidad, asimetría y presencia de colas pesadas.

Determinar la magnitud del efecto de las variables explicativas sobre la densidad típica ($Q_{0.5}$) y las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta en el mar peruano.

Diagnosticar los patrones residuales y el desempeño predictivo de los modelos de regresión clásicos centrados en la media y la regresión distribucional en la escala original.

Evaluar la aplicabilidad de los modelos ajustados como base metodológica de un servicio pesquero de predicción de la densidad de los cardúmenes de anchoveta para ecosondas comerciales no calibradas.

1.4 Justificación de la investigación

Desde el punto de vista científico, el análisis predictivo constituye una de las aplicaciones más relevantes del aprendizaje automático y del modelamiento estadístico moderno, al permitir la estimación de variables de interés a partir de conjuntos de variables explicativas.

En la predicción de datos marinos se han aplicado modelos basados en variables ambientales y espaciales, tales como modelos lineales generalizados (Alfatinah et al., 2023), modelos aditivos generalizados (Ospina-Álvarez et al., 2013; Morane et al., 2020), árboles de decisión (Alfatinah et al., 2023), bosques aleatorios (Morane et al., 2020), redes neuronales artificiales (Armas, 2024) e inferencia bayesiana (Quiroz et al., 2014; Castillo, 2017).

No obstante, los datos pesqueros presentan características estadísticas particulares: alta heterogeneidad, marcada asimetría y fuerte sesgo hacia valores elevados, lo que supone un desafío para los enfoques tradicionales. En las ciencias ambientales, ecológicas y pesqueras, muchos fenómenos no pueden estudiarse mediante experimentación controlada; por ello, el análisis estadístico de datos observacionales constituye una herramienta central para la inferencia científica (Maynard Smith, 1978; Harvey & Pagel, 1991).

La aplicación acrítica de métodos clásicos ha mostrado limitaciones cuando se emplea en sistemas complejos caracterizados por alta variabilidad, dependencia estructural y procesos no lineales.

En particular, la práctica de basar la inferencia científica en valores p y en el coeficiente de determinación (R^2) ha sido ampliamente cuestionada en la literatura

metodológica contemporánea (Hilborn & Mangel, 1997; Burnham & Anderson, 2002; Freckleton, 2009; Halsey, 2019). En conjuntos de datos grandes y complejos pueden detectarse efectos estadísticamente significativos aun cuando estos sean débiles o ecológicamente irrelevantes.

En tales casos, la significancia estadística pierde valor informativo y la interpretación dicotómica basada en valores p puede conducir a conclusiones engañosas, sobredimensionando predictores con contribución marginal y desplazando el análisis hacia la mera detección de significancia en lugar de evaluar la magnitud y relevancia ecológica del efecto (Burnham & Anderson, 2002; Freckleton, 2009; Nakagawa & Cuthill, 2007; Halsey, 2019).

De manera similar, el uso del coeficiente de determinación (R^2) como medida principal de ajuste o desempeño resulta problemático. Se ha documentado que modelos con valores bajos de R^2 pueden presentar términos estadísticamente "significativos", lo que evidencia una limitada capacidad explicativa pese a la aparente validez estadística (Freckleton, 2009). Además, R^2 no constituye una medida adecuada para la comparación entre modelos, ya que no penaliza apropiadamente la complejidad y puede incrementarse artificialmente al incorporar un mayor número de parámetros. Este problema se acentúa cuando los datos se analizan en escalas logarítmicas que abarcan varios órdenes de magnitud, donde valores elevados de R^2 pueden ocultar una variabilidad sustancial en la escala original y generar una falsa percepción de buen ajuste (Smith, 1980; Nee et al., 2005; Freckleton, 2009).

Frente a estas limitaciones, la literatura metodológica propone priorizar el análisis del tamaño del efecto, la comparación entre modelos alternativos, el uso de criterios de información (AIC/GAIC) y pesos de evidencia para evaluar el soporte relativo de los

modelos, en lugar de depender de R^2 y valores p como ejes centrales de validación (Hilborn & Mangel, 1997; Burnham & Anderson, 2002; Freckleton, 2009).

Este enfoque reconoce que, en sistemas ecológicos reales, la heterocedasticidad, la asimetría y las colas pesadas no constituyen ruido, sino componentes intrínsecos del proceso observado; por tanto, la inferencia debe centrarse en la magnitud, consistencia e importancia relativa de los efectos (Nakagawa & Cuthill, 2007; Halsey, 2019).

En la estadística contemporánea, una limitación fundamental de los enfoques clásicos de regresión (incluidos los modelos lineales, GLM y GAM) es que restringen la relación entre la variable respuesta y los predictores exclusivamente al valor esperado, asumiendo implícitamente que la varianza, la asimetría y la forma de la distribución son constantes o irrelevantes para la inferencia (Stasinopoulos et al., 2017; Rigby et al., 2019; Stasinopoulos et al., 2024).

Esta suposición resulta especialmente problemática en sistemas ambientales y pesqueros, donde la heterogeneidad, la asimetría, las colas pesadas, los extremos y la inflación de ceros constituyen rasgos estructurales del proceso.

La regresión distribucional moderna ha demostrado que, cuando estas propiedades no se modelan explícitamente, la inferencia basada en modelos centrados en la media puede ser estructuralmente incorrecta, incluso cuando los modelos presentan buen ajuste aparente o términos estadísticamente significativos. En tales casos, los residuos exhiben patrones sistemáticos que indican desajustes en la media, la varianza y los extremos de la distribución, lo que compromete la validez de la inferencia estadística y ecológica derivada del modelo (Stasinopoulos et al., 2017; Rigby et al., 2019; Stasinopoulos et al., 2024).

En consecuencia, el uso exclusivo de modelos centrados en la media y transformaciones logarítmicas para describir la densidad acústica de los cardúmenes de

anchoveta puede conducir a modelos mal especificados, inferencias parciales y predicciones sesgadas en un sistema caracterizado por agregación extrema y alta variabilidad ambiental.

El marco de los modelos aditivos generalizados para localización, escala y forma (GAMLSS) surge para superar estas limitaciones, al permitir modelar la distribución condicional completa de la variable respuesta y no únicamente su media. En este enfoque, los parámetros de localización, dispersión, asimetría y curtosis pueden depender explícitamente de covariables ambientales, espaciales y temporales, representando la heterogeneidad estructural, las colas pesadas, la inflación de ceros y los extremos como componentes centrales del análisis (Stasinopoulos et al., 2017; Rigby et al., 2019; Stasinopoulos et al., 2024; Flores et al., 2026).

Bajo este enfoque distribucional, la validación del modelo se fundamenta en criterios basados en la verosimilitud, el diagnóstico distribucional de residuos (residuos cuantílicos normalizados, diagramas de gusano, diagramas de cubo) y la comparación formal entre modelos mediante el criterio de información generalizado de Akaike (GAIC).

Desde el punto de vista técnico y operativo, la elevada variabilidad ambiental del mar peruano exige intensificar el monitoreo del ecosistema marino para comprender con mayor precisión la distribución y abundancia de los recursos pesqueros. En el ámbito internacional, organismos como la FAO y el ICES promueven enfoques colaborativos de gestión que integran a científicos, administradores y a la propia industria, reconociendo el valor estratégico de los datos generados por embarcaciones comerciales como complemento de las campañas científicas.

Los datos recopilados por la flota pesquera comercial pueden ser tan relevantes como los obtenidos en cruceros de investigación, ya que proporcionan información detallada sobre la ubicación espacial de los recursos, la distribución espacio-temporal de presas (eufáusidos, zooplancton, peces mesopelágicos y micronecton), así como variables ecológicas

(termoclina, oxiclina y zona mínima de oxígeno) y oceanográficas (temperatura, salinidad y oxígeno disuelto).

Al estar equipadas con sistemas acústicos, las embarcaciones comerciales se convierten en plataformas permanentes de observación, capaces de generar información digital útil para modelos ecológicos y para la evaluación de poblaciones de peces.

En el caso peruano, la flota industrial dedicada a la captura de anchoveta está compuesta por aproximadamente 320 embarcaciones equipadas con ecosondas comerciales de última generación, como SIMRAD (modelos ES60, ES70 y ES80), FURUNO (incluida la versión científica FSV-30) y Sonic Corporation (Sonic Kaijo). Estos sistemas, que operan con tecnología similar a las ecosondas científicas, tienen capacidad para generar información acústica de alta resolución susceptible de análisis cuantitativo.

Desde el punto de vista económico, aunque las ecosondas comerciales pueden calibrarse mediante blancos estándar (Demer et al., 2015), este procedimiento debe realizarse de forma individual para cada equipo y resulta complejo de implementar en el gran número de embarcaciones existentes. La calibración puede verse afectada por factores operativos como la presencia de peces en la zona, la intensidad del oleaje o del viento (Flores et al., 2025), lo que incrementa la dificultad técnica y operativa del proceso.

Además, el desplazamiento de cada embarcación hacia zonas adecuadas para realizar la calibración implica costos adicionales de operación, tiempo y logística. Como consecuencia, aproximadamente el 80 % de los sistemas acústicos utilizados por la flota industrial no pueden calibrarse en condiciones óptimas, lo que limita el aprovechamiento cuantitativo de la información generada.

En este contexto, el desarrollo de un marco metodológico que permita utilizar de manera indirecta la información proveniente de ecosondas comerciales no calibradas representa una alternativa económicamente eficiente, al incrementar el valor informacional

de infraestructura ya instalada, reducir costos asociados a procesos de calibración masiva y optimizar el uso de datos disponibles para la gestión pesquera.

El estudio aborda un cambio de paradigma de la regresión clásica centrada en la media condicional, donde la varianza, asimetría y curtosis quedan restringidas por la familia asumida, hacia la regresión distribucional, la cual permite especificar un predictor propio para cada parámetro de la respuesta, como base para el desarrollo de un marco de modelamiento estadístico flexible y coherente con la estructura real de los datos acústicos. Dicho marco se entrena a partir de información procedente de un conjunto reducido de embarcaciones equipadas con ecosondas científicas calibradas, y posteriormente puede extenderse, de manera indirecta, al aprovechamiento de descriptores morfológicos e información espacio-temporal derivados de los registros generados por ecosondas comerciales no calibradas.

Este enfoque puede permitir ampliar de manera consistente la base de datos disponible, mejorar la estimación de la densidad acústica de anchoveta a nivel de cardúmenes y establecer una base metodológica sólida para el desarrollo futuro de un servicio pesquero predictivo, orientado a optimizar las operaciones de pesca y a fortalecer la gestión sostenible del recurso.

1.5 Delimitaciones del estudio

La presente investigación se delimita al análisis de datos acústicos recopilados mediante ecosondas científicas SIMRAD EK60, operadas a una frecuencia de 120 kHz, considerando exclusivamente información proveniente de plataformas científicas previamente calibradas, y excluyendo el uso de datos de ecosondas comerciales tanto en la etapa de entrenamiento como en la validación del modelo.

El estudio no aborda la predicción de la distribución espacial de la densidad acústica ni la identificación de zonas de pesca de anchoveta ni el nicho ecológico que son

condicionados explícitamente por variables oceanográficas o ambientales. Los datos analizados corresponden a cardúmenes individuales y no a la integración del eco en unidades básicas de muestreo (UBM) de 1 milla náutica, las cuales combinan ceros ($s_A = 0$) y valores positivos ($s_A > 0$) y son la base para representar procesos de presencia–ausencia y patrones de distribución espacial de la especie. En consecuencia, el modelo se restringe a la densidad acústica a escala de cardumen ($s_A > 0$), entidad que no constituye una unidad espacial de muestreo y que puede extenderse parcial o totalmente sobre una o varias UBM; por tanto, las variables oceanográficas no forman parte del análisis.

Asimismo, no se incorporan procesos de validación biológica directa, evaluaciones poblacionales independientes ni estimaciones operativas de biomasa. En consecuencia, el alcance de la investigación se circunscribe al análisis metodológico y predictivo de la estructura estadística de la densidad acústica observada, en el marco de la comparación entre enfoques de modelamiento centrados en la media y enfoques distribucionales.

CAPÍTULO II.

2. MARCO TEÓRICO

2.1 Antecedentes de la investigación

2.1.1. Investigaciones internacionales

Ospina-Alvarez et al. (2013) aplicaron un modelo aditivo generalizado (GAM, por sus siglas en inglés) con un vínculo logarítmico y una estructura de error cuasi-Poisson para estimar la variación espacial en la abundancia de peces y huevos de la anchoveta europea (*Engraulis encrasicolus*).

Para construir el modelo predictivo de la producción interanual de huevos, se integraron datos biológicos y acústicos (biomasa y estructura de tallas, proporción de sexos, modelo de fecundidad por lotes de peso, peso medio, número de peces y desove medio). Los resultados indican que el modelo se contrastó satisfactoriamente con los datos de campo entre 2003 y 2010 obtenidos mediante el método de producción diaria de huevos (MPD) en el Mediterráneo noroccidental.

Este estudio aporta evidencia metodológica respecto a que integra biología y acústica mediante GAM. No obstante, la validación del modelo se centra en métricas relativas, complementándose con comparaciones descriptivas frente a estimaciones independientes obtenidas mediante el método de producción diaria de huevos (MPD). En consecuencia, la evaluación del desempeño del modelo se mantiene dentro del marco clásico de modelamiento centrado en la media, sin una validación explícita de los supuestos distribucionales de los residuos.

Morane et al. (2020) utilizaron GAM y bosques aleatorios (RF, por sus siglas en inglés) para evaluar la distribución espacial y temporal de la biomasa de peces mesopelágicos en el mar Mediterráneo entre 1994 y 2011. Este estudio fue pionero en

estimar la biomasa y los patrones espacio-temporales de estos peces usando variables ambientales como predictoras.

Los resultados indican que ambos modelos (GAM y RF) sugirieron una dinámica temporal diferente en la biomasa total de peces mesopelágicos. El modelo GAM mostró un aumento de la biomasa desde 1995, seguido de un descenso suave hasta 2007, tras el cual subió abruptamente de nuevo. Por su parte, el modelo RF mostró un descenso de la biomasa en 1998 que se mantuvo estable entre 2001 y 2011.

Este estudio aporta evidencia metodológica al emplear enfoques basados GAM y RF; sin embargo, la inferencia se restringe al comportamiento del valor esperado en la escala transformada, sin incorporar una validación basada en el diagnóstico formal de los residuos ni una evaluación explícita de los posibles sesgos introducidos por el cambio de escala simultáneo de la respuesta y los predictores.

Alfatinah et al. (2023) desarrolló un método para predecir las áreas de pesca más productivas (puntos calientes) de atún listado (*Katsuwonus pelamis*) mediante el uso de árboles de decisión (DT, por sus siglas en inglés) y modelos lineales generalizados (GLM, por sus siglas en inglés) basados en diferentes «escenas» o condiciones ambientales históricas, como datos de sensores remotos y observaciones de campo, que incluyen la temperatura y la altura de la superficie del mar, así como la concentración de clorofila-a.

El estudio concluye que el uso de modelos de conjunto basados en escenas, que adaptan la elección del modelo predictivo a las condiciones ambientales específicas del momento, es un método eficaz para pronosticar las áreas de pesca del atún listado. Estos resultados proporcionan una base científica sólida para optimizar las operaciones de pesca, reducir costes y promover una gestión pesquera más sostenible.

Desde el punto de vista metodológico, el aporte de esta investigación radica en la predicción espacial discreta de áreas de pesca, formulada como un problema de

clasificación, mediante GLM, que asume una relación funcional lineal entre la variable de respuesta y los predictores ambientales, y DT, que particionan el espacio de los predictores en regiones discretas mediante reglas jerárquicas del tipo "si–entonces", generando predicciones categóricas en función de umbrales en las covariables.

La inferencia se centra en la clasificación espacial y en la evaluación del desempeño predictivo mediante métricas de clasificación y concordancia. Por lo cual, este enfoque no incorpora un diagnóstico formal de los residuos ni un análisis explícito de la variabilidad, asimetría o comportamiento extremo de la variable de interés, tratada como un proceso continuo con estructura probabilística explícita.

Armas (2024) propuso un modelo basado en redes neuronales artificiales (NN**,** por sus siglas en inglés) para predecir la ubicación de la anchoveta, utilizando datos georreferenciados de capturas diarias de barcos cerqueros industriales entre 2003 y 2020, e información de variables oceanográficas (temperatura superficial del mar, salinidad, profundidad de la capa de mezcla, altura del mar y corrientes), obtenidas a partir de imágenes satelitales.

El modelo de NN se utilizó para generar pronósticos a corto plazo de las áreas con mayor potencial de pesca e identificar la distribución espacial y temporal de las probabilidades de pesca de anchoveta, especialmente en los años El Niño (2015), La Niña (2007, 2013 y 2020) y neutro (2004). El modelo demostró ser eficaz para generar pronósticos a corto plazo de las zonas de pesca con mayor potencial, lo que puede contribuir a una pesca más eficiente y a la gestión sostenible de los recursos.

Desde el punto de vista metodológico, el aporte de este antecedente radica en la aplicación de redes neuronales artificiales para la predicción espacial de áreas con potencial de pesca, integrando información oceanográfica y registros georreferenciados de capturas a escala operativa. La validación del modelo se sustentó en métricas de clasificación (p. ej.,

exactitud, sensibilidad, especificidad y curvas ROC), orientadas a evaluar la capacidad predictiva de presencia–ausencia de pesca.

En consecuencia, el alcance del trabajo se restringe al ámbito operativo de la predicción de capturas, sin inferencia directa sobre la intensidad del proceso ni sobre la estructura estadística subyacente de la variable biológica, ni un análisis explícito de la variabilidad, asimetría o comportamiento extremo del fenómeno observado.

Izquierdo (2026) analizó el efecto de la complejidad en modelos de evaluación de stocks mediante estudios empíricos y simulaciones aplicadas a la merluza europea (*Merluccius merluccius*) y al atún aleta amarilla (*Thunnus albacares*), mediante modelos bayesianos implementados con Aproximación de Laplace Integrada Anidada (INLA, por sus siglas en inglés) para refinar los datos de entrada, incluyendo modelos delta jerárquicos con dependencia espaciotemporal para estimar índices de reclutamiento y modelos multivariantes para corregir sesgos en parámetros biológicos.

Asimismo, evalúa configuraciones estructurales en Stock Synthesis 3 (SS3), como modelos de uno y dos sexos y formulaciones espaciales con procesos de movimiento y reclutamiento regional. Sus resultados muestran que la mayor complejidad puede mejorar la representación de la dinámica poblacional, pero también generar sesgos e identificabilidad cuando no está bien sustentada. En conjunto, concluye que el aumento de complejidad no garantiza mejoras en el desempeño, siendo útil solo cuando está guiado por hipótesis biológicas, se introduce de forma gradual y cuenta con datos suficientes.

Este antecedente aporta al presente estudio al evidenciar que incluso modelos complejos, como los implementados en INLA, cuando se evalúan mediante métricas relativas de desempeño y sin diagnóstico residual con herramientas avanzadas, pueden conducir a la falsa conclusión de que modelos simples, aparentemente bien ajustados, son superiores a modelos más complejos.

2.1.2. Investigaciones nacionales

Joo (2008) propuso utilizar NN y datos de posicionamiento satelital para identificar automáticamente las zonas de pesca de la anchoveta. La investigación demostró que las NN son una herramienta eficaz para identificar automáticamente los puntos de actividad pesquera a partir de los datos de posicionamiento satelital de las embarcaciones pesqueras de anchoveta.

Este método automatiza el proceso de identificación, lo que mejora la eficiencia y proporciona información sobre la distribución espacial del esfuerzo pesquero, contribuyendo así a la gestión sostenible de los recursos marinos.

Desde el punto de vista metodológico, el aporte de este antecedente al presente estudio radica en la aplicación de redes neuronales artificiales para la identificación automática de zonas de pesca, a partir de datos de posicionamiento satelital de embarcaciones, orientada a caracterizar la distribución espacial del esfuerzo pesquero. La validación del modelo se sustentó en métricas de clasificación, enfocadas en la detección de eventos de pesca y en la capacidad del algoritmo para identificar correctamente puntos de actividad extractiva. En consecuencia, el alcance del estudio se restringe al análisis del comportamiento operativo de la flota y a la predicción de eventos de pesca, sin modelar variables biológicas continuas, como la densidad acústica o la biomasa, ni realizar inferencia estadística sobre la estructura ecológica subyacente de la anchoveta.

Quiroz et al. (2014) desarrollan un enfoque bayesiano basado en INLA para la estimación y predicción de la biomasa de anchoveta (*Engraulis ringens*) frente a la costa del Perú, utilizando datos acústicos recolectados por IMARPE durante un crucero científico en 2005. El estudio propone un modelo jerárquico Bayesiano tipo hurdle, que separa el proceso de ausencia–presencia (masa puntual en cero) del componente positivo de la biomasa,

modelado mediante distribuciones continuas alternativas (gamma, lognormal y log-logística).

La inferencia se realiza mediante INLA, incorporando efectos espaciales latentes con estructura Matérn y permitiendo la predicción de la biomasa en ubicaciones no muestreadas sobre una malla espacial regular. Los predictores fueron distancia a la costa, profundidad del fondo marino y posición geográfica.

Los resultados indican que la ausencia de anchoveta es mayor cuando la distancia a la costa aumenta, es decir, en aguas más alejadas de la costa donde el océano es más profundo. La presencia de anchoveta es mayor en la región central, un área caracterizada por oxígeno superficial y agua fría, dulce y somera en las capas superficiales costeras que proporcionan un hábitat ideal para la anchoveta (Bertrand et al., 2011).

La ausencia/presencia de anchoveta en algunos lugares no es independiente de la ausencia/presencia en lugares vecinos. Se obtiene la misma conclusión para la biomasa positiva de anchoveta, lo que respalda la necesidad de efectos espaciales.

Este estudio constituye un antecedente relevante, ya que a pesar de aplicar un modelo avanzado basado en INLA, replica las limitaciones de la estadística clásica al emplear en el componente positivo una distribución simple como la Gamma en la escala logarítmica. Asimismo, la adecuación del modelo se asume a partir de criterios globales relativos, como el criterio de información de desviación (DIC) y log pseudo-verosimilitud marginal (LPML), métricas que evalúan el ajuste relativo pero no la correcta especificación del modelo.

Castillo (2017) aplicó una combinación de modelos aditivos generalizados (GAM), geoestadística, modelos jerárquicos bayesianos espacio–temporales implementados mediante INLA, y modelos de series temporales (Box–Jenkins y espacio–estado) para describir la influencia de variables ambientales como temperatura, salinidad y oxígeno sobre la distribución y el área total de ocupación de la anchoveta peruana (*Engraulis ringens),*

integrando datos provenientes de 67 cruceros hidroacústicos realizados por IMARPE entre 1985 y 2015.

El estudio empleó un modelo GAM con distribución Tweedie para analizar el s_A transformado logarítmicamente en función de variables ambientales y espaciales, así como un modelo jerárquico bayesiano formulado mediante dos componentes separados (Bernoulli para la presencia de ceros y Gamma para los valores positivos), manteniendo en ambos casos una inferencia centrada en la media condicional.

Los resultados evidenciaron una marcada dependencia espacial y temporal, una estructura semicontinua con presencia de ceros y una elevada variabilidad asociada a forzantes oceanográficos estacionales y a eventos El Niño, las cuales fueron abordadas principalmente mediante transformaciones logarítmicas del s_A y el uso de distribuciones pertenecientes a la familia exponencial. Asimismo, el estudio demostró que la incorporación explícita de autocorrelación espacial y efectos dinámicos mejora la capacidad predictiva frente a enfoques puramente descriptivos.

Este estudio constituye un antecedente metodológico relevante, ya que, pese a incorporar mayor cobertura de datos y modelamiento espacio–temporal, no logra mejoras sustantivas respecto a Quiroz et al. (2014), al mantener esquemas de validación de enfoques anteriores sin evaluar la correcta especificación ni el desempeño predictivo. Al centrarse en la media condicional mediante transformaciones logarítmicas y distribuciones simples (Gamma, Tweedie), refuerza la necesidad de enfoques distribucionales en la escala original de la densidad acústica (s_A).

Flores et al. (2025) ajustaron un modelo para predecir la densidad de los cardúmenes de anchoveta peruana (*Engraulis ringens*) a partir de descriptores morfológicos del cardumen (alto, largo y volumen 3D) y factores espacio-temporales del ambiente marino (periodicidad horaria, profundidad media, distancia a la costa y posición geográfica). Para

ello, emplearon modelos aditivos generalizados para localización, escala y forma (GAMLSS) y su extensión modelos aditivos bayesianos para localización, escala y forma (BAMLSS), estimando simultáneamente todos los parámetros de la distribución (μ, σ, ν y τ) del logaritmo de la densidad acústica.

Este estudio es el precursor directo de la presente investigación, al introducir explícitamente un enfoque distribucional en el análisis de la densidad acústica de los cardúmenes. Los resultados demostraron que, aun bajo transformación logarítmica, la variable de respuesta no satisface los supuestos de normalidad y que la complejidad estructural del s_A persiste después de la transformación.

En particular, se evidenció que para representar adecuadamente dicha estructura es necesaria una familia distribucional flexible, como la Box-Cox Power Exponencial de cuatro parámetros (μ, σ, ν y τ), capaz de capturar asimetría y curtosis. Esta evidencia motiva el desarrollo posterior de enfoques formulados directamente en la escala original de la densidad acústica, evitando las distorsiones inducidas por transformaciones logarítmicas.

Flores et al. (2026) analizaron datos de 25 cruceros hidroacústicos del IMARPE entre 2007 y 2017, para evaluar el efecto de variables oceanográficas y factores espacio-temporales sobre la densidad acústica (s_A) de la anchoveta peruana (*Engraulis ringens*). Para ello, aplicaron modelos completos GAMLSS ajustados a cero mediante la extensión gamlss.inf, combinando dos modelos: un modelo para los valores positivos del s_A basado en las distribuciones flexibles de 3 y 4 parámetros, Gaussiana Inversa Generalizada y Box-Cox t, respectivamente, que permiten representar una cola pesada derecha; y un modelo logístico binario para la probabilidad de ausencia (π_0).

El estudio demostró que modelar s_A en su escala original, sin transformaciones logarítmicas, permite preservar la asimetría, leptocurtosis, heterogeneidad estructural y la proporción de ceros, aspectos que se distorsionan bajo enfoques tradicionales basados en

transformaciones logarítmicas como $log(s_A + 1)$. Asimismo, la comparación entre modelos en escala original y escala logarítmica mediante el criterio de información generalizado de Akaike (GAIC) corregido con el jacobiano evidenció que las transformaciones inducen pérdida sustancial de información (ΔGAIC > 10) y de soporte empírico, afectando la interpretación ecológica de los procesos de agregación extrema y de la probabilidad de ausencia (π_0).

Este antecedente proporciona la base metodológica que sustenta el enfoque distribucional del presente estudio, al justificar el uso de modelos formulados en la escala original para caracterizar adecuadamente la estructura ecológica de la densidad acústica, evitando la pérdida de información inducida por transformaciones logarítmicas propias de los enfoques estadísticos tradicionales.

2.2 Bases teóricas

2.2.1. Sistemas de detección acústica

La detección acústica de recursos pesqueros depende de ecosondas, equipos también usados en mapeo batimétrico y estudios de sedimentos, entre otras aplicaciones. En el Perú, dos tipos de estas ecosondas operan según su función: las de uso científico y las de uso comercial.

Las instituciones dedicadas a la investigación pesquera trabajan con ecosondas SIMRAD de la serie científica (modelos EK60 y EK80) construidas específicamente para evaluar organismos vivos mediante tecnología digital.

La flota que captura especies pelágicas, entre ellas la anchoveta, opera en cambio con ecosondas SIMRAD de la serie comercial (ES60, ES70, ES80). Aunque destinadas a la actividad pesquera, estas comparten los transductores Split beam de los equipos científicos, lo que permite calibrarlas con el mismo procedimiento y nivel de precisión (Demer et al., 2015).

2.2.2. Ecoconteo y Ecointegración

Aplicaciones específicas de la Acústica Pesquera son la determinación de la abundancia por métodos como el ecoconteo y ecointegración (Ona, 1999). El principio fundamental de la ecointegración es la denominada linearidad acústica, la cual consiste en lo siguiente: si la cantidad de energía que se emite durante la transmisión de sonido es conocida, y se puede medir con precisión la que se recibe como eco, entonces la diferencia entre ambos niveles de energía es atribuida a los dispersores (peces, por ejemplo) que existen en la columna de agua.

Dichas diferencias, constituidas por los ecos que provienen de los peces insonificados son acumulados por un eco integrador, que es el dispositivo a cargo del conteo de la acumulación de energía sonora en intervalos de tiempo o distancia predefinidos. Así, a mayor cantidad de peces la ecointegración aumenta linealmente. El equipo fundamental para realizar esta labor es un ecosonda (Simmonds y MacLennan, 2005). Anteriormente la ecointegración requería de un equipo específico (eco integrador); hoy en día esta labor puede realizarse a través de un programa informático a costos comparativos mucho menores.

2.2.3. Calibración de sistemas acústicos

La calibración periódica de los instrumentos acústicos es fundamental para garantizar la fiabilidad y precisión de los datos obtenidos, que pueden utilizarse de manera cuantitativa para calcular la densidad de recursos marinos, como peces o zooplancton. Además, ayuda a supervisar el funcionamiento del instrumento a lo largo del tiempo, ya que permite detectar cualquier desviación o mal funcionamiento e identificar cambios en el rendimiento debidos a factores como el deterioro, fallos en los componentes o influencias ambientales (Demer et al., 2015).

2.2.4. Detección acústica

El principio de funcionamiento de la ecosonda consiste en generar, amplificar y transmitir un pulso acústico que viaja a través de la columna de agua. Cuando este pulso encuentra un blanco (ya sea zooplancton, peces o el fondo marino), se genera una onda de retorno, reflejada o dispersada, cuyas características (amplitud, forma y frecuencia) quedan determinadas por las propiedades del objeto detectado. Es precisamente en estas variaciones donde reside el fundamento de la detección remota.

El proceso se completa cuando una porción de esa onda dispersada llega nuevamente al equipo: allí se transforma en una señal de voltaje, se amplifica, se filtra el ruido y se digitaliza, obteniéndose así la señal de retrodispersión (Furusawa, 1991).

2.2.5. Coeficiente del área de dispersión

El coeficiente de dispersión del área s_A, comúnmente conocido como Coeficiente Náutico de Dispersión Acústica (NASC por sus siglas en inglés), es una medida fundamental utilizada en hidroacústica cuantitativa para estimar la abundancia relativa o la biomasa de los organismos acuáticos (Rogers et al., 2022).

El s_A representa la integral de la intensidad del volumen de retrodispersión S_v en un intervalo de profundidad o rango específico, normalmente toda la columna de agua o una capa específica (por ejemplo, desde unos pocos metros por debajo de la superficie hasta el lecho marino). Cuantifica la energía acústica total retrodispersada por los organismos dentro de esa zona definida, normalizada por la zona insonificada en ese rango (Nottestad et al., 2021; Baptist et al., 2018).

Esencialmente, es una medida de la "reflectividad" acústica por unidad de superficie marina (m^2 de la sección transversal de dispersión por m^2 de superficie marina) y sirve como indicador de la densidad o biomasa de los peces cuando se escala mediante los factores de conversión adecuados, como la fuerza de blanco (TS) y la información específica de cada

especie. La unidad estándar del s_A es metros cuadrados por milla náutica al cuadrado (m^2/mn^2).

El coeficiente de dispersión del área (s_A en m^2/mn^2 para una capa de profundidad específica) se define como la integral del coeficiente del volumen de retrodispersión (s_v en m^{-1}) en el intervalo de rango (profundidad) desde:

$$s_A = \int_{r_1}^{r_2} s_v \, dr$$

Donde r_1 y r_2 son los límites superior e inferior (profundidades) del rango de la capa integrada (m). s_v es el coeficiente del volumen de retrodispersión, que está relacionado con la intensidad media del volumen de retrodispersión (s_v) en decibeles (dB) mediante la ecuación $s_v = 10 \log_{10}(\frac{S_v}{A_0})$, donde $A_0 = 1$ m^2 (Korneliussen et al. (2018).

En la práctica, la cantidad utilizada para la evaluación de las poblaciones de peces suele ser el coeficiente de dispersión del área náutica (NASC), que es s_A escalado a una base por milla náutica cuadrada:

NASC (o s_A en m^2/mn^2) = s_A (en m^2/m^2) × $(1852)^2$

Donde 1852 es el número de metros en una milla náutica. El valor NASC puede entonces relacionarse directamente con la densidad de peces (ρ, número por unidad de área) utilizando la sección transversal de retrodispersión media ($\bar{\sigma}_{bs}$) de los peces individuales de la población (Doray et al., 2010):

$$\rho = \frac{NASC}{\bar{\sigma}_{bs}}$$

Esto permite estimar la abundancia total de peces y la biomasa en una zona determinada.

2.2.6. Ecosondas científicas

Se trata de dispositivos que emiten y reciben ondas acústicas, diseñados en hardware y software para la investigación científica. Al ser calibrados, miden componentes biológicos subacuáticos, como los peces, la vegetación acuática y el plancton (Demer et al., 2015).

2.2.7. Ecosondas comerciales

Se trata de dispositivos que emiten y reciben ondas acústicas, diseñados tanto en hardware como en software, para su uso en la pesca profesional o comercial (Demer et al., 2015).

2.2.8. Servicio pesquero predictivo

Se define como un sistema analítico aplicado orientado a generar predicciones evaluables y reproducibles para optimizar las operaciones de pesca, mejorar la eficiencia y respaldar la toma de decisiones en la actividad pesquera, superando los enfoques puramente descriptivos (Hilborn & Mangel, 1997; Burnham & Anderson, 2002; Vásquez-Quispesivana et al., 2022; Chen et al., 2024; Pantraki et al., 2024). Este enfoque se enmarca dentro de la evaluación de stock moderna, donde el objetivo no es solo describir el estado histórico de un recurso, sino generar pronósticos verificables sobre su comportamiento futuro bajo distintos escenarios de explotación (Hilborn & Walters, 1992). La capacidad predictiva del sistema permite, además, su integración en marcos de evaluación de estrategias de manejo (management strategy evaluation), donde modelos alternativos se contrastan explícitamente por su desempeño predictivo antes de ser adoptados como base para decisiones de gestión (Punt et al., 2016). Bajo esta lógica, la selección de modelos deja de basarse únicamente en el ajuste a los datos observados y pasa a evaluarse por su capacidad de generalización y reproducibilidad frente a nueva información (Burnham & Anderson, 2002).

2.2.9. Densidad acústica de los cardúmenes de anchoveta

La densidad acústica de los cardúmenes de anchoveta se define conceptualmente como una variable continua que representa la intensidad relativa de biomasa agregada detectada por sistemas acústicos, expresada mediante el coeficiente del área de dispersión acústica (s_A o NASC). En este estudio, dicha densidad se asume como el resultado de procesos ecológicos de agregación espacial intensa, caracterizados por alta variabilidad, asimetría positiva y presencia de valores extremos, lo que exige enfoques estadísticos flexibles para su adecuada representación (Simmonds & MacLennan, 2005; Clark, 2007).

2.2.10. Agregación extrema de cardúmenes

La agregación extrema se conceptualiza como el patrón ecológico mediante el cual una proporción sustancial de la biomasa total se concentra en un número reducido de eventos de alta densidad espacial. En especies pelágicas como la anchoveta, este comportamiento genera distribuciones altamente asimétricas en las variables acústicas, donde los valores extremos constituyen una parte estructural del proceso ecológico y no observaciones atípicas (Fréon & Misund, 1999; Clark, 2007).

2.2.11. Análisis predictivo

Es un proceso que utiliza datos históricos y actuales, combinado con técnicas de inteligencia artificial, aprendizaje automático (machine learning) y modelado estadístico, para identificar patrones y hacer predicciones sobre eventos, comportamientos o resultados futuros con alta precisión. Mediante un proceso iterativo, se desarrolla el modelo predictivo empleando un conjunto de datos de entrenamiento y es validado con datos de prueba, para determinar su precisión y mejorar los pronósticos. El objetivo general es predecir resultados futuros basándose en datos conocidos (Flores et al., 2025).

2.2.12. Variable de Respuesta

También conocida como variable dependiente, es el efecto o resultado que se mide y que se espera que cambie debido a la variable explicativa.

2.2.13. Variable Explicativa

También conocida como variable independiente, es la causa, o el factor que el investigador manipula o cambia para observar sus efectos.

2.2.14. Inferencia estadística

La parte estocástica de un modelo estadístico suele surgir de la suposición de que los datos observados son una muestra de una población mayor (desconocida), cuyas propiedades intentamos estudiar. El modelo estadístico es una simplificación de la población y su comportamiento. Para comprender los conceptos básicos de la inferencia estadística, son esenciales los conceptos de población, muestra y modelo.

2.2.14.1. Población

La población es el conjunto de sujetos que pretendemos estudiar. Por lo general, nos interesan algunas de sus características, las cuales se representan mediante un conjunto de variables, por ejemplo, *Y*. Las poblaciones pueden ser reales, como la población de anchoveta considerada en esta tesis, o conceptuales. El número de elementos de la población es finito, por ejemplo, *N*, aunque puede ser muy grande.

El rango de valores que puede asumir la característica poblacional *Y* depende del tipo de variable, es decir, de si esta es continua, discreta o categórica. En la práctica, incluso para las variables continuas, las mediciones observadas en la población solo pueden asumir un número finito de valores, dado que *N* es finito. Por lo tanto, el conjunto de todos los valores observados posibles de *Y* puede considerarse discreto, incluso cuando Y sea una variable

continua en el sentido matemático convencional, aunque dicho conjunto puede ser muy grande.

En el caso de una variable discreta, algunos valores de Y en la población pueden ser idénticos. Sea D el número de valores distintos de Y en la población y N_i el número de veces que aparece el valor distinto Y_i. Entonces, el tamaño total de la población es

$$\mathrm{N} = \sum_{I=1}^{D} N_i$$

Si

$$P_I = \frac{N_i}{N}$$

la distribución poblacional está definida por una función de probabilidad (pf) dada por:

$$f_P(y) = \mathrm{P}(\mathrm{Y} = \mathrm{y}) = \begin{cases} P_i & si\ y = \mathrm{Y}_I, \quad I = 1, 2, \ldots, D \\ 0 & en\ caso\ contrario \end{cases}$$

Es decir, la probabilidad de observar y en la población es el número de veces que y aparece en la población dividido entre el tamaño total de la población. En el caso especial (pero común) donde Y es una variable continua, todos sus valores son distintos, y la función de probabilidad es simplemente un punto de referencia en los valores distintos de Y con la misma probabilidad de altura 1/N.

La función de distribución acumulada poblacional (cdf, por sus siglas en inglés), $f_P(y) = P\ (Y \leq y)$, la suma de P_i sobre todos los valores i para los cuales $Y_i \leq y$, es una función escalonada que aumenta cada vez que aparece un valor distinto en la población.

2.2.14.2.Muestra

La verdadera distribución de probabilidad poblacional $f_P(y)$ suele ser desconocida. La muestra es un subconjunto de la población y se representa mediante un vector **y** de

longitud n. Sea d el número de valores distintos de Y en la muestra y, y n_i el número de veces que el valor distinto y_i aparece en ella. Entonces, el tamaño total de la muestra es

$$n = \sum_{i=1}^{d} n_i$$

Si

$$p_i = \frac{n_i}{n}$$

la distribución muestral queda definida mediante una función de probabilidad empírica (epf) asociada a la muestra **y** dada por

$$f_E(y) = \hat{P}(\mathrm{Y} = y) = \begin{cases} p_i & si\ y = y_i,\ \mathrm{I} = 1, 2, \ldots, d \\ 0 & en\ caso\ contrario \end{cases}$$

La función de distribución acumulativa empírica (ecf),

$$F_E\ (y)\ = \widehat{P}\ (\mathrm{Y} \leq\ y),$$

se obtiene sumando p_i para todos los índices i que cumplen $y_i \leq y$. Esta es una función escalonada que aumenta cada vez que aparece un valor distinto en la muestra (Stasinopoulos et al., 2017).

2.2.14.3. Modelo

La inferencia estadística es el proceso mediante el cual se utiliza una muestra para inferir la distribución de la población. Un modelo estadístico implica una serie de supuestos acerca de cómo se genera la muestra a partir de la población. Estos supuestos pueden ser paramétricos o no paramétricos, lo que da lugar a inferencias estadísticas paramétricas o no paramétricas, respectivamente.

2.2.14.4. Modelos paramétricos

La inferencia estadística paramétrica clásica supone que la verdadera distribución poblacional $f_P(y)$ pertenece a una familia particular de funciones de probabilidad o

densidad, representada por $f_Y(y/\theta)$, para un conjunto de valores posibles de θ, y que $f_P(y) = f_Y(y/\theta)$ para todo y, dado un valor particular de θ, donde θ es un vector de parámetros desconocidos.

La inferencia clásica ha puesto mayor énfasis en realizar inferencias sobre θ, suponiendo una función teórica particular $f_Y(y/\theta)$, que en seleccionar, en primer lugar, una familia de distribuciones apropiada para representar $f_P(y)$.

Si $f_Y(y/\theta)$ no constituye una buena aproximación de la verdadera distribución poblacional para ningún valor de θ, esta parte de la inferencia estadística clásica puede resultar irrelevante para el problema particular. La dependencia de las conclusiones inferenciales respecto de la especificación de la familia de distribuciones $f_Y(y/\theta)$, para un conjunto de valores posibles de θ, ha llevado a los estadísticos a desarrollar métodos de inferencia no paramétricos.

2.2.14.5. Modelos no paramétricos

Los supuestos estadísticos no paramétricos no implican parámetros desconocidos. Ejemplos de modelos donde se invocan supuestos no paramétricos son las denominadas pruebas no paramétricas. En estas pruebas, se supone que la muestra se genera a partir de una población desconocida con una distribución independiente e idéntica (iid), pero la distribución de la población no se define explícitamente. Los estimadores de densidad forman parte de las técnicas estadísticas denominadas técnicas de suavizado no paramétrico. En sentido estricto, en estas técnicas se determinan parámetros a partir de los datos, denominados parámetros de suavizado.

Las funciones empíricas de probabilidad y de distribución acumulativa son de suma importancia en la inferencia estadística no paramétrica. Esto se debe a que, para una muestra aleatoria de tamaño n procedente de la verdadera función de probabilidad o densidad poblacional (pdf) $f_P(y)$, la función de distribución acumulativa empírica ecf $F_E\ (y)$ es un

estimador, cuando $n \to \infty$, de la verdadera función de distribución acumulativa poblacional $F_P(y)$. Esto supone un muestreo aleatorio de una población de tamaño infinito, $N = \infty$, o un muestreo aleatorio con reemplazo de una población finita de tamaño N. La ecf también proporciona, mediante el principio de sustitución (Efron y Tibshirani, 1993, p. 35), un método sencillo de estimación no paramétrica.

2.2.15. Modelos lineales generalizados (GLM)

Los modelos lineales generalizados (Nelder y Wedderburn, 1972) surgen como una extensión del modelo de regresión lineal clásico, liberándolo de los supuestos de normalidad y homocedasticidad que este último exige para la variable respuesta. A diferencia del modelo lineal tradicional, que asume errores normales con varianza constante, los GLM admiten que la respuesta siga cualquier distribución de la familia exponencial (gaussiana, binomial, Poisson, gamma o inversa gaussiana, entre otras), lo que amplía considerablemente su utilidad frente a datos reales de naturaleza discreta, asimétrica o acotada.

Esta flexibilidad se articula a través de tres componentes que estructuran el modelo. El componente aleatorio especifica la distribución de probabilidad de la variable respuesta, definida por su media y por la relación funcional que esta guarda con la varianza. El componente sistemático corresponde al predictor lineal, es decir, la combinación lineal de las variables explicativas ponderadas por sus coeficientes. El componente de enlace, por su parte, aporta una función monótona que conecta la media esperada de la respuesta con dicho predictor lineal, lo que permite representar relaciones no lineales entre la respuesta y los regresores sin abandonar la linealidad en los parámetros del modelo.

Matemáticamente, el modelo se formula como $g(\mu) = X^T\beta$, donde $\mu = E(Y)$, $g(\cdot)$ es la función de enlace y β es el vector de parámetros desconocidos. Esta estructura hace posible interpretar los coeficientes en términos del efecto de los predictores sobre una transformación de la media, manteniendo una interpretación estadística clara y consistente

con la distribución asumida. Casos particulares de esta formulación incluyen el modelo lineal clásico, cuando se emplea la distribución gaussiana con enlace identidad, y la regresión logística o de Poisson, cuando se utilizan distribuciones binomial o Poisson con enlaces logit y log, respectivamente.

Desde una perspectiva inferencial, la estimación de los parámetros en los GLM se lleva a cabo mediante máxima verosimilitud. Únicamente en el caso gaussiano con función de enlace identidad, esta estimación admite una solución analítica, equivalente a la minimización de la suma de cuadrados de los residuos. En los demás casos, la función de verosimilitud carece de solución cerrada, lo que exige el uso de algoritmos iterativos de optimización.

2.2.16. Modelos aditivos generalizados (GAM)

A pesar de la flexibilidad que ofrecen los modelos lineales generalizados (GLM), estos mantienen como restricción fundamental que la relación entre los predictores y la media de la variable respuesta sea lineal y constante en la escala del predictor lineal. Esta limitación resulta restrictiva en muchos problemas reales, donde los efectos de las covariables sobre la respuesta presentan comportamientos no lineales. Con el objetivo de superar esta restricción, Hastie y Tibshirani (1986) introdujeron los modelos aditivos generalizados (GAM), los cuales extienden el marco de los GLM permitiendo incorporar relaciones no lineales de forma flexible y controlada.

En los modelos GAM, la media de la variable respuesta, a través de la función de enlace g(μ), se modela como la suma de funciones suaves de las covariables, en lugar de una combinación lineal estricta. De este modo, el efecto de cada predictor x_i sobre la respuesta no se introduce directamente, sino mediante una función $f_i(x_i)$, que es estimada a partir de los datos. Esta formulación permite capturar patrones complejos sin imponer una forma

funcional paramétrica previa, manteniendo al mismo tiempo interpretabilidad y coherencia probabilística. La expresión general de un GAM es:

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$$

Las funciones $f_i(x_i)$ pueden adoptar distintas formas, lineales o no lineales, aunque en la práctica se emplean principalmente funciones suaves (smooths). Entre las más utilizadas se encuentran los cubic regression splines, los thin plate regression splines y los penalized splines, cuya complejidad se controla mediante penalizaciones que evitan el sobreajuste y permiten un equilibrio entre ajuste y suavidad.

2.2.17. Redes neuronales

Las redes neuronales artificiales (NN, por sus siglas en inglés) constituyen una clase fundamental de algoritmos de aprendizaje automático cuya inspiración deriva del funcionamiento del sistema nervioso biológico, particularmente de la forma en que las neuronas humanas procesan y transmiten información (Haykin, 2009). Desde un punto de vista computacional, una red neuronal se conforma por un conjunto de nodos interconectados que operan en capas, donde cada neurona aplica una transformación a sus entradas ponderadas y transmite el resultado hacia las neuronas de la siguiente capa. Este proceso posibilita que la red capture relaciones complejas, incluso altamente no lineales, presentes en los datos (Goodfellow et al., 2016).

Históricamente, los primeros modelos neuronales se remontan al perceptrón de Rosenblatt (1958), diseñado originalmente para clasificar patrones linealmente separables. Posteriormente, investigaciones posteriores demostraron que redes con múltiples capas podían aproximar cualquier función continua bajo ciertas condiciones, lo que se conoce como el teorema de aproximación universal (Cybenko, 1989; Hornik, 1991). Este hallazgo sentó las bases teóricas para el desarrollo de arquitecturas más profundas y complejas.

Desde la perspectiva matemática, el funcionamiento de una NN se basa en ajustar los pesos y sesgos que controlan la influencia de cada neurona. La salida de una neurona se obtiene mediante una función de activación (como la sigmoide, ReLU o tanh) que introduce no linealidad y permite que la red capture patrones intrincados (Nielsen, 2015). El entrenamiento implica minimizar una función de pérdida mediante algoritmos de optimización basados en gradientes. La técnica de retropropagación, presentada por Rumelhart et al. (1986), revolucionó el campo al permitir calcular eficientemente los gradientes de redes profundas, facilitando su entrenamiento masivo.

Las redes neuronales pueden clasificarse en diversos tipos según su arquitectura y propósito. Las redes feedforward constituyen la estructura más básica, donde la información fluye en una única dirección sin ciclos (Haykin, 2009). Por otro lado, las redes neuronales convolucionales (CNN) han demostrado un desempeño sobresaliente en el procesamiento de imágenes y visión por computadora, debido a su capacidad para extraer características espaciales jerárquicas (Lecun et al., 2015). Las redes neuronales recurrentes (RNN) y sus variantes modernas como LSTM o GRU se han aplicado ampliamente en tareas que requieren modelar secuencias temporales, tales como reconocimiento de voz y análisis de series temporales (Goodfellow et al., 2016).

En el contexto del aprendizaje profundo (deep learning), las NN han ganado terreno gracias a los avances en potencia computacional, disponibilidad de grandes volúmenes de datos y nuevos métodos de regularización y optimización, como dropout, normalización por lotes (batch normalization) y el optimizador Adam (Kingma y Ba, 2015). Estos avances han permitido entrenar modelos compuestos por millones o incluso miles de millones de parámetros, con resultados cercanos o incluso superiores al rendimiento humano en tareas específicas.

Las ventajas de las redes neuronales residen en su capacidad para aprender directamente de los datos sin necesidad de ingeniería manual de características, su adaptabilidad y su elevado rendimiento en problemas complejos. Sin embargo, también presentan limitaciones importantes: requieren grandes cantidades de datos etiquetados, sus modelos suelen ser opacos y difíciles de interpretar, y la selección de hiperparámetros puede ser costosa computacionalmente (Zhang et al., 2017). Pese a ello, su utilidad en múltiples dominios ha impulsado su integración en sistemas de apoyo a la toma de decisiones, análisis predictivo y automatización inteligente.

Actualmente, las NN se emplean en áreas tan diversas como la medicina, para el diagnóstico asistido por imágenes; las finanzas, para la detección de anomalías y predicción de mercados; la ingeniería, para el control adaptativo; y las ciencias ambientales, para la modelación de sistemas climáticos complejos (Lecun et al., 2015; Goodfellow et al., 2016). Como menciona Ripley (2008), las redes neuronales proporcionan un marco robusto capaz de capturar patrones ocultos en conjuntos de datos de alta dimensionalidad, lo que las convierte en una herramienta indispensable dentro del aprendizaje automático moderno.

2.2.18. Bosques aleatorios

Los bosques aleatorios (RF, por sus siglas en inglés), propuestos por Breiman (2001), constituyen un método de aprendizaje ensemble que funciona generando numerosos árboles de decisión y combinando sus resultados. En problemas de clasificación, la predicción final corresponde a la clase más frecuente entre los árboles, mientras que en regresión se utiliza el promedio de sus salidas. La fortaleza principal de este enfoque radica en la combinación de múltiples predictores: un conjunto de árboles poco correlacionados tiende a producir un error de predicción menor que cualquier árbol individual.

Para construir estos bosques, Breiman incorporó dos fuentes fundamentales de aleatoriedad. La primera es el muestreo con reemplazo (bagging), mediante el cual cada

árbol se entrena con una muestra aleatoria (y repetida) del conjunto de entrenamiento, lo que contribuye a disminuir la varianza. La segunda es la selección aleatoria de variables, donde en cada división de un nodo el algoritmo evalúa únicamente un subconjunto aleatorio de atributos en lugar de todas las características disponibles.

Los bosques aleatorios se destacan por su resistencia al ruido y su capacidad para abordar tanto tareas de clasificación como de regresión. Entre sus ventajas se encuentra la reducción del sobreajuste: a diferencia de un árbol único, que puede ajustarse excesivamente a los datos, la combinación de múltiples árboles independientes mitiga este problema. Además, el algoritmo permite estimar el error de generalización mediante el método out-of-bag sin requerir un conjunto de validación adicional. Finalmente, es capaz de calcular la importancia de las variables, facilitando la identificación de los atributos más determinantes en las predicciones (Breiman, 2001).

2.2.19. Aproximación de Laplace integrada anidada

La aproximación de Laplace integrada anidada (Integrated Nested Laplace Approximation, INLA) es un método de inferencia bayesiana aproximada desarrollado para modelos latentes gaussianos. Su objetivo es obtener aproximaciones rápidas y precisas de las distribuciones posteriores marginales de los parámetros y efectos latentes, evitando el uso intensivo de métodos de simulación como MCMC. INLA fue propuesto por Rue, Martino y Chopin (2009) como una alternativa computacional eficiente para modelos bayesianos jerárquicos con estructura latente gaussiana.

En este marco, se asume que las observaciones y_i son condicionalmente independientes dado un campo latente gaussiano x y un conjunto de hiperparámetros θ. De forma general, el modelo puede expresarse como:

$$\pi(x, \theta|y) \propto \pi(\theta)\pi(x|\theta) \prod_{i=1}^{n} \pi(y_i|x_i, \theta)$$

donde $\pi(\theta)$ representa la distribución previa de los hiperparámetros, $\pi(x \mid \theta)$ corresponde a la distribución del campo latente gaussiano y $\pi(\ _{yi} \mid x_i, \theta)$ representa la contribución de cada observación a la verosimilitud.

El campo latente x se modela como una distribución normal multivariada con matriz de precisión $Q(\theta)$:

$$x \mid \theta \sim N\left(0, Q^{-1}(\theta)\right)$$

La eficiencia computacional de INLA se basa en el uso de matrices de precisión dispersas, lo que permite trabajar con modelos complejos, incluidos modelos espaciales, temporales y espacio-temporales, sin depender directamente de cadenas MCMC. En R, esta metodología se implementa mediante el paquete INLA, cuyas funcionalidades han sido ampliadas para diferentes clases de modelos bayesianos aplicados.

El objetivo principal de INLA es aproximar las distribuciones posteriores marginales de los componentes del campo latente y de los hiperparámetros:

$$\pi(\mathrm{x_j} \mid y)$$

$$\pi(\theta_k \mid y)$$

Para ello, INLA utiliza una estrategia anidada: primero aproxima la distribución posterior de los hiperparámetros y luego, condicionada a estos, aproxima las distribuciones marginales de los efectos latentes. Este procedimiento permite obtener inferencia bayesiana aproximada de manera rápida, especialmente en modelos donde el número de efectos latentes es elevado.

2.2.20. Modelo vectorial autorregresivo espacio-temporal (VAST)

El modelo vectorial autorregresivo espacio-temporal, conocido como VAST (Vector Autoregressive Spatio-Temporal model), es un enfoque estadístico desarrollado para analizar datos pesqueros con estructura espacial y temporal. Este marco permite estimar la densidad poblacional, la biomasa, la abundancia relativa o la distribución de organismos

marinos a partir de datos observados en diferentes ubicaciones y periodos, integrando información de encuentro/no encuentro, conteos o biomasa (Thorson & Barnett, 2017; Thorson, 2019). VAST ha sido propuesto como una herramienta común para evaluaciones de stock, hábitat, ecosistemas y cambio climático, porque permite representar la variación de la densidad en el espacio, el tiempo y, en algunos casos, entre múltiples especies o categorías biológicas.

Desde el punto de vista estadístico, VAST se basa en modelos mixtos generalizados espacio-temporales. Su estructura combina una distribución observacional para la variable respuesta con efectos fijos, covariables, campos espaciales latentes y componentes espacio-temporales. En términos generales, el predictor del modelo puede expresarse como:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\,\beta_j + \Omega(s_i) + \varepsilon(s_i, t_i)$$

donde μ_i representa el valor esperado condicional de la respuesta, $g(\cdot)$ es la función de enlace, x_{ij} son covariables asociadas a la observación i, $\Omega(s_i)$ representa el campo espacial latente en la ubicación s_i, y $\varepsilon(s_i, t_i)$ corresponde al efecto espacio-temporal asociado a la ubicación s_i y al tiempo t_i. Esta formulación permite modelar dependencia espacial y variación temporal no explicada únicamente por las covariables observadas (Thorson, 2019).

En aplicaciones pesqueras, VAST se usa frecuentemente bajo una estructura delta, separando el proceso de encuentro del componente positivo de la respuesta. Esta formulación es útil cuando la variable contiene muchos ceros, como ocurre con datos de captura, biomasa, CPUE o densidad, donde una parte del modelo describe la probabilidad de presencia y otra parte describe la magnitud positiva observada (Thorson & Barnett, 2017; Thorson, 2019). En forma general, esta estructura puede representarse como:

$$P(Y_i = 0) = 1 - p_i$$

$$Y_i \mid Y_i > 0 \sim F(\mu_i, \theta), P(Y_i > 0) = p_i$$

donde p_i representa la probabilidad de encuentro y $F(\mu_i, \theta)$ es una distribución continua o discreta para el componente positivo de la respuesta. De este modo, VAST permite separar la ocurrencia espacial del recurso y la intensidad positiva del proceso observado.

El campo espacial latente de VAST permite representar correlación espacial entre observaciones cercanas, mientras que el componente espacio-temporal permite describir cambios temporales en la distribución o densidad. Esta estructura puede mejorar la estimación de índices de abundancia o biomasa cuando existe dependencia espacial, muestreo irregular o variación temporal en la distribución del recurso. Por ello, VAST se ha utilizado ampliamente en evaluaciones pesqueras modernas para estandarizar índices, estimar distribución espacial, evaluar desplazamientos de población y construir indicadores ecológicos espacio-temporales (Thorson, 2019; Grüss & Thorson, 2019).

Sin embargo, la incorporación de campos espaciales o espacio-temporales no elimina la necesidad de especificar adecuadamente la distribución observacional de la respuesta. VAST permite introducir estructura espacial y temporal, pero la validez del modelo sigue dependiendo de que la familia probabilística elegida represente correctamente la naturaleza de los datos analizados. Por ello, en variables pesqueras con fuerte asimetría, heterogeneidad, sobredispersión, exceso de ceros o colas pesadas, la selección de la distribución observacional y la evaluación residual siguen siendo elementos centrales para determinar la adecuación del modelo.

2.2.21. Modelos aditivos generalizados para locación escala y forma

Los modelos aditivos generalizados para localización, escala y forma (GAMLSS) constituyen un tipo de modelo de regresión distribucional propuesto por Rigby y Stasinopoulos (2005) como alternativa frente a las limitaciones que presentan los modelos

estadísticos clásicos (regresión lineal simple, regresión lineal múltiple, modelos lineales generalizados y modelos aditivos generalizados) ante la heterogeneidad, asimetría y curtosis de la variable respuesta. Estos modelos resultan pertinentes cuando el interés no se limita a los cambios en la media, sino que se extiende a otras características de la distribución, como la varianza (volatilidad), la asimetría y la curtosis (colas pesadas), así como a los cuantiles (Stasinopoulos y Rigby, 2007). Su principal aporte radica en permitir que todos los parámetros de la distribución de la respuesta (no solo la media) se modelen en función de las variables explicativas: μ (parámetro de localización, aproximadamente la media, la mediana o la moda), σ (parámetro de escala, aproximadamente el coeficiente de variación), ν (parámetro de forma, asociado a la asimetría) y τ (parámetro de forma, asociado a la curtosis) (Stasinopoulos et al., 2017; Stasinopoulos et al., 2024).

2.2.21.1. GAMLSS como modelo de regresión distribucional

En un modelo de regresión distribucional, la relación entre la respuesta y y las covariables x es de naturaleza estocástica. La respuesta y depende de x a través de la distribución condicional $f(y \mid x)$, que es el principal objeto de interés, ya que proporciona información detallada sobre cómo las covariables x afectan diversos aspectos de la distribución condicional de y (Stasinopoulos et al., 2024).

GAMLSS proporciona un marco paramétrico para la inferencia estadística en regresión distribucional, en el que la distribución condicional $f(y \mid x)$ se aproxima mediante una distribución paramétrica $f(y \mid \theta(x))$, donde

$$\theta(x) = \left(\theta_1(x), \theta_2(x), \dots, \theta_K(x)\right)^T$$

es un vector K-dimensional de parámetros desconocidos del modelo que, a su vez, dependen de términos explicativos. La idea básica de la inferencia estadística es utilizar $f(y \mid \theta(x))$ para extraer conclusiones razonables sobre la distribución poblacional $f(y \mid x)$ (Stasinopoulos et al., 2024).

La notación $\theta(x)$ enfatiza que cualquiera de los parámetros del modelo contenidos en θ puede ser función de cualquiera de las variables explicativas x, y no únicamente la media. Este enfoque implica que la forma de la distribución del modelo para y puede variar en función de los valores de las variables explicativas x. Al modelar todos los parámetros de $f(y \mid \theta(x))$ como funciones de los términos explicativos, se modelan explícita y simultáneamente todas las características de la distribución, incluida la ubicación, la escala (variabilidad), los cuantiles, los momentos, la asimetría y la curtosis (Stasinopoulos et al., 2024).

Modelar solo la media permite cambios exclusivamente en la ubicación de la distribución, manteniendo todos los demás parámetros de la distribución constantes. En consecuencia, modelar todos los parámetros permite varios tipos de cambios en la forma de la distribución de una variable de respuesta, en función de una o más variables explicativas (Stasinopoulos et al., 2024).

2.2.21.2.Función de verosimilitud

El concepto de función de verosimilitud es fundamental en la inferencia estadística paramétrica. Se basa en el principio de que los valores de los parámetros que hacen que los datos observados sean relativamente más probables bajo el modelo asumido deben considerarse más plausibles que aquellos que hacen que los datos resulten menos probables.

La función de verosimilitud, $\mathcal{L}(\theta)$, se define como la probabilidad de observar la muestra, considerada no como una función de los datos, sino como función del parámetro θ.

Sea $y = (y_1, y_2, \dots y_n)$ una muestra aleatoria observada, proveniente de una distribución poblacional paramétrica con función de densidad (o función de probabilidad, en el caso discreto) $f_Y(y_i|\theta)$, cuya forma funcional es conocida excepto por el parámetro desconocido θ. Se asume que las observaciones son independientes e idénticamente

distribuidas, lo cual es consistente con un muestreo aleatorio de una población infinita o con reemplazo desde una población finita.

Bajo estas condiciones, la función de verosimilitud se expresa como:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta)$$

donde $f_Y(y_i|\theta)$ representa la función de densidad de la variable aleatoria Y, evaluada en la observación $y_i y\theta$ denota el conjunto de parámetros del modelo.

Es decir, la probabilidad conjunta de la muestra es el producto de las probabilidades individuales. Obsérvese que, una vez observados los datos y, la verosimilitud no es una función de los datos (que se consideran fijos), sino del parámetro θ.

El método clásico para ajustar un modelo paramétrico a una muestra observada es el método de máxima verosimilitud, que consiste en encontrar el valor de θ que maximiza la función de verosimilitud. En la práctica, resulta más conveniente trabajar con el logaritmo de la verosimilitud, definido como:

$$\ell(\theta) = \sum_{i=1}^{n} \log f_Y\,(\,y_i \mid \theta\,)$$

Dado que transforma el producto en una suma y facilita tanto la optimización numérica como el análisis teórico.

2.2.21.3. Función de verosimilitud de una variable continua

Si la distribución de *Y* es continua, en la práctica se observa un valor específico $_{yi}$ con cierto nivel de precisión, digamos $y_i \pm \Delta_i$. Por ejemplo, si y_i se redondea al primer decimal más cercano, entonces Δ_i= 0.05 y, por ejemplo, un valor observado y_i = 5.7 corresponde a 5.65 < y < 5.75. Por lo tanto, la verosimilitud real (es decir, la probabilidad real de observar los datos y) puede definirse como:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[f_Y\left(y_i + \frac{\Delta_i}{\theta}\right) - f_Y(y_i - \Delta_i \mid \theta) \right]$$

donde $F_Y(\cdot \mid \theta)$ es la función de distribución acumulada asociada a $f_Y(\cdot \mid \theta)$

2.2.21.4. Estimación de máxima verosimilitud

La estimación de máxima verosimilitud (MLe) es el valor de θ, que maximiza la función de verosimilitud $\mathcal{L}(\theta)$ o, equivalentemente, maximiza la función de log-verosimilitud $\ell(\theta)$. La estimación de máxima verosimilitud se denomina estimación puntual en la inferencia estadística clásica. Las estimaciones puntuales responden al problema inferencial de adivinar el valor real θ_T de los parámetros θ. Los estimadores puntuales son bastante ingenuos, ya que incluso si creemos que existe un parámetro verdadero θ_T, nuestra suposición sería inevitablemente errónea. La MLe $\hat{\theta}$ de θ_T es una función de la muestra observada y, mientras que el estimador de máxima verosimilitud (MLE) de θ_T es la misma función de la muestra aleatoria (de variables aleatorias independientes e idénticamente distribuidas) Y. Por lo tanto, el MLe de θ_T es un vector de valores numéricos constantes calculados a partir de la y observada, mientras que el MLE de θ_T es un vector de variables aleatorias que son funciones de Y.

2.2.21.5. Propiedades estadísticas del estimador de máxima verosimilitud (MLE) en un modelo correcto

Bajo condiciones de regularidad y suponiendo que el modelo está correctamente especificado, el estimador de máxima verosimilitud $\hat{\theta}$ presenta las siguientes propiedades fundamentales:

- Invariancia: Establece que, si $\emptyset$ es una transformación uno a uno del parámetro θ, digamos $\emptyset = g(\theta)$, entonces los estimadores de máxima verosimilitud $\widehat{\emptyset}$ y $\hat{\theta}$ están relacionados por

$$\widehat{\emptyset} = g(\hat{\theta}).$$

Esto implica que el valor de la verosimilitud en el punto de máximo es independiente de la parametrización utilizada, y que un cambio de parametrización no afecta las estimaciones puntuales. No obstante, los errores estándar no son invariantes, ya que dependen de la curvatura local de la función de verosimilitud. La aproximación cuadrática de la log-verosimilitud en el entorno del máximo permite la construcción de intervalos de confianza y pruebas de Wald para θ_T.

- Consistencia: El estimador de máxima verosimilitud $\hat{\theta}$ es, en ciertas condiciones, un estimador (débilmente) consistente del parámetro verdadero θ_T, es decir, $\hat{\theta}$ converge en probabilidad a θ_T cuando $n \to \infty$.

Esto significa que para todo $\varepsilon > 0$,

$$\lim_{n\to\infty} p\left(\left|\hat{\theta} - \theta_T\right| > \varepsilon\right) = 0$$

- Normalidad asintótica: Bajo condiciones de regularidad, el estimador de máxima verosimilitud presenta normalidad asintótica, de modo que:

$$\sqrt{n}\left(\hat{\theta} - \theta_T\right) \xrightarrow{d} \mathcal{N}\left(0, I^{-1}(\theta_T)\right)$$

donde $I(\theta_T)$ es la matriz de información de Fisher evaluada en el verdadero valor del parámetro. Esta propiedad justifica el uso de aproximaciones normales para la inferencia estadística basada en la MLE, tales como intervalos de confianza y pruebas de hipótesis.

En el marco de los modelos GAMLSS, la estimación de los parámetros de localización, escala y forma se realiza mediante máxima verosimilitud penalizada, lo que permite conservar las propiedades asintóticas de la MLE incorporando términos de suavizado que controlan la complejidad de los efectos no lineales.

2.2.21.6. Evaluación del ajuste de los modelos GAMLSS: PIT y z-score

GAMLSS utiliza dos principios fundamentales de diagnóstico distribucional: la Transformada Integral de Probabilidad (Probability Integral Transform, PIT) y las puntuaciones tipo z-score.

La PIT permite evaluar el ajuste del modelo independientemente de la escala original de los datos o de la forma de la distribución, y constituye un diagnóstico universal para modelos con parámetros de localización, escala, asimetría y curtosis (como GAMLSS).

Para evaluar la adecuación del modelo, se calcularon los residuos derivados de la Transformada Integral de Probabilidad (PIT) como:

$$u_i = F_Y\left(\mathrm{y}_i ; \widehat{\theta}_i\right)$$

donde $F(\cdot)$ es la función de distribución acumulada condicional ajustada (cdf).

El principio matemático subyacente es que, si el modelo reproduce correctamente la distribución condicional de la variable respuesta, estos valores transformados deberían comportarse como variables aleatorias uniformes en el intervalo (0,1):

$$U(0,1)$$

Sin embargo, los residuos expresados en escala uniforme no son intuitivos para la inspección gráfica. Por ello, se aplica la transformación normal estándar inversa:

$$z_i = \Phi^{-1}(u_i)$$

obteniendo residuos tipo z-score normalizados, los cuales deberían aproximarse a una distribución normal estándar con media 0 y varianza 1 si el modelo está bien especificado:

$$N(0,1)$$

Estos residuos permiten utilizar herramientas gráficas de diagnóstico (histogramas, Q–Q plots, worm plots) para detectar faltas de ajuste. En particular:

Una media distinta de 0 refleja sesgo o incorrecta modelación del parámetro de localización (μ).

Una varianza distinta de 1 sugiere mal modelado del parámetro de escala (σ) o sobredispersión.

Asimetrías sistemáticas revelan insuficiente modelación del parámetro de sesgo (ν).

Colas excesivas o deficientes indican errores en curtosis o parametrización de τ.

Además, cuando se analizan los z-scores frente a covariables o mediante diagramas de gusano, es posible identificar en qué rangos o valores de los predictores el modelo falla, facilitando la mejora estructural del modelo (inclusión de términos no lineales, efectos por temporada, funciones spline, o covariables adicionales).

Así, el uso conjunto de residuos PIT y z-scores constituye una estrategia robusta y ampliamente recomendada en modelación distribucional, permitiendo evaluar no solo el ajuste global del modelo, sino también la consistencia y relevancia de cada parámetro (μ, σ, ν, τ) a lo largo de los predictores. Esta metodología ha sido desarrollada y aplicada extensivamente en modelos distribucionales (Dunn y Smyth, 1996; Rigby y Stasinopoulos, 2005).

2.2.21.7. Selección de modelos

2.2.21.7.1. Desviación global

Es una medida o cantidad muy importante para realizar pruebas entre diferentes modelos, se expresa como:

$$GD = -2 \log \mathcal{L}(\hat{\theta}) = -2\, \ell\, (\hat{\theta})$$

2.2.21.7.2. Selección de modelos mediante el Criterio de Información Generalizado de Akaike (GAIC)

Para comparar modelos GAMLSS no anidados, se puede utilizar el Criterio de Información de Akaike generalizado (GAIC) para penalizar el ajuste excesivo. Esto se obtiene añadiendo a la desviación global ajustada una penalización fija k por cada grado de libertad efectivo utilizado en un modelo, es decir:

$$GAIC(k) = GD + (k\ x\ df),$$

donde *df* representa el total de grados de libertad efectivos utilizados en el modelo y GD es la desviación global ajustada. Se selecciona entonces el modelo con el menor valor del criterio GAIC(k). El criterio de información de Akaike (AIC) (Akaike, 1983) y el criterio bayesiano de Schwartz (SBC) (Schwarz, 1978) son casos especiales del criterio GAIC(k) correspondientes a k = 2 y k = log(n) respectivamente:

$$GAIC(k) = GD + (2\ x\ df),$$

$$GAIC(k) = GD + (log(n)\ x\ df),$$

Los dos criterios, AIC y SBC, se justifican asintóticamente como predictores del grado de ajuste en un nuevo conjunto de datos, es decir, como aproximaciones al error predictivo promedio. La justificación para el uso de SBC también se presenta como una aproximación burda a los factores de Bayes (Raftery, 1996; Raftery, 1999). En la práctica, se suele observar que, si bien el AIC original es muy generoso en la selección del modelo, el SBC es demasiado restrictivo. La experiencia indica que un valor de la penalización k en el rango $2.5 \leq k \leq 3$ funciona bien para la mayoría de los datos. El uso de GAIC(k) permite probar diferentes penalizaciones k para diferentes propósitos de modelado. También se puede investigar la sensibilidad del modelo seleccionado a la elección de k. Una selección de diferentes valores de k, por ejemplo, k = 2, 2.5, 3, 3.5, 4 podrían utilizarse a su vez para

investigar la sensibilidad o robustez de la selección del modelo a la elección del valor de la penalización k.

2.2.21.8. Distribuciones de respuesta

En un modelo GAMLSS, se asume que las respuestas se generan a partir de una familia K -paramétrica de distribuciones con parámetros (dependientes de la covariable) $\theta(x) = \left(\theta_1(x), \theta_2(x), \dots, \theta_K(x)\right)^T$.Posteriormente permitimos que esos K parámetros difieran posiblemente para cada observación i, para $i = 1, \dots., n$, por lo que introducimos la notación $\theta_i(x_i) = \left(\theta_{i1}(x_i), \theta_{i2}(x_i), \dots, \theta_{iK}(x_i)\right)^T$ para $i = 1, \dots., n$. Al suprimir la dependencia explícita de ik en xi para mantener la notación corta, tenemos $\theta_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{iK})^T$. Se supone que hay un tipo común de distribución que se aplica a todas las observaciones (por ejemplo, normal, Poisson, etc.) pero que se permite que los K parámetros de esta distribución varíen en las observaciones individuales, es decir:

$$\mathrm{y_i} \overset{ind}{\sim} \mathcal{D}(\theta_i 1, \dots, \theta_i K), \qquad \text{para } i = 1, \dots., n.$$

La densidad y la función de distribución acumulada de esta distribución se denotan como $f(y/\theta_i)$ y $F(y/\theta_i)$, respectivamente. Cabe destacar que cada θ_{ik} para $k = 1, \dots., K$ puede depender de diferentes subconjuntos de x_i

$\mathcal{D}$ indica la distribución de la respuesta. El marco GAMLSS permite una gran variedad de tipos de respuesta, entre ellos: (i) modelos para respuestas continuas, que nos permiten no solo tratarlas, sino también estudiarlas sistemáticamente, junto con fenómenos como la heterocedasticidad y la asimetría, (ii) modelos para respuestas continuas no negativas, que pueden presentar una masa puntual discreta en cero, (iii) respuestas de recuento, que pueden presentar inflación cero y/o sobredispersión, (iv) respuestas fraccionarias o limitadas continuas (por ejemplo, proporciones), que también pueden

presentar masas puntuales discretas en uno o ambos extremos, y (v) distribuciones de respuesta multivariantes (Stasinopoulos et al., 2024).

2.2.21.9. Funciones de enlace

En el marco de los modelos GAMLSS, para cada parámetro de la distribución θ_k se define una función de enlace monótona $g_k()$ y su correspondiente función de respuesta $h_k(\cdot) = g_k^{-1}(\cdot)$, que relacionan el predictor lineal con el parámetro de distribución. Formalmente, esta relación se expresa como:

$$\eta_k = g_k(\theta_k) \text{ y } \theta_k = h_k(\eta_k).$$

donde η_k representa el predictor de regresión asociado al parámetro θ_k.

La función de respuesta $h(\cdot)$ se selecciona de modo que transforme el predictor lineal, definido sobre la recta real, al soporte adecuado del parámetro de distribución correspondiente. Esto garantiza que los parámetros estén correctamente restringidos. Por ejemplo, parámetros como la desviación estándar o la varianza deben ser estrictamente positivos, mientras que los parámetros que representan probabilidades deben estar acotados al intervalo unitario [0, 1].

Esta es una característica importante y útil de la función de enlace, pero tenga en cuenta que las funciones de enlace reflejan la relación entre el parámetro y las covariables. Por ejemplo, en un modelo con un enlace de identidad para θ_k, la contribución de cada variable explicativa al parámetro de distribución θ_k es aditiva, mientras que para un modelo con un enlace logarítmico el efecto es multiplicativo (Stasinopoulos et al., 2024).

Más allá de esta función de restricción, la elección de la función de enlace determina la forma funcional de la relación entre las covariables y el parámetro de distribución. En particular, cuando se utiliza una función de enlace identidad, la contribución de cada covariable al parámetro θ_k es aditiva. En contraste, bajo una función de enlace logarítmica,

los efectos de las covariables actúan de manera multiplicativa sobre el parámetro (Stasinopoulos et al., 2024).

Si bien existen elecciones predeterminadas comúnmente empleadas (como el enlace logarítmico para parámetros positivos o los enlaces logit y probit para parámetros restringidos al intervalo unitario), es importante destacar que la selección de la función de enlace constituye una decisión explícita de modelado. Esta elección define la relación exacta entre las covariables y la distribución condicional de la variable respuesta, influyendo tanto en el ajuste del modelo como en la interpretación de los efectos de regresión estimados.En consecuencia, resulta metodológicamente apropiado considerar especificaciones alternativas de funciones de enlace y evaluar su desempeño durante el proceso de construcción, comparación y verificación del modelo, tal como se recomienda en el marco metodológico de GAMLSS (Stasinopoulos et al., 2024).

2.2.21.10. Notación de un modelo GAMLSS

El modelo GAMLSS se expresa en notación matricial como:

$$\mathrm{y} \overset{ind}{\sim} \mathcal{D}(\theta^1, \ldots, \theta_K)$$

$$g_k(\theta_k) = \eta_k$$

$$\eta_k = \beta_{0k} 1_n + B_{1k}\gamma_{1k} + \cdots + B_{J_k,k}\gamma_{J_k,k}$$

$$\gamma_{jk} \sim N\left(0, \tau_{jk}^2 K_{jk}^{-}\right)$$

donde $\mathcal{D}(\theta^1, \ldots, \theta_K)$ es una distribución K-paramétrica y los vectores θ_k y η_k tienen una longitud n, es decir, $\theta_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})^T$ para k = 1,…, K. Observe la diferencia entre el vector K-dimensional $\theta = (\theta_1, \theta_2, \ldots, \theta_K)^T$, que representa los parámetros de distribución en general para cualquier **x**, el vector K-dimensional $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})^T$, que representa los parámetros de distribución para la i-ésima observación, y el vector n-dimensional $\theta_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})^T$, que representa el k-ésimo parámetro de distribución para n observaciones (Stasinopoulos et al. 2024).

2.2.21.11. Distribuciones continuas

Las distribuciones continuas se pueden subdividir en tres categorías, según su soporte: la línea real $\mathbb{R} = (-\infty, \infty)$, la línea real positiva $\mathbb{R}+= (0, \infty)$, y el intervalo unitario $\mathbb{R}_{(0,1)} = (0, 1)$.

2.2.21.12. Distribuciones continuas en la línea real $\mathbb{R}(-\infty, +\infty)$

Distribución de un parámetro en $\mathbb{R}$

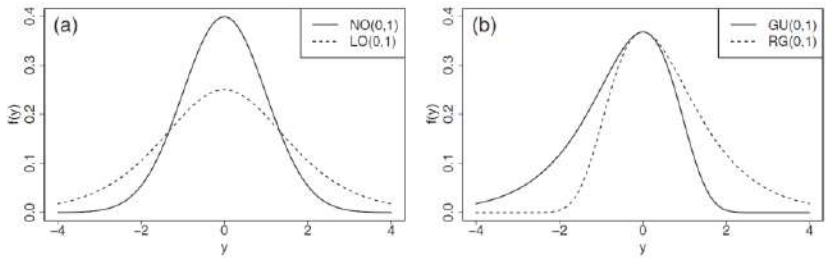
La distribución continua más conocida después de la normal es la distribución t (o t de Student). Fue desarrollada con éxito en el contexto del muestreo como la distribución subyacente a la prueba t (Gosset, 1908). En su forma estándar, es simétrica respecto a cero y tiene un único parámetro (grados de libertad, generalmente denotados como ν) que controla su dispersión (y curtosis). Dado que la distribución t no tiene un parámetro de ubicación, no es útil como distribución de respuesta. Sin embargo, al introducir parámetros de ubicación y escala en la t simple, la distribución resultante de la familia t de tres parámetros resulta útil en el contexto de la regresión (Stasinopoulos et al. 2024).

Distribuciones de dos parámetros en $\mathbb{R}$

Las distribuciones normal y logística son simétricas bien conocidas, siendo la logística leptocúrtica (más puntiaguda en el centro y con una cola más gruesa) en comparación con la normal. Las distribuciones Gumbel e inversa son sesgadas a la izquierda y a la derecha, respectivamente (Figura 1).

Figura 1

Distribuciones de dos parámetros en la línea real $\mathbb{R}$



Nota. (a) Distribuciones normal $NO(0,1)$ y logística $LO(0,1)$; (b) Distribuciones Gumbel $GU(0,1)$ e inversa Gumbel $RG(0,1)$. Tomada de Stasinopoulos et al. (2024).

Distribuciones de tres parámetros en $\mathbb{R}$

- Exponencial de potencia es simétrica respecto a su media μ, con varianza σ^2 y parámetro ν controlando la curtosis. Los casos especiales son las distribuciones de Laplace ($\nu = 1$) y normal ($\nu = 2$); y la distribución uniforme es el caso límite cuando $\nu \to \infty$ (Stasinopoulos et al. 2024).

- Familia t es una extensión de la distribución t de Student original con simetría respecto del parámetro de ubicación μ, y con el parámetro de escala σ y el parámetro ν de grados de libertad original (Stasinopoulos et al. 2024).

- Normales asimétrica de tipo 1 y 2 tienen como parámetro de ubicación μ (en el tipo 2, μ es la moda), parámetro de escala σ y parámetro de sesgo ν, donde $\nu \in$ R. Por lo tanto, es posible el sesgo a la izquierda. El caso límite, como $\nu \to \infty$ es la seminormal y, con $\to -\infty$ la seminormal reflejada (Stasinopoulos et al. 2024).

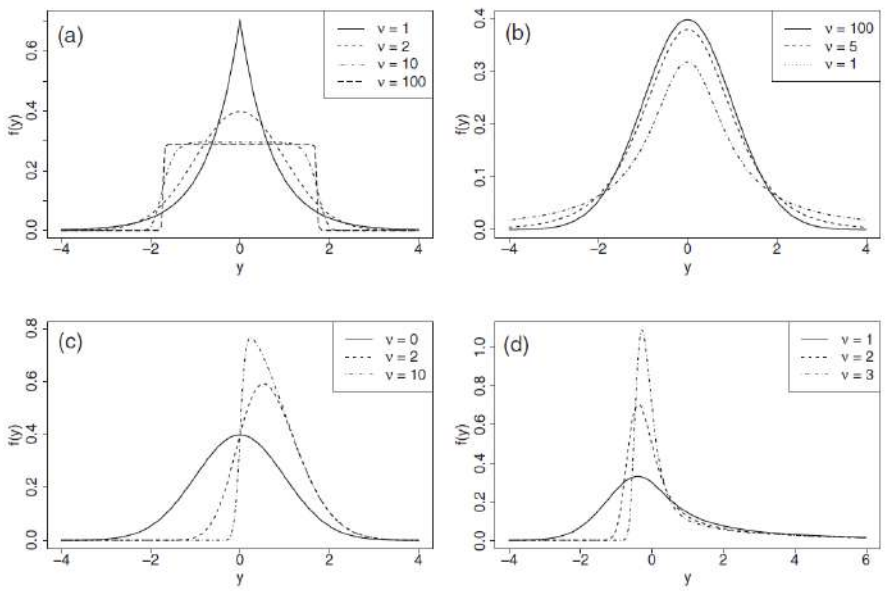
- Familia normal es una distribución normal, pero con su relación media-varianza controlada por ν. Es simétrica respecto a su media y mesocurtótica, con varianza proporcional a una potencia de la media: $\mathbb{V}(\mathrm{y}) = \sigma^2\mu^2$. Nótese que ν no está diseñada para ser modelada con covariables, sino que se utiliza para regular la relación media-varianza. Nótese también que debemos tener $\mu > 0$ (Stasinopoulos et al. 2024).

- Gaussiana exponencial (o gaussiana exponencialmente modificada) se deriva como la distribución de la suma de las variables aleatorias exponenciales y normales independientes, es decir, $y = y_1 + y_2$, donde $y_1 \sim Exp(\nu), y_2 \sim \mathcal{N}(\mu, \sigma^2)$, independientemente. Es sesgada positivamente, lo cual se hereda de su componente exponencial, y tiene media $\mu + \nu$ (Stasinopoulos et al. 2024).

En la Figura 2 se muestra las funciones de probabilidad de las distribuciones gaussianas exponenciales, de familia t, normales sesgadas y exponenciales.

Figura 2

Distribuciones de tres parámetros en la línea real $\mathbb{R}$



Nota. (a) exponencial de potencia $(PE(0, 1, \nu))$, (b) familia t $(TF(0, 1, \nu))$, (c) normal asimétrica tipo 1 $(SN1(0, 1, \nu))$, y (d) gaussiana exponencial $(exGAUS(0, 1, \nu))$. Tomada de Stasinopoulos et al. (2024).

Distribuciones de cuatro parámetros en $\mathbb{R}$

Se pueden clasificar en los siguientes grupos de distribuciones:

Distribuciones que modelan asimetría y curtosis:

- Potencia exponencial asimétrica (SEP1-SEP4).
- Sinh-arcsinh (SHASH, SHASHo, SHASHo2) (Figura 3).

Distribuciones que modelan asimetría y leptocurtosis:

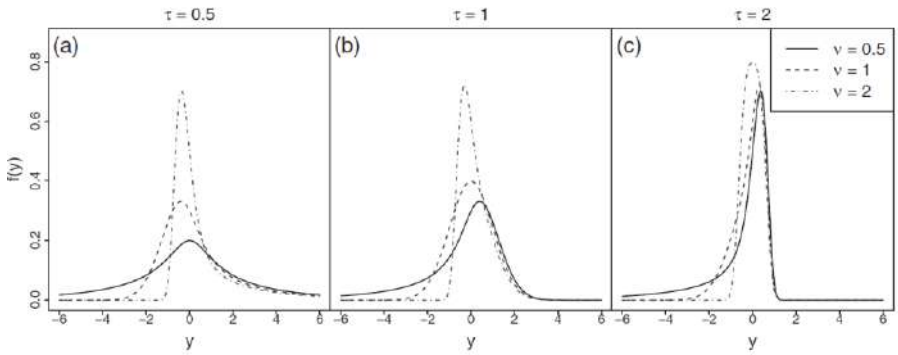
- T asimétricas (ST1-ST5, SST).
- SU de Johnson (JSU, JSUo).
- Beta generalizada exponencial tipo 2(EGB2).

La distribución t generalizada (GT) es simétrica. La curtosis se puede modelar, pero no la asimetría (Stasinopoulos et al. 2024).

La distribución t exponencial normal (NET) tiene cuatro parámetros, pero en gamlss se trata como una distribución de dos parámetros, con ν y τ fijos. Es simétrica respecto a μ (que es la media, si existe) y se introdujo para una estimación robusta de la ubicación y la escala (Rigby et al., 2019).

Figura 3

Distribución de cuatro parámetros en la línea real $\mathbb{R}$



Nota**.** sinh-arcsinh SHASH $(0, 1, \tau)$. Panel (a) $\tau = 0.5$, (b) $\tau = 1$ y (c) $\tau = 2$. Tomada de Stasinopoulos et al. (2024).

2.2.21.13. Distribuciones continuas en la línea real positiva $\mathbb{R}_+(0, +\infty)$

Distribución de un parámetro en $\mathbb{R}_+$

La distribución exponencial, un caso especial de la gamma y la Weibull, es una distribución de un parámetro (Stasinopoulos et al. 2024).

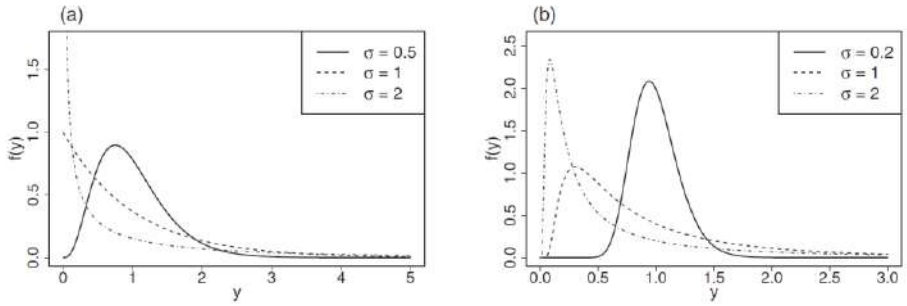
Distribución de dos parámetros en $\mathbb{R}_+$

La distribución gamma y la gaussiana inversa (Figura 4) pertenecen a la familia exponencial de distribuciones. La gaussiana inversa presenta una mayor asimetría positiva y una cola derecha más pesada que la gamma. Otras distribuciones de dos parámetros disponibles para modelar son:

- Logarítmica normal.
- Pareto, la conocida distribución de cola pesada para modelar eventos extremos.
- Weibull, utilizada en el análisis paramétrico de supervivencia.
- Gamma inversa, ampliamente utilizada como distribución a priori en el análisis bayesiano (Stasinopoulos et al. 2024).

Figura 4

Distribuciones de dos parámetros en la línea real positiva $\mathbb{R}_+$



Nota. (a) distribución gamma (GA), con $\mu = 1$ y $\sigma = 0.5$, 1 y 2; (b) distribución gaussiana inversa (IG), $\mu = 1$ y $\sigma = 0.5$, 1 y 2. Tomada de Stasinopoulos et al. (2024).

Distribución de tres parámetros en $\mathbb{R}_+$

- Gaussiana inversa generalizada (GIG) es una distribución flexible de tres parámetros, de la cual la gaussiana inversa es un caso especial (Jørgensen, 1982).

- Familia gamma es una versión de tres parámetros de la distribución gamma, en la que el tercer parámetro controla la relación media-varianza.

Otras distribuciones de tres parámetros en $\mathbb{R}_+$ en el sistema gamlss se derivan mediante transformación:

- Box-Cox, Cole y Green (BCCG) se define asumiendo que la variable aleatoria z tiene una distribución normal estandarizada truncada, donde z es una transformación de Box-Cox de y (Figura 5). La variable z aleatoria se define como:

$$z = \begin{cases} Y^{\nu} - 1 & si\ \nu \neq 0 \\ log(Y) & si\ \nu = 0 \end{cases}$$

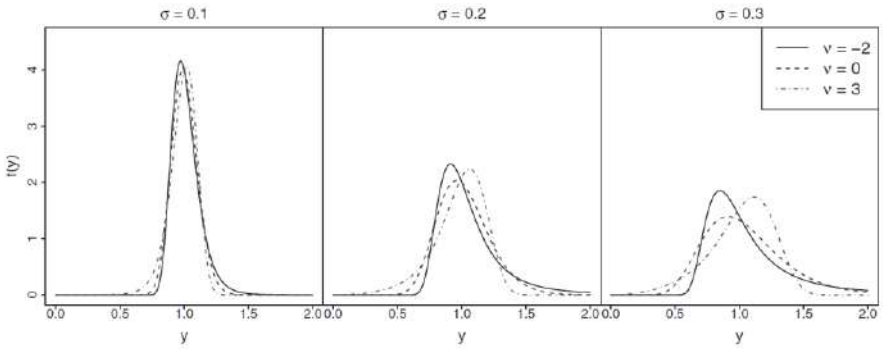
Donde ν es el parámetro de asimetría y se asume que la variable transformada Z tiene una distribución normal truncada $NO(\mu, \sigma)$. Cabe señalar que cuando $\nu = 0$ corresponde a la distribución logarítmica normal (LNO) y cuando $\nu = 1$ a la distribución normal (NO) (Stasinopoulos et al. 2017; Rigby et al., 2019).

- Gamma generalizada (GG) se deriva por transformación, asumiendo que $z = (y/\mu)^{\nu}$ tiene una distribución gamma con media 1.

- Familia logarítmica normal (LNO) se obtiene de forma similar a la distribución BCCG (Stasinopoulos et al. 2024).

Figura 5

Distribución de tres parámetros en la línea real positiva $\mathbb{R}_+$



Nota. Distribución de Box-Cox Cole y Green (BCCG), $\mu = 1$ y $\sigma = 0.1$, 0.2 y 0.3; y $\nu = -2$, 0 y 3. Tomada de Stasinopoulos et al. (2024).

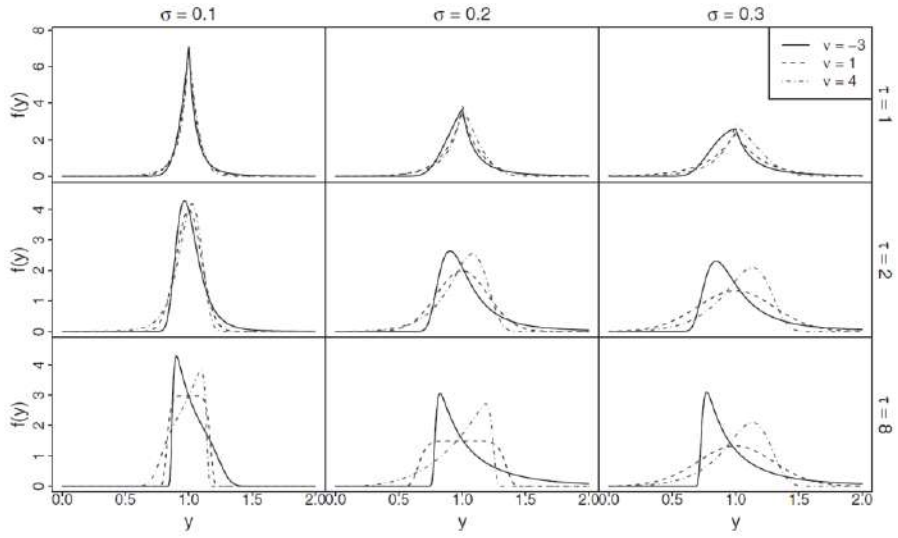
Distribución de cuatro parámetros en $\mathbb{R}_+$

- Potencia exponencial de Box-Cox (BCPE) se obtiene asumiendo que z tiene una distribución exponencial de potencia estándar truncada (Rigby y Stasinopoulos, 2004) (Figura 6).
- T de Box-Cox (BCT) se obtiene asumiendo que z (como se indicó anteriormente) tiene una distribución de familia t truncada (Rigby y Stasinopoulos, 2006).
- Beta generalizada tipo 2 (GB2) es una distribución flexible que contiene varias distribuciones conocidas como casos especiales. Por ejemplo, si se establece $\sigma = 1$, se obtiene una forma específica de la distribución de Pearson tipo VI. La distribución de Burr XII (o Singh-Maddala) se da cuando $\nu = 1$. La distribución de Burr III (o Dagum) se da cuando $\tau = 1$. Para $\sigma = \tau = 1$, se obtiene la distribución de Pareto (Rigby et al. 2019).

Tanto la distribución BCT como la BCPE son flexibles, fiables y se ajustan bien a diversos datos. La distribución GB2 es flexible y puede modelar la asimetría, así como la platicurtosis y la leptocurtosis (Rigby et al. 2019).

Figura 6

Distribución de cuatro parámetros en la línea real positiva $\mathbb{R}_+$



Nota. Potencia exponencial de Box-Cox (BCPE), con $\mu = 1$, σ = 0.1, 0.2; 0,3, $\nu = -3$, 1, 4 y $\tau = 1$, 2, 8. Tomada de Rigby et al. (2019).

2.2.21.14. Términos del modelo aditivo

Genéricamente, en un modelo GAMLSS cada parámetro distribucional θ_k se relaciona con sus covariables mediante una función de enlace $g_k(\cdot)$, de modo que:

$$g_k(\theta_{ik}) = \eta_{ik}$$

Genéricamente, se asume que cada uno de los predictores en un modelo GAMLSS es de la forma aditiva:

$$\eta_{ik} = \beta_0^{\theta_k} + s_1^{\theta_k}(x_{i1}) + \cdots + s_j^{\theta_k}(x_{ij}) + \cdots + s_{J_k}^{\theta_k}(x_{iJ_k})$$

es decir, compuesto por un término de intersección $\beta_0^{\boldsymbol{\theta}_k}$ y J_k funciones potencialmente no lineales $\mathrm{s}_\mathrm{j}^{\theta_k} + (x_{iJ_k})$, donde x_{ij} denota el vector de covariables para el j-ésimo efecto en el parámetro de distribución θ_k. Esta formulación permite que cada parámetro de la distribución, como ubicación, escala, asimetría o curtosis, tenga su propia estructura aditiva en función de las covariables consideradas (Stasinopoulos et al., 2024).

2.2.21.15. Splines penalizados

En el marco de los Modelos Aditivos Generalizados para Localización, Escala y Forma (GAMLSS), los efectos no lineales de las covariables continuas se modelaron mediante splines penalizados utilizando la función pb(). Este tipo de suavizador implementa P-splines (Eilers y Marx, 1996; Eilers y Marx, 2010; Eilers y Marx, 2021), los cuales combinan una base de B-splines con una penalización sobre los coeficientes del modelo, permitiendo capturar relaciones no lineales de manera flexible y estable.

La formulación de los P-splines se basa en expresar el efecto suave de una covariable continua x como una combinación lineal de funciones base B-spline, representada por una matriz de diseño B, y un vector de coeficientes γ. En este contexto, el ajuste del suavizador pb() puede interpretarse como la solución del siguiente problema de mínimos cuadrados penalizados:

$$S = (y - B\gamma)^T W (y - B\gamma) + \lambda \gamma^T D^T D \gamma$$

donde W es una matriz diagonal de pesos, D es una matriz que define diferencias finitas de segundo orden entre los coeficientes, y λ es el parámetro de suavizado que controla el compromiso entre ajuste y suavidad de la función estimada. La solución al problema de mínimos cuadrados penalizados viene dada por:

$$\hat{\gamma} = (B^T W B + \lambda D^T D)^{-1} B^T W y$$

y los valores ajustados de la función suave se obtienen como:

$$\hat{\mu} = B\hat{\gamma}$$

En GAMLSS, los suavizadores pb() pueden incorporarse de forma independiente en cada uno de los parámetros de la distribución (por ejemplo, μ, σ, ν y τ), lo que permite modelar explícitamente cómo una misma covariable influye de manera diferenciada sobre la media, la dispersión, la asimetría y la probabilidad de ausencia de la densidad acústica.

Esta característica constituye una de las principales ventajas del enfoque GAMLSS frente a modelos aditivos tradicionales centrados únicamente en la media.

En la práctica, el suavizador pb() utiliza por defecto una base de B-splines cúbicos y una penalización basada en diferencias de segundo orden, lo que garantiza la continuidad de la primera y segunda derivada de la función suave estimada. El número de nodos de la base se fija a un valor suficientemente grande (usualmente entre 20 y 50), de modo que la complejidad efectiva del suavizador esté controlada principalmente por el parámetro de suavizado λ, el cual es estimado automáticamente durante el proceso de ajuste del modelo (Stasinopoulos et al., 2023).

2.2.22. Modelos aditivos bayesianos para locación escala y forma

Los modelos aditivos bayesianos para localización, escala y forma (BAMLSS por sus siglas en inglés) introducidos por Umlauf et al. (2018), son una herramienta para el modelado de regresión bayesiana compleja con predictores aditivos; el enfoque totalmente bayesiano utiliza técnicas de simulación de Monte Carlo en cadena de Markov (MCMC) y proporciona intervalos creíbles válidos para los estimadores en situaciones en las que fallan los intervalos de confianza para los correspondientes estimadores de máxima verosimilitud basados en propiedades asintóticas. Este es específicamente el caso en modelos más complejos, por ejemplo, con distribuciones de respuesta fuera de la familia exponencial o cuando múltiples predictores contienen varios efectos suaves (Klein et al., 2015).

A diferencia de los modelos de regresión tradicionales (que suelen modelar solo la media o locación), los BAMLSS permiten modelar todos los parámetros de la distribución de la variable de respuesta: locación (μ), escala (σ) y forma (ν, τ) en función de variables explicativas. Asumen una distribución paramétrica para la variable de respuesta que puede provenir de una familia muy general, incluyendo distribuciones asimétricas y con curtosis

elevada, no limitándose a la familia exponencial (como la distribución normal, Poisson o binomial). Esto es útil cuando los datos no siguen una distribución estándar.

Cada parámetro de la distribución puede depender de las covariables a través de una suma de funciones flexibles (lineales, no lineales, suavizadas, espaciales o efectos aleatorios). Estas funciones suaves se suelen construir utilizando funciones base, como los splines penalizados (p-spline).

Los modelos BAMLSS utilizan métodos de inferencia bayesiana, típicamente técnicas de Cadenas de Markov Monte Carlo (MCMC), para estimar los parámetros del modelo. El enfoque bayesiano permite incorporar conocimiento previo (priors) y proporciona una medida completa de la incertidumbre en las estimaciones a través de las distribuciones posteriores.

2.2.22.1. Estructura del modelo BAMLSS

Basados en datos para $i = 1, \ldots, n$ observaciones, los modelos BAMLSS asumen la independencia condicional de las observaciones de respuesta individuales dadas las covariables. Al igual que en las clases de GAMLSS (Rigby y Stasinopoulos, 2005) o en los modelos de regresión distribucional (Klein, Kneib, Lang y Sohn, 2015c), todos los parámetros de la distribución de respuesta pueden modelarse mediante variables explicativas de tal manera que:

$$y \sim \mathcal{D}(h_1(\theta_1) = \eta_1, h_2(\theta_2) = \eta_2, \ldots, h_K(\theta_K) = \eta_K)$$

Donde $\mathcal{D}$ denota una distribución paramétrica para la variable de respuesta y con K parámetros θ_k, $k = 1, \ldots, K$, que están vinculados a predictores aditivos mediante funciones monótonas y dos veces diferenciables conocidas $h_k(\cdot)$. Nótese que la respuesta también puede ser un vector q-dimensional $y = (y_1, \ldots, y_q)$, por ejemplo, cuando $\mathcal{D}$ es una distribución multivariante (Klein, Kneib, Klasen y Lang 2015a). El k-ésimo predictor aditivo viene dado por:

$$\eta_k = \eta_k(x;\beta_k) = f_{1k}(x;\beta_{1k}) + \cdots + f_{J_k,k}\left(x;\beta_{J_k,k}\right)$$

Con funciones no especificadas (posiblemente no lineales) $f_{jk}(\cdot)$ de subvectores de un vector **x** que recopilan toda la información de covariables disponible, $j = 1, \ldots, J_k$ y $k = 1, \ldots, K$ y $\beta_k = \left(\beta_{1k}, \ldots, \beta_{J_k,k}\right)^T$ son parámetros, típicamente coeficientes de regresión que deben estimarse a partir de los datos.

2.2.22.2. Respuesta y distribución posterior del modelo BAMLSS.

El componente principal de los algoritmos de modelos de regresión es la función de densidad de probabilidad $\mathrm{d}y(y|\theta_1, \ldots, \theta_K)$ o, por motivos computacionales, su logaritmo. La estimación suele requerir evaluar la función de verosimilitud logarítmica (Umlauf et al., 2018):

$$\ell(\beta; \mathrm{y}, \mathrm{X}) = \sum_{i=1} \log \mathrm{d}y(\,_{\mathrm{y1}}; \theta_{\mathrm{i1}} = h_1^{-1}\left(\eta_{i1}(x_i;\beta_1)\right), \ldots, \theta_{\mathrm{iK}} = h_K^{-1}\left(\eta_{ik}(x_i;\beta_K)\right))$$

varias veces, donde el vector $\beta = (\beta_1^\top, \ldots, \beta_K^\top)$ comprende todos los coeficientes/parámetros de regresión que deben estimarse, $X = (X_1, \ldots, X_K)$ son las matrices covariables respectivas cuya i-ésima fila se denota $\mathrm{x_i}$ y θ_k son vectores de parámetros de distribución de longitud n. La asignación de distribuciones a priori $p_{jk}(\cdot)$ a los componentes individuales del modelo da como resultado el logaritmo a posteriori (Umlauf et al., 2018):

$$\log \pi\,(\beta, \tau;\ \mathrm{y}, \mathrm{X}, \alpha)\ \alpha\ \ell(\beta;\ \mathrm{y}, \mathrm{X}) + \sum_{K=1}^{K}\ \sum_{j=1}^{J_k}\left[\log \mathrm{p}_{jk}\left(\beta_{jk}; \tau_{jk}, \alpha_{jk}\right)\right]$$

Donde $\tau = (\tau_1^\top, \ldots, \tau_K^\top)^\top = \left(\tau_{11}^\top, \ldots, \tau_{J_1 1}^\top, \ldots \tau_{1K}^\top, \ldots \tau_{J_K K}^\top\right)^\top$ es el vector de todos los hiperparámetros asignados utilizados en las funciones a priori $p_{jk}(\cdot)$ y α de forma similar, es el conjunto de todas las especificaciones a priori fijas. Más precisamente, la distribución a priori bastante general para el término jk del modelo viene dada por (Umlauf et al., 2018).

$$p_{jk}(\beta_{jk};\tau_{jk},\alpha_{jk})\alpha d_{\beta_{jk}}\left(\beta_{jk}|\tau_{jk};\alpha_{\beta_{jk}}\right).d_{\tau_{jk}}\left(\tau_{jk}\middle|\alpha_{\tau_{jk}}\right)$$

Con densidades a priori (o combinaciones de densidades) $d_{\beta_{jk}}(\cdot)$ y $d_{\tau_{jk}}(\cdot)$ que dependen del tipo de covariable y de supuestos previos sobre f_{jk} $(\cdot)$. En este marco, suelen ser varianzas que, por ejemplo, tienen en cuenta el grado de suavidad de $f_{jk}(\cdot)$ o la cantidad de correlación entre las observaciones. Por ejemplo, al usar una representación spline de f_{jk} $(\cdot)$ en combinación con una distribución a priori normal para $d_{\beta_{jk}}(\cdot)$, las varianzas pueden interpretarse como los parámetros de suavizado inversos en un contexto de regresión penalizada; es decir, desde una perspectiva frecuentista, puede considerarse como una log-verosimilitud penalizada. Además, las especificaciones previas fijas α_{jk}= $\{\alpha_{\beta_{jk}},\alpha_{\tau_{jk}}\}$ pueden controlar aún más la forma de $d_{\beta_{jk}}$ $(\cdot)$ y $d_{\tau_{jk}}(\cdot)$, incorporar creencias previas sobre β_{jk}, o para los modelos de tipo GAM α_{jk} generalmente contiene las llamadas matrices de penalización, entre otras.

2.3 Bases filosóficas

La investigación se sustenta en un enfoque cuantitativo de carácter aplicado, inscrito dentro del paradigma positivista, el cual concibe el conocimiento científico como resultado del análisis sistemático de datos observacionales mediante métodos formales, replicables y verificables (Creswell, 2014; Hernández, Fernández & Baptista, 2014). Este marco es coherente con los objetivos del estudio, orientados al modelamiento estadístico predictivo de la densidad acústica (s_A) de los cardúmenes de anchoveta, a partir de información numérica obtenida mediante sistemas acústicos.

Desde esta perspectiva, el conocimiento se genera a partir de la observación empírica del fenómeno, el uso de modelos estadísticos como representaciones formales del proceso observado y la evaluación crítica de su desempeño mediante criterios objetivos de validación. En este estudio, los registros acústicos y los descriptores morfológicos de los

cardúmenes constituyen la base empírica para la inferencia, mientras que el razonamiento estadístico y computacional permite identificar regularidades, evaluar supuestos y contrastar enfoques alternativos de modelamiento.

Asimismo, el estudio adopta una concepción predictiva e inferencial de la estadística, en la que la validez del conocimiento no se fundamenta exclusivamente en pruebas de significancia o en la explicación del valor medio, sino en la correcta especificación del modelo, la coherencia distribucional de los residuos y la capacidad predictiva condicionada a las covariables. Este posicionamiento es consistente con los desarrollos contemporáneos de la estadística aplicada en ciencias ambientales y pesqueras, donde los datos observacionales suelen presentar heterogeneidad, asimetría y colas pesadas que deben ser modeladas explícitamente (Stasinopoulos et al., 2017; Rigby et al., 2019; Stasinopoulos et al., 2024).

En consecuencia, el enfoque filosófico adoptado respalda el uso de modelos distribucionales modernos, como los GAMLSS y sus extensiones, los cuales permiten representar de manera más fiel la estructura estadística de los datos acústicos (Flores et al., 2025; Flores et al., 2026) y constituyen una base metodológica coherente para el desarrollo de un servicio pesquero predictivo, orientado al aprovechamiento indirecto de información generada por ecosondas comerciales no calibradas.

2.4 Definición de términos básicos

2.4.1. Servicio

Fisher y Navarro (1994) definen los servicios como un tipo de bien económico perteneciente al sector terciario, caracterizado por la provisión de actividades intangibles que satisfacen necesidades específicas sin generar un bien físico tangible. En este sentido, toda actividad orientada a la generación de valor a través del conocimiento, la información o la toma de decisiones se enmarca dentro del concepto de servicio.

Desde una perspectiva de gestión, Kotler (1997) define un servicio como una acción, desempeño o beneficio que una parte ofrece a otra, el cual es esencialmente intangible, no transferible en propiedad y puede o no estar asociado a un producto físico. En el contexto pesquero, un servicio puede materializarse como un sistema analítico que transforma datos en información útil para la toma de decisiones operativas y estratégicas.

2.4.2. Industria pesquera

La industria pesquera se define como el conjunto de actividades económicas vinculadas a la extracción, procesamiento, conservación, transporte, comercialización y distribución de recursos hidrobiológicos, incluyendo tanto la pesca de captura como la acuicultura. Este sector desempeña un rol estratégico a nivel global al contribuir de manera significativa a la seguridad alimentaria, la generación de empleo y el desarrollo económico de numerosas regiones costeras.

En países con ecosistemas altamente productivos, como el Perú, la industria pesquera constituye uno de los pilares de la economía nacional, lo que exige el uso de herramientas científicas y tecnológicas que permitan optimizar la explotación de los recursos bajo criterios de sostenibilidad y eficiencia.

2.4.3. Modelado estadístico

El modelado estadístico es un proceso mediante el cual se construyen representaciones matemáticas del comportamiento de un fenómeno aleatorio, asignando probabilidades a los posibles resultados observados. Un modelo estadístico puede adoptar la forma de una distribución de probabilidad discreta o continua, dependiendo de la naturaleza de los datos, y permite describir, explicar y predecir la variabilidad inherente al proceso bajo estudio.

De acuerdo con Tadikamalla y Durcikova (2003), el uso de modelos probabilísticos constituye la base de la inferencia estadística moderna, al permitir comparar explicaciones

alternativas de los datos observados y evaluar la plausibilidad de distintos supuestos sobre la estructura del proceso generador de datos. En aplicaciones predictivas, el modelado estadístico cumple un rol central al proporcionar estimaciones cuantificables de incertidumbre y riesgo.

2.4.4. Cardumen de peces

Un cardumen se define como un grupo de tres o más peces que se desplazan de manera coordinada, ajustando continuamente su posición y trayectoria en función del comportamiento de los individuos vecinos (Partridge, 1982). Desde una perspectiva ecológica y pesquera, un cardumen también puede entenderse como una alta concentración de individuos de una misma especie que ocupa un espacio relativamente delimitado y mantiene cohesión durante intervalos temporales prolongados (Fréon y Misund, 1999).

En estudios acústicos, el cardumen se caracteriza por su densidad, extensión espacial y morfología, propiedades que determinan su detectabilidad y la intensidad de la retrodispersión acústica registrada por los sistemas de observación.

2.4.5. Descriptores morfológicos

Los descriptores morfológicos son variables métricas que permiten cuantificar la forma, tamaño, estructura interna y organización espacial de un cardumen de peces. Estas métricas describen aspectos como la altura, largo, área y volumen del cardumen, y reflejan el grado de agregación y complejidad espacial del grupo (Robotham et al., 2012).

En el contexto de la acústica pesquera, los descriptores morfológicos constituyen una fuente clave de información, ya que influyen directamente en la intensidad y variabilidad de la señal acústica registrada, y permiten establecer relaciones cuantitativas entre la estructura del cardumen y la densidad acústica observada.

2.4.6. Factores espacio-temporales

Los factores espacio-temporales son variables que describen la ubicación geográfica y el momento temporal en que ocurre un fenómeno biológico o ecológico, y se utilizan para analizar patrones de distribución, movimiento y dinámica poblacional. Estos factores integran información espacial y temporal con el fin de comprender cómo los procesos ecológicos varían a lo largo del espacio y del tiempo (Murakami et al., 2023).

Las escalas temporales hacen referencia al intervalo de tiempo en el cual se analizan dichos procesos, pudiendo abarcar desde variaciones horarias hasta patrones interanuales o decádicos. La adecuada selección de la escala temporal es fundamental para interpretar correctamente la dinámica de los ecosistemas y la respuesta de las especies a los forzantes ambientales (Quijas y Balvanera, 2013).

2.5 Hipótesis de la investigación

2.5.1. Hipótesis general

H_0: La regresión clásica centrada en la media con transformaciones logarítmicas es equivalente a la regresión distribucional en la escala original, en el modelado de la densidad acústica de los cardúmenes de anchoveta, en términos de correcta especificación residual, representación de la estructura de los datos y desempeño predictivo.

H_1: La regresión clásica centrada en la media con transformaciones logarítmicas no es equivalente a la regresión distribucional en la escala original, en el modelado de la densidad acústica de los cardúmenes de anchoveta, debido a diferencias en la especificación residual, representación de la estructura de los datos y desempeño predictivo.

2.5.2. Hipótesis específicas

H_0: La regresión clásica centrada en la media con transformaciones logarítmicas y la regresión distribucional en la escala original no difieren en su capacidad para representar la estructura de los datos, en términos de variabilidad, asimetría y colas pesadas.

H_1: La regresión distribucional en la escala original representa de manera más adecuada la estructura de los datos, en términos de variabilidad, asimetría y colas pesadas, que la regresión clásica centrada en la media con transformaciones logarítmicas.

H_0: Ninguna variable explicativa presenta efecto sobre la densidad típica ($Q_{0.5}$) ni sobre las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta.

H_1: Al menos una variable explicativa presenta efecto sobre la densidad típica ($Q_{0.5}$) o sobre las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta.

H_0: No existe evidencia de diferencias en la especificación entre la regresión clásica centrada en la media y la regresión distribucional en la escala original; los residuos cuantílicos son compatibles con una distribución normal estándar $N(0,1)$, sin patrones sistemáticos persistentes en los diagramas de gusano o de cubo, y no existen diferencias sustanciales en las métricas de desempeño predictivo.

H_1: La regresión clásica centrada en la media presenta mala especificación, evidenciada por residuos cuantílicos con desviaciones sistemáticas respecto a $N(0,1)$; en contraste, la regresión distribucional en la escala original reduce dichas desviaciones y presenta mejor desempeño predictivo que los modelos clásicos.

H_0: Los modelos ajustados no son aplicables como base metodológica de un servicio pesquero predictivo de la densidad acústica (s_A) a escala de cardumen.

H_1: Los modelos ajustados son aplicables como base metodológica de un servicio pesquero predictivo de la densidad acústica (s_A) a escala de cardumen.

2.6 Operacionalización de las variables

Tabla 1

Operacionalización de variables

VARIABLE	DEFINICIÓN	DIMENSIÓN	INDICADOR
Servicio pesquero predictivo	Sistema analítico aplicado orientado a generar predicciones evaluables y reproducibles para optimizar las operaciones de pesca, mejorar la eficiencia y respaldar la toma de decisiones en la actividad pesquera, superando los enfoques puramente descriptivos (Hilborn & Mangel, 1997; Burnham & Anderson, 2002; Vásquez-Quispesivana et al., 2022; Chen et al., 2024; Pantraki et al., 2024).	Descriptores morfológicos del cardumen	Altura media (m) Largo corregido (m) Área corregida (m^2) Volumen 3D (m^3)
		Factores espacio-temporales	Profundidad media (m) Distancia a la costa (mn) Posición geográfica (°) Periodicidad horaria (h)
		Validez estadística del servicio	Residuos cuantílicos ~ $N(0,1)$ Diagramas de gusano y cubo Funciones de autocorrelación y autocorrelación parcial
		Capacidad predictiva del servicio	Error Absoluto Medio (m^2/mn^2) Raíz del Error Cuadrático Medio (m^2/mn^2) Puntuación probabilística continua ordenada (m^2/mn^2)
Densidad de los cardúmenes de anchoveta	Densidad acústica asociada a cardúmenes individuales ($s_A > 0$), expresada en m^2/mn^2 (Nottestad et al., 2021).	Densidad relativa	Coeficiente del área de dispersión (m^2/mn^2)

Nota. Elaboración propia.

CAPÍTULO III.

3. METODOLOGÍA

3.1 Diseño metodológico

La investigación presenta un enfoque cuantitativo de carácter aplicado y un diseño no experimental, retrospectivo y transversal, al analizar observaciones independientes de cardúmenes de anchoveta por estaciones del año.

El nivel de investigación es predictivo, al evaluar el desempeño predictivo de la densidad acústica en función de covariables como base metodológica de un servicio pesquero.

3.2 Población y muestra

3.2.1. Población

La población estuvo compuesta de 356457 observaciones de los descriptores morfológicos de los registros acústicos de anchoveta (incluye peces individuales y cardúmenes) extraídos de los datos en bruto (raw) de las ecosondas científicas SIMRAD EK60, con transductores de la frecuencia de 120 kHz, y de los factores espacio-temporales del medio marino, obtenidos de las evaluaciones acústicas realizadas por el IMARPE entre 2010 y 2017 a lo largo de la costa peruana, desde los 3° 20' S hasta los 18° 20' S.

3.2.2. Muestra

La muestra estuvo compuesta de 53088 observaciones de los descriptores morfológicos de los registros acústicos de los cardúmenes de anchoveta extraídos de los datos en bruto (raw) de las ecosondas científicas SIMRAD EK60, con transductores de la frecuencia de 120 kHz, y de los factores espacio-temporales del medio marino, obtenidos de las evaluaciones acústicas realizadas por el IMARPE entre 2010 y 2017 a lo largo de la costa peruana, desde los 3° 20' S hasta los 18° 20' S.

3.2.3. Recolección de datos de la densidad de los cardúmenes de anchoveta

Los datos sobre la densidad de los cardúmenes de anchoveta se obtuvieron con ecosondas científicas SIMRAD EK60 previamente calibradas, según lo descrito en Demer et al. (2015), utilizando un transductor de haz dividido de 120 kHz de frecuencia. El escrutinio de los ecotrazos se realizó en el programa Echoview, donde se transforma el S_v en el coeficiente de dispersión del área náutica (NASC o s_A) expresado en m^2/mn^2. La información sobre los cardúmenes de anchoveta se exportó a una base de datos que contiene los valores s_A de peces individuales y de cada cardumen, así como de las dimensiones de los descriptores morfológicos que los conforman. Los descriptores de la morfología de los cardúmenes de anchoveta se obtienen mediante el módulo de detección 2D del software Echoview, que permite identificar o reconocer un cardumen. En este módulo se han implementado algunos algoritmos del software SHAPES (sistema de estimación de parches y análisis de cardúmenes) desarrollado por Sea Fisheries Research Institute (SFRI) que define una serie de variables básicas de análisis capaces de detectar, extraer y caracterizar estadísticamente cardúmenes de peces de forma automatizada y objetiva a partir de datos acústicos (Barange, 1994; Coetzee, 2000; Lawson et al., 2001). Coetzee (2000) utilizó el sistema SHAPES para caracterizar datos acústicos de los cardúmenes de sardina. Diner (2001) describió correcciones para la geometría y la densidad de los cardúmenes teniendo en cuenta el patrón del haz y los efectos del umbral S_v.

3.2.4. Descriptores morfológicos utilizando algoritmos SHAPES

A continuación, se describe la formulación del cálculo de las variables extraídas de los cardúmenes utilizando el sistema SHAPES.

3.2.4.1 Altura media

La altura media es la proyección de espesor medio sobre el eje vertical teniendo en cuenta la geometría del transductor (Echoview Software Pty Ltd, 2025). El espesor medio es definido como:

$$T = \sum_{p}^{in\ A_p} t_p \Big/ N_p$$

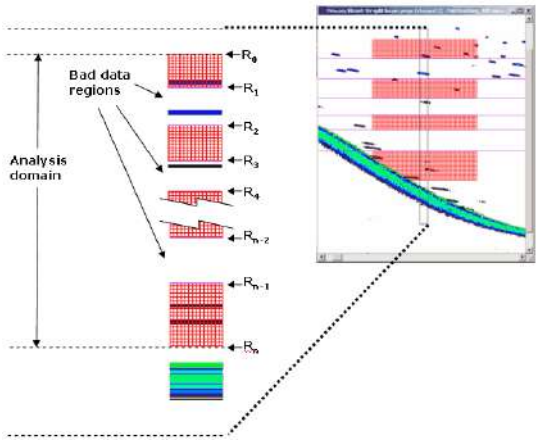
Donde A_p es el conjunto de pings p en el dominio de análisis, T es el espesor medio (m), N_p es el número de pings en el dominio de análisis y t_p es el espesor del ping p (m), calculado. Para ecogramas de S_v:

$$t_p = \Delta t_p \sum_{s}^{in\ A_s} \varepsilon_s$$

Donde A_s es el conjunto de muestras s en ping p en el dominio de análisis, Δt_p es el espesor de una muestra o el espacio entre muestras para ping p (m) y ε_s es el factor o valor de la muestra (s). (Figura 7).

Figura 7

Límites en un ping para el cálculo de t_p



Nota: Tomada de Echoview Software Pty Ltd, 2025

3.2.4.2 Profundidad media

La distancia desde la superficie del mar hasta el centro geométrico del banco de peces (Reid, 2000).

$$r = \sum_{s}^{in\ A} \varepsilon_s r_s \Big/ \sum_{s}^{in\ A} \varepsilon_s$$

Donde r es el rango medio (profundidad media en metros), r_s es el rango de muestras s en el dominio (m), *A* el conjunto de muestras *s* en el dominio de análisis que se está considerando y ε_s es el factor o valor de la muestra (Echoview Software Pty Ltd, 2025).

3.2.4.3 Largo corregido

Dimensión horizontal en el plano del ecograma corregida según la geometría conocida del haz (Reid, 2000).

$$Lc = (\mathrm{L} - 2 \cdot \mathrm{D} \cdot tan(\Phi/2))$$

Donde L es el largo sin corregir del rectángulo delimitador de la región.

$$L = Distancia(pe) - Distancia(ps) + Tamaño\ del\ ping(pe)$$

Donde pe y ps son los números del último y primer ping (final e inicio) del cardumen, la Distancia (pi) es la distancia (m) a recorrer hasta el i-ésimo ping (pi), el tamaño del ping es el espesor de pe (Echoview Software Pty Ltd, 2025).

3.2.4.4 Área de imagen corregida

Área de sección transversal (en el plano del ecograma) de un cardumen representada por una región en un ecograma corregido según la geometría conocida del haz (Reid, 2000).

$$A_{2D} = \frac{A\,(L_c \cdot T_c)}{L \cdot T} \text{ si } L \cdot T \neq 0$$

$$A_{2D} = 0 \text{ si } L \cdot T = 0$$

$$A = \text{Área de imagen 2D sin corregir}$$

$$A = \sum_i a_i$$

Donde la suma es sobre todas las muestras de datos en la región.

$a_i =$ *área de la muestra*

$i =$ $r_i w_i$ para el ping p_i

$w_i =$ *ancho del ping i (m)*

$r_i =$ *espacio entre muestras (m) de i*

3.2.4.5 Volumen del cardumen 3D

Volumen (m^3) estimado de un cardumen suponiendo que es cilíndrico.

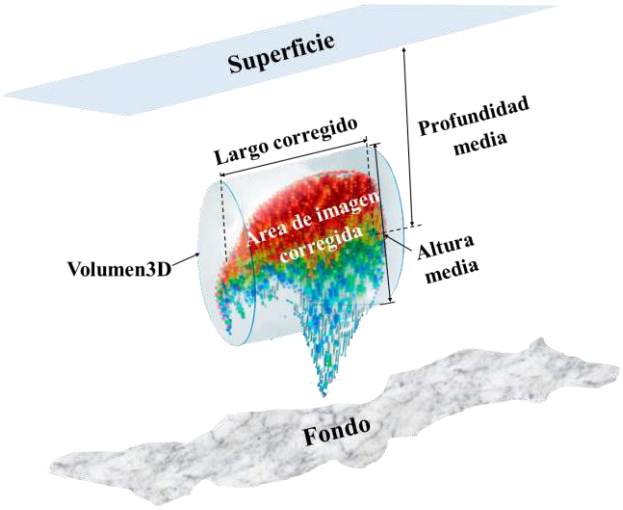
$$V = \sum j\, \pi . (Tj/2)^2 \, . Lj$$

Donde j es el número de ping, Tj es el espesor de la región en el ping j expresado como Σ_i T_i y T_i es el tamaño del intervalo de muestra i en ping j (Echoview Software Pty Ltd, 2025).

Los datos exportados mediante el módulo de detección 2D del software Echoview contienen información correspondiente a peces individuales y cardúmenes de anchoveta. Para extraer los descriptores morfológicos correspondientes a un cardumen, se aplicaron los siguientes criterios: s_A mínimo = 100 m^2/mn^2, altura media mínima = 1 m, largo corregido mínimo = 1 m, área corregida mínima = 1 m^2, volumen3D mínimo = 1 m^3 (Figura 8).

Figura 8

Esquema tridimensional *de un cardumen de anchoveta*



Nota. Descriptores morfológicos y espaciales derivados de ecogramas. Elaboración propia.

3.2.5. Variables espaciales y temporales

- Distancia a la costa (dc) mínima estimada en millas náuticas (mn) entre la posición geográfica registrada durante el muestreo de campo y la costa peruana.
- Posición geográfica (longitud y latitud) registrada durante el muestreo de campo a lo largo de la costa desde los 03° hasta los 18° de latitud sur.
- Periodicidad horaria expresada en horas y minutos registrada durante el muestreo de campo.

3.3 Técnicas para procesamiento de la información

3.3.1. Análisis exploratorio

Se realizó un análisis exploratorio de los datos separados por estaciones utilizando el software estadístico R (Ihaka & Gentleman, 1996). Esta separación se realizó dado que estudios recientes demuestran que la dependencia de la anchoveta respecto a los predictores

ambientales, espaciales y temporales difiere estacionalmente. Dicha heterogeneidad se refleja en variaciones en la distribución de la densidad acústica (s_A) a lo largo del año, la cual puede ser adecuadamente representada mediante distribuciones Box–Cox t ajustada a cero (ZABCTo) o Gaussiana Inversa Generalizada ajustada a cero (ZAGIG); en consecuencia, los modelos estacionales describen de forma más adecuada la relación entre s_A y las variables explicativas (Flores et al., 2026).

La estadística descriptiva básica se obtuvo mediante las funciones mean y sd, empleadas para calcular la media y la desviación estándar de la variable respuesta. La forma de la distribución empírica se evaluó mediante los coeficientes de asimetría y curtosis, estimados con las funciones skewness y kurtosis, con el fin de explorar desviaciones respecto a la normalidad.

La adecuación de la distribución empírica a una distribución normal teórica se examinó adicionalmente mediante la prueba de Jarque–Bera, aplicada con la función jarque.bera.test del paquete tseries. La homocedasticidad se evaluó mediante la prueba de Breusch–Pagan, implementada con la función bptest del paquete lmtest. Estas pruebas se utilizaron con un propósito exploratorio, orientado a caracterizar la naturaleza asimétrica, leptocúrtica y heterocedástica de los datos.

Adicionalmente, se evaluó la independencia residual mediante las funciones de autocorrelación y autocorrelación parcial de los residuos cuantílicos normalizados. Para ello, se construyeron gráficos ACF y PACF por estación, con el fin de identificar la posible persistencia de dependencia serial no explicada por el modelo. La ausencia de autocorrelaciones significativas dentro de las bandas de confianza fue considerada como evidencia complementaria de que la estructura temporal de los residuos no presentaba patrones sistemáticos remanentes.

3.3.2. División de los datos en conjuntos de entrenamiento, validación y prueba

El conjunto total de datos fue dividido en cada estación y en tres subconjuntos independientes: un conjunto de entrenamiento (60 %), utilizado para la selección de las covariables de los modelos; un conjunto de prueba (20 %), empleado para el ajuste y diagnóstico residual; y un conjunto de validación (20 %), para la evaluación del desempeño predictivo.

3.3.3. Selección del modelo

Para seleccionar el modelo que describe adecuadamente la distribución de la variable de respuesta, se emplearon diferentes herramientas estadísticas implementadas en el lenguaje R: el paquete stats, mediante la función glm(); mgcv, para modelos aditivos generalizados (Wood, 2006); INLA, para la aproximación de Laplace anidada integrada en modelos latentes gaussianos (Rue et al., 2009; Martins et al., 2013); VAST, para modelos vectoriales autorregresivos espacio-temporales con campos espaciales y espacio-temporales latentes (Thorson y Barnett, 2017; Thorson, 2019); gamlss, para modelos aditivos generalizados para localización, escala y forma (Rigby y Stasinopoulos, 2005); gamlss2, su versión extendida (Stasinopoulos et al., 2024); gamlss.dist, para las familias de distribuciones (Stasinopoulos y Rigby, 2017); gamlss.add, para términos aditivos adicionales de GAMLSS (Stasinopoulos y Rigby, 2024); y bamlss, para modelos aditivos bayesianos para localización, escala y forma (Umlauf et al., 2018).

Las variables explicativas incluyeron cuatro descriptores morfológicos de los cardúmenes: altura (m), largo (m), área (m^2), volumen3D (m^3), cuatro predictores espaciales que indican su ubicación en el espacio marino: profundidad media (m), distancia a la costa (mn), Longitud (°) y Latitud (°) y un predictor temporal que indica la periodicidad horaria: tiempo (horas).

3.3.3.1 Regresión clásica centrada en la media condicional

Para modelar la respuesta transformada Y=log(s_A+1), se ajustaron GLM asumiendo una distribución Gamma con función de enlace identidad. El modelo lineal será:

$$E[Y] = \beta_0 + \beta_1 \text{Altura} + \beta_2 \text{Largo} + \beta_3 \text{Área} + \beta_4 \text{Volumen}_{3D} + \beta_5 \text{Latitud} + \beta_6 \text{Longitud} + \beta_7 \text{Profundidad} + \beta_8 \text{Hora} + \beta_9 dc$$

donde E{·} denota el valor esperado o media.

Para permitir efectos no lineales, se ajustaron GAM utilizando el paquete mgcv (Wood, 2006). La variable de respuesta fue la transformación Y=log(s_A+1) con una distribución Gamma y función de enlace identidad. Todos los predictores se representaron mediante splines cúbicos de regresión penalizados (cs). El modelo con efectos no lineales será:

$$E[Y] = s(\text{Altura}) + s(\text{Largo}) + s(\text{Área}) + s(\text{Volumen3D}) + s(\text{Latitud}) + s(\text{Longitud}) + s(\text{Profundidad}) + s(\text{Hora}) + s(dc)$$

donde E{·} denota el valor esperado o media, y s(·) representa una función suave estimada mediante splines cúbicos de regresión con penalización (shrinkage), lo que permite que términos no informativos sean penalizados hacia cero. Los parámetros de suavizamiento se seleccionaron mediante Máxima Verosimilitud Restringida (REML).

Se ajustaron dos modelos bayesianos mediante INLA: un modelo INLA-Gamma sin campo espacial latente y un modelo INLA-Gamma con campo espacial latente continuo mediante ecuaciones diferenciales parciales estocásticas (SPDE). Ambos modelos permitieron representar la inferencia dentro de la clase de modelos latentes gaussianos, estimando las distribuciones posteriores marginales mediante aproximaciones de Laplace integradas anidadas (Rue, Martino, & Chopin, 2009; Martins, Simpson, Lindgren, & Rue, 2013).

Para el modelo INLA-Gamma sin campo latente espacial, la especificación fue:

$$Y_i \mid \mu_i, \phi \sim \text{Gamma}(\mu_i, \phi), i = 1, \ldots, n$$

$$g(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + f_1(\text{Altura}_i) + f_2(\text{Largo}_i) + f_3(\text{Área}_i) + f_4(\text{Volumen3D}_i) + f_5(\text{Latitud}_i) + f_6(\text{Longitud}_i) + f_7(\text{Profundidad}_i) + f_8(\text{Hora}_i) + f_9(dc_i)$$

Para el modelo INLA-Gamma con campo espacial latente, la estructura espacial bidimensional se incorporó mediante un campo gaussiano $u(s_i)$, definido sobre la localización $s_i = (\text{Longitud}_i, \text{Latitud}_i)$. En este caso, latitud y longitud no se incluyeron como suavizadores independientes, para evitar duplicar la estructura espacial:

$$\eta_i = \beta_0 + f_1(\text{Altura}_i) + f_2(\text{Largo}_i) + f_3(\text{Área}_i) + f_4(\text{Volumen3D}_i) + f_5(\text{Profundidad}_i) + f_6(\text{Hora}_i) + f_7(dc_i) + u(s_i)$$

$$u(s) \sim GF(0, C_{\text{Matérn}})$$

$$(\kappa^2 - \Delta)^{\alpha/2} \tau u(s) = W(s)$$

donde η_i representa el predictor lineal latente, β_0 es el intercepto, $f_j(\cdot)$ son efectos suavizados de las covariables, ϕ es el parámetro de precisión de la distribución Gamma, $u(s_i)$ es el campo espacial latente gaussiano, κ controla el alcance espacial, τ controla la varianza marginal del campo, Δ es el operador laplaciano y $W(s)$ representa ruido blanco espacial gaussiano.

Al incorporar los efectos suavizados $f_j(\cdot)$ de las covariables en el predictor latente, ambos modelos INLA-Gamma corresponden a GAM bayesianos, es decir, la misma estructura aditiva centrada en la media condicional del GAM ajustado con mgcv, pero con inferencia bayesiana aproximada (Rue et al., 2009; Blangiardo y Cameletti, 2015).

La formulación SPDE permite aproximar el campo latente continuo mediante un campo aleatorio gaussiano de Markov sobre una malla triangular, haciendo viable la inferencia bayesiana mediante INLA (Lindgren, Rue, & Lindström, 2011). La comparación

entre ambos modelos permitió evaluar si la incorporación explícita de dependencia espacial corrige los patrones residuales observados en el modelo INLA-Gamma sin campo latente.

La validación residual se basó en residuos cuantílicos normalizados $z_i = \Phi^{-1}(u_i) \sim \mathrm{N}(0,1)$, evaluados mediante diagramas de gusano y de cubo. En INLA-Gamma, al disponer de CDF cerrada, u_i se obtuvo analíticamente por la transformada integral de probabilidad (PIT).

Se ajustó un modelo VAST-Gamma con campo espacial latente como Delta-Gamma, pero con probabilidad de encuentro fija en 1, por lo que el ajuste correspondió al componente positivo Gamma.

$$Y_i = s_{A,i} > 0$$

$$Y_i \mid \mu_i, \phi \sim \mathrm{Gamma}(\mu_i, \phi)$$

$$g(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + f_1(\mathrm{Altura_i}) + f_2(\mathrm{Largo_i}) + f_3(\mathrm{Área_i}) + f_4(\mathrm{Volumen3D_i}) + f_5(\mathrm{Profundidad_i}) + f_6(\mathrm{Hora_i}) + f_7(dc_i) + \Omega_2(s_i)$$

donde $s_i = (\mathrm{Longitud_i}, \mathrm{Latitud_i})$, $f_j(\cdot)$ corresponde a splines naturales de las covariables estandarizadas y $\Omega_2(s_i)$ representa el campo espacial latente gaussiano del componente positivo. El campo espacial Ω_2 fue activado, mientras que el componente espacio-temporal ϵ_2 fue desactivado para mantener el ajuste por estación.

Al representar las covariables mediante splines naturales como base fija del predictor lineal, el modelo VAST-Gamma corresponde a un GLM con campo espacial latente gaussiano definido sobre una malla triangular, y no a un modelo aditivo penalizado (Thorson y Barnett, 2017).

La validación residual se realizó con residuos simulados tipo DHARMa transformados a residuos cuantílicos normalizados:

$$z_i = \Phi^{-1}(u_i)$$

y evaluados mediante diagramas de gusano y diagramas de cubo.

3.3.3.2 Regresión distribucional

El paquete estadístico GAMLSS, utilizó el procedimiento descrito por Stasinopoulos et al. (2017) para seleccionar los términos del modelo mediante la metodología "step-GAIC", basada en el criterio de información generalizado de Akaike (GAIC).

Este procedimiento evalúa de manera iterativa la incorporación y eliminación de términos, comparando en cada etapa el modelo actual con modelos candidatos que difieren en un único término, y reteniendo únicamente aquellos cambios que producen una reducción del GAIC. En este estudio se adoptó una penalización conservadora (k = 4), con el objetivo de controlar el sobreajuste y favorecer modelos parsimoniosos, especialmente en la modelación de parámetros no asociados a la media. La selección se aplicó de forma independiente a los parámetros de locación (μ), escala (σ) y asimetría (ν), ya que cada uno describe distintos aspectos de la distribución de la variable respuesta. En consecuencia, algunos términos son excluidos aun cuando resultan relevantes, debido a que su contribución marginal al ajuste no compensa el incremento en complejidad del modelo, o porque su efecto es redundante respecto de otros predictores ya incluidos.

Esto es coherente en modelos con alta colinealidad entre las variables explicativas y refleja que distintos procesos pueden influir diferencialmente en la media, la variabilidad y la forma de la distribución. En particular, la selección de términos para el parámetro de asimetría (ν) tendió a ser más restrictiva, ya que este parámetro controla el comportamiento de las colas de la distribución y no la estructura central de los datos. De este modo, el procedimiento step-GAIC permitió identificar, para cada parámetro distribucional, el conjunto mínimo de covariables necesarias para describir adecuadamente la estructura de los datos, garantizando un equilibrio entre calidad de ajuste, estabilidad numérica e interpretabilidad del modelo.

Inicialmente, se utilizó la distribución Box-Cox t, que es una distribución flexible de cuatro parámetros (μ, σ, ν, τ) para $Y > 0$, y que permite una cola derecha pesada (Rigby et al., 2006). Los predictores continuos del modelo incorporan splines de suavizado penalizados (Eilers y Marx 1996). Para comprobar la distribución adecuada en el rango (0, ∞), se utilizó el método "choose-D" (Rigby et al., 2019). Los términos seleccionados para los parámetros μ, σ, ν, τ mediante la metodología "step-GAIC" y la distribución adecuada para la respuesta elegida con el método "choose-D", se utilizaron para ajustar modelos GAMLSS con la función "gamlss", redes neuronales (NN) con la función aditiva "nn" del paquete gamlss.add, que permite aplicar en el contexto de los modelos GAMLSS la función "nnet" del paquete nnet (Ripley, 2008) de R ampliamente utilizado para construir y entrenar redes neuronales, y bosques aleatorios (RF) con la función "cf" del paquete gamlss2 (Stasinopoulos et al., 2024) que es la versión más reciente y extendida de GAMLSS.

Con el paquete estadístico BAMLSS se ajustó un modelo de regresión bayesiano con predictores aditivos, donde las medidas de variación se basan en la simulación de Cadena de Markov Monte Carlo (MCMC). Se empleó el método predeterminado que es mínimos cuadrados ponderados iterativamente (IWLS), en el que las muestras de varianzas de suavizado se calculan utilizando un muestreo por sectores (Umlauf et al. 2018). En BAMLSS se utilizó el modelo y la distribución seleccionada con los métodos "step-GAIC" y "choose-D", ya que admite familias GAMLSS mediante una función de transferencia que construye el objeto bamlss.family a partir de la función de distribución gamlss.dist (Stasinopoulos y Rigby, 2007). BAMLSS también admite funciones de suavizado del paquete mgcv (Wood, 2006). Los predictores continuos utilizaron splines penalizados, propuestos por Eilers y Marx (1996).

3.3.4. Evaluación del modelo ajustado

La comprobación de los supuestos distribucionales sobre los residuos de los modelos GLM, GAM, INLA, GAMLSS, NN, RF y BAMLSS se efectuó mediante herramientas de diagnóstico visual: diagramas Q-Q normalizados (Dunn y Smyth, 1996), diagramas de gusano (van Buuren y Fredriks, 2001) y diagramas de cubo (De Bastiani et al., 2022). Los diagramas de gusano son gráficas Q-Q sin línea de tendencia, que permiten identificar patrones sistemáticos lineales o cuadráticos en los residuos cuantílicos, indicativos de desajustes en la media, la varianza, la asimetría y la curtosis, y que evidencian que no siguen una distribución normal estándar independiente (Stasinopoulos et al., 2024).

Los diagramas de cubo verifican la presencia de asimetría y curtosis del momento en los residuos cuantílicos del modelo, como indicadores de un modelo incorrecto (van Buuren y Fredriks, 2001; van Buuren, 2007; Stasinopoulos et al., 2017).

3.3.5. Representación y estimación de los efectos de las variables explicativas sobre el s_A de los cardúmenes de anchoveta

El efecto de cada covariable sobre la densidad acústica (s_A) de los cardúmenes de anchoveta se interpretó mediante gráficas de Efectos Locales Acumulados (ALE; Apley y Zhu, 2020). A diferencia de las gráficas de términos y las gráficas de dependencia parcial (PDP) (que promedian las predicciones sobre combinaciones de covariables potencialmente inexistentes y se distorsionan cuando las variables explicativas están correlacionadas), las ALE particionan el rango de cada covariable en intervalos locales, promedian dentro de cada intervalo las diferencias de predicción condicionando a la distribución observada del resto de covariables, y las acumulan a lo largo del rango (Apley y Zhu, 2020; Molnar, 2022). Las curvas resultantes se centran verticalmente en cero, de modo que el recorrido vertical de cada una cuantifica la magnitud del efecto acumulado de la covariable. Se calcularon gráficas ALE para los cuantiles condicionales $Q_{0.5}$ y $Q_{0.95}$ en cada estación, con el fin de

explicar cómo influyen las variables explicativas en la densidad típica y en las densidades más altas.

3.3.6. Representación de la función de densidad de probabilidad (pdf) y estimación de la excedencia de umbrales ecológicamente relevantes del s_A

A partir de los parámetros μ, σ, ν y τ predichos por los modelos GAMLSS para cada estación, se derivaron la función de densidad de probabilidad (pdf) que representa la forma completa de la estructura de los datos, la función de distribución acumulativa (cdf) y la probabilidad de excedencia (1 − cdf) que permiten estimar la probabilidad que la variable aleatoria continua $Y = S_A$ sea inferior o superior respectivamente, a un valor dado y = s_A.

La pdf condicional $f(y)$, se estimó con las funciones dlogSEP3() y dlogSEP4() correspondiente a las distribuciones SEP3 y SEP4 adaptadas a escala logarítmica sobre el soporte positivo. La función de densidad ajustada $f(y)$ se evaluó sobre el soporte positivo, verificando que integra a la unidad, $\int_0^\infty f(y)dy = 1$.

La cdf se evaluó con las funciones plogSEP3() y plogSEP4() y se estimó la excedencia $P(S_A > s_A) = 1 - F(s_A)$, a umbrales fijos ($s_A$ = 250, 500, 1000, 2000, 5000 m²/mn²). La excedencia permite cuantificar la probabilidad de superar densidades ecológicamente relevantes, como diferencias en el extremo superior, lo que un modelo centrado en la media no captura.

Adicionalmente, se calculó la excedencia muestral (empírica) como la proporción de observaciones que superan cada umbral, $\hat{S}(s_A) = (1/n)\sum 1(s_A, i > s_A)$, como referencia observada frente a la cual contrastar la excedencia teórica derivada del modelo.

3.3.7. Predicción del parámetro de localización de la densidad de los cardúmenes de anchoveta

Los modelos GAMLSS, BAMLSS, NN-GAMLSS y RF-GAMLSS previamente entrenados permitieron predecir los parámetros de la distribución para nuevos valores de las

variables explicativas. Para ello, se utilizaron los descriptores morfológicos y factores espacio-temporales del conjunto de validación, correspondiente al 20% del total de datos. Las predicciones se realizaron mediante la función predictAll en GAMLSS (Rigby & Stasinopoulos, 2005; Rigby et al., 2019) y la función predict en los paquetes bamlss (Umlauf et al., 2018) y gamlss2 (Stasinopoulos et al., 2024).

Para cada observación i, el modelo estimó el vector de parámetros distribucionales:

$$\widehat{\theta}_i = (\widehat{\mu}_i, \widehat{\sigma}_i, \widehat{\nu}_i, \widehat{\tau}_i)$$

En la parametrización de las distribuciones logarítmicas de Potencia Exponencial Sesgada tipo 3 (logSEP3) y Potencia Exponencial Sesgada tipo 4 (logSEP4), el parámetro μ representa el logaritmo de la moda condicional. Por ello, la moda estimada del s_A en la escala original se obtuvo como:

$$\widehat{\text{Moda}}(s_{A,i}) = \exp(\widehat{\mu}_i)$$

Debido a que $\exp(\hat{\mu})$ representa la moda condicional y no la media, no es directamente comparable con la media empírica de los datos de validación. Por esta razón, para comparar las predicciones de los modelos con el promedio observado del s_A, se estimó la media teórica de la distribución predictiva:

$$\hat{E}(s_{A,i}) = E(Y_i \mid \widehat{\mu}_i, \widehat{\sigma}_i, \widehat{\nu}_i, \widehat{\tau}_i)$$

Dado que las familias logSEP3 y logSEP4 no presentan una media cerrada simple para la interpretación directa del parámetro μ, la media teórica se aproximó numéricamente mediante la función cuantil de la distribución predictiva:

$$\hat{E}(s_{A,i}) \approx \frac{1}{K}\sum_{k=1}^{K} Q(u_k \mid \widehat{\mu}_i, \widehat{\sigma}_i, \widehat{\nu}_i, \widehat{\tau}_i)$$

donde:

$$u_k = \frac{k - 0.5}{K}, \quad k = 1, \ldots, K$$

De esta manera, $\exp(\hat{\mu})$ se utilizó para representar la moda condicional del s_A, mientras que $\hat{E}(s_{A,i})$ permitió evaluar si los modelos reproducen adecuadamente el nivel promedio observado en el conjunto de validación.

3.3.8. Evaluación del desempeño predictivo de los modelos

Para evaluar el desempeño predictivo de los modelos se emplearon dos métricas de error puntual, el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE), y una métrica distribucional, la puntuación de probabilidad clasificada continua (CRPS). El MAE y el RMSE permitieron cuantificar la diferencia entre los valores observados de s_A y la media teórica estimada por cada modelo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

La puntuación CRPS evalúa la calidad de la distribución predictiva completa al comparar la función de distribución acumulada predicha $F_i(z)$ con el valor observado y_i. A diferencia de una métrica puntual, el CRPS penaliza distribuciones predictivas desplazadas, demasiado dispersas o mal concentradas alrededor del valor observado. Por tanto, valores menores de CRPS indican mejor desempeño distribucional del modelo.

$$CRPS(F_i, y_i) = \int_{-\infty}^{\infty}[F_i(z) - I(z \geq y_i)]^2 dz$$

Para cuantificar la incertidumbre asociada a las métricas de desempeño predictivo, se aplicó un procedimiento bootstrap no paramétrico sobre el conjunto de validación. En cada réplica bootstrap se remuestrearon con reemplazo las observaciones y sus

correspondientes predicciones, calculándose nuevamente el RMSE, el MAE y el CRPS. A partir de la distribución empírica de cada métrica se estimaron los percentiles 2.5% y 97.5%, los cuales fueron utilizados como límites inferior y superior del intervalo bootstrap al 95%.

Este procedimiento permitió evaluar no solo el valor puntual de las métricas, sino también su estabilidad frente a la variabilidad muestral. En consecuencia, la comparación entre GAMLSS, BAMLSS, NN-GAMLSS y RF-GAMLSS no se basó exclusivamente en el menor valor puntual de RMSE, MAE o CRPS, sino también en la amplitud y superposición de los intervalos bootstrap. Esta estrategia fue necesaria debido a la alta asimetría, heterogeneidad y presencia de valores extremos en la distribución del s_A.

CAPÍTULO IV.

4. RESULTADOS

4.1 Análisis de resultados

4.1.1. Análisis exploratorio de los datos

La estadística descriptiva del s_A muestra valores medios elevados y desviaciones estándar extremadamente grandes en todas las estaciones, lo que evidencia una variabilidad muy alta y la presencia de valores atípicos estructurales. Los coeficientes de asimetría y curtosis confirman que el s_A es una variable fuertemente asimétrica y con colas derechas pesadas. Las pruebas de Jarque–Bera y Breusch–Pagan indican desviaciones severas de la normalidad y heterocedasticidad significativa ($p < 0.001$). Estos resultados confirman que el s_A presenta una distribución altamente no normal, de cola pesada y con fuerte dispersión, lo que justifica el uso de modelos distribucionales definidos sobre el soporte positivo y capaces de representar colas derechas pronunciadas (Tabla 2).

Tabla 2

Estadística exploratoria de la densidad de la anchoveta (s_A) por estación

Estación	Media ± desviación estándar	Asimetría	Curtosis	JB χ^2	$\log_{10}(p_{JB})$	BP χ^2	$\log_{10}(p_{BP})$
Verano	7593.60 ± 13787.67	3.47	16.78	149030.00	-32361.40	890.30	-185.13
Otoño	6507.79 ± 12474.68	3.76	19.74	138545.00	-30084.70	665.10	-136.66
Invierno	5685.63 ± 12109.09	4.15	23.05	296373.00	-64356.50	821.50	-170.30
Primavera	7294.41 ± 13800.24	3.37	16.06	63628.00	-13816.60	427.00	-85.63

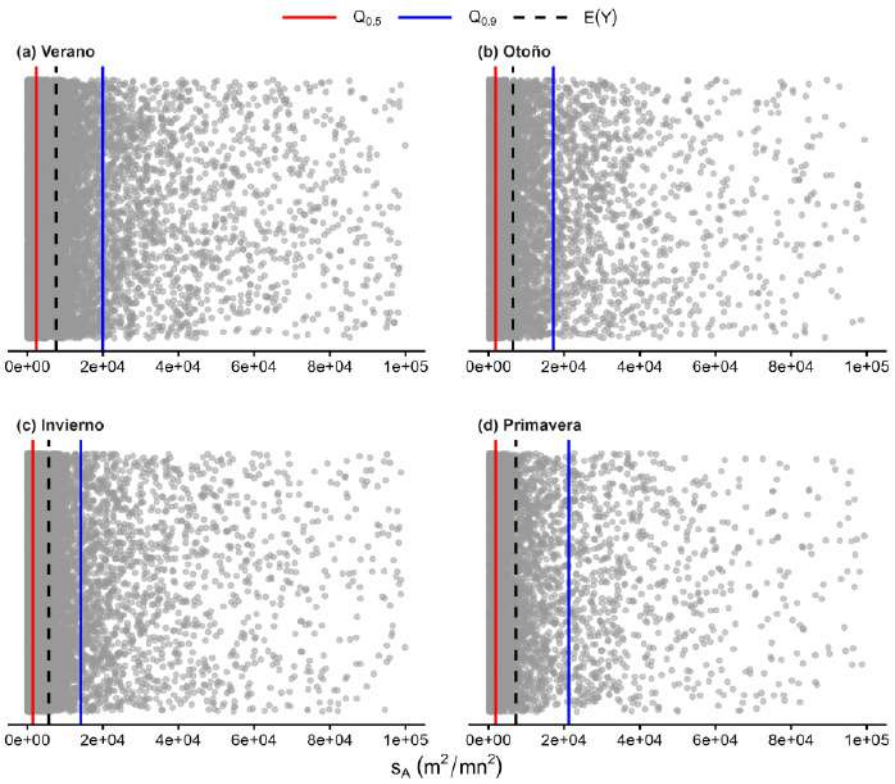
Nota. JB χ^2 = estadístico de la prueba de Jarque–Bera (normalidad); BP χ^2 = estadístico de la prueba de Breusch–Pagan (homocedasticidad), los valores p extremadamente pequeños expresan como $\log_{10}(p)$. Elaboración propia.

Los valores de la densidad acústica (s_A) de los cardúmenes de anchoveta presentan en todas las estaciones un predominio de valores bajos, con la mediana ($Q_{0.50}$) cercana al origen. En las cuatro estaciones la media [E(Y)] superó ampliamente a la mediana y se ubicó en el

tramo superior de la distribución ($Q_{0.75}$-$Q_{0.80}$), lo que confirma que el valor medio no representa la densidad típica de los cardúmenes y evidencia una fuerte asimetría positiva (Tabla 2). Esta característica desafía los supuestos de la regresión clásica y requiere el uso de marcos estadísticos flexibles que incorporen una distribución definida sobre la línea real positiva con una cola larga hacia la derecha (Figura 9).

Figura 9

Valores observados de la densidad acústica (s_A) de los cardúmenes de anchoveta por estación



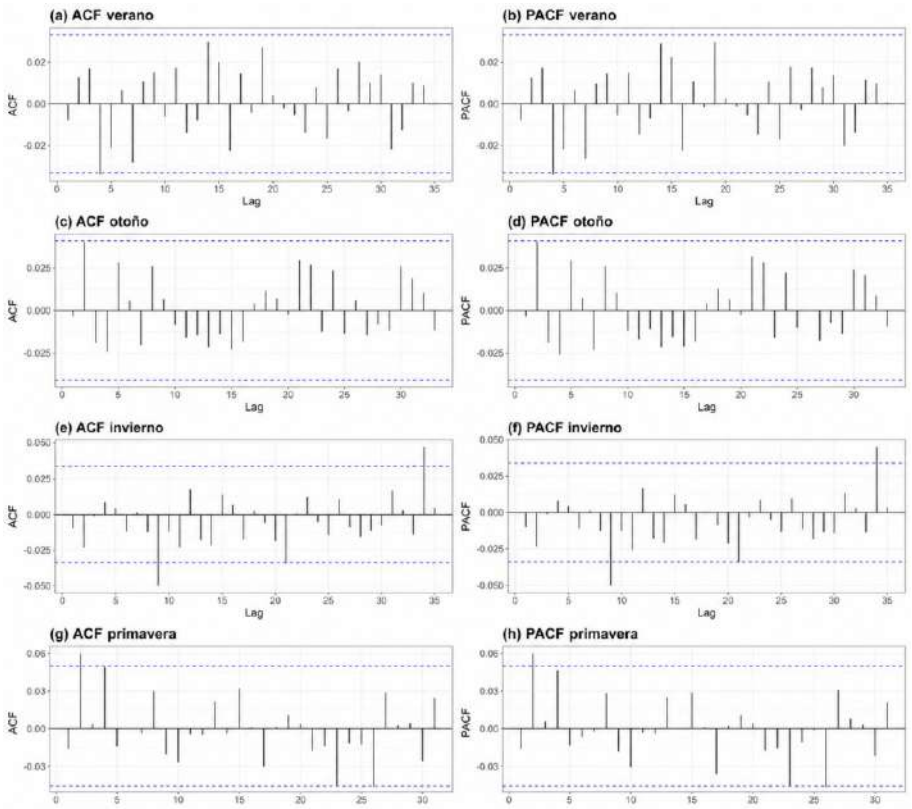
Nota. (a) verano, (b) otoño, (c) invierno y (d) primavera. Cada panel muestra la mediana ($Q_{0.5}$, línea roja), la media (línea negra) y el cuantil 90% ($Q_{0.9}$, línea azul). Elaboración propia.

La Figura 10 no muestra autocorrelación en los residuos cuantílicos normalizados de los modelos GAMLSS ajustados por estación. En verano y otoño un solo valor alcanza el límite del intervalo de confianza (retardos 4 y 1). En invierno y primavera algunos picos lo

superan (retardos 9 y 34; 1, 4, 23 y 26), pero son aislados y esperables por azar. Los residuos se comportan como ruido blanco, por lo que no fue necesario incorporar componentes autorregresivos.

Figura 10

Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de los residuos de los modelos GAMLSS por estaciones



Nota. verano (a, b), otoño (c, d), invierno (e, f) y primavera (g, h). Elaboración propia.

4.1.2. Regresión clásica centrada en la media

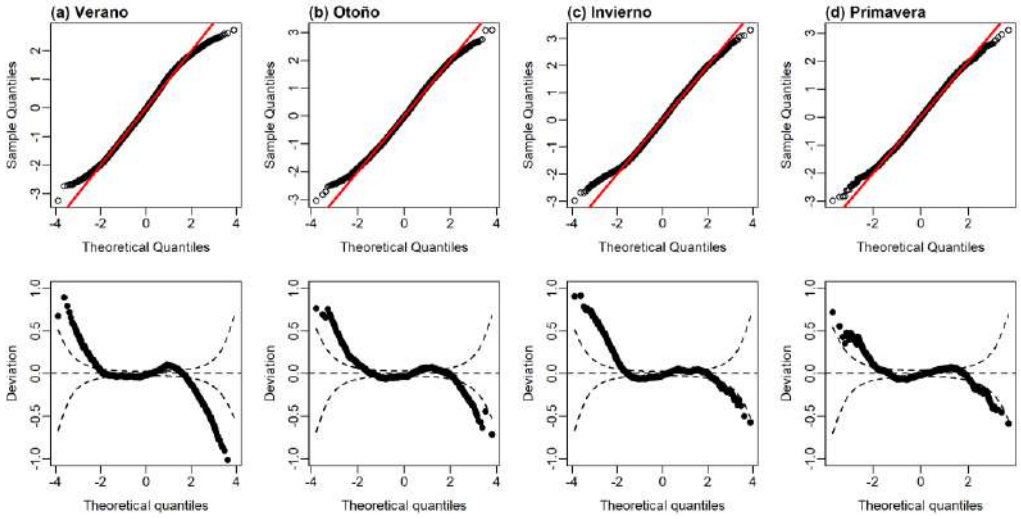
4.1.2.1 Modelos lineales generalizados (GLM)

En los diagramas de cuantiles normalizados, los modelos GLM con distribución Gamma para las cuatro estaciones presentan desviaciones en los residuales, principalmente en los extremos inferior y superior de la distribución ajustada por el modelo. Los diagramas de gusano exhiben residuales que exceden las bandas de confianza del 95%, representadas

por las hipérbolas superior e inferior (línea discontinua), lo que indica que las colas de la distribución ajustada son demasiado pesadas (van Buuren & Fredriks, 2001); es decir, la variabilidad teórica en los extremos está sobreestimada respecto a los datos observados, lo que evidencia un sesgo en la estimación de la curtosis o el parámetro de forma del modelo (Figura 11).

Figura 11

Diagramas cuantil cuantil normalizados y diagramas de gusano de los GLM-Gamma por estaciones



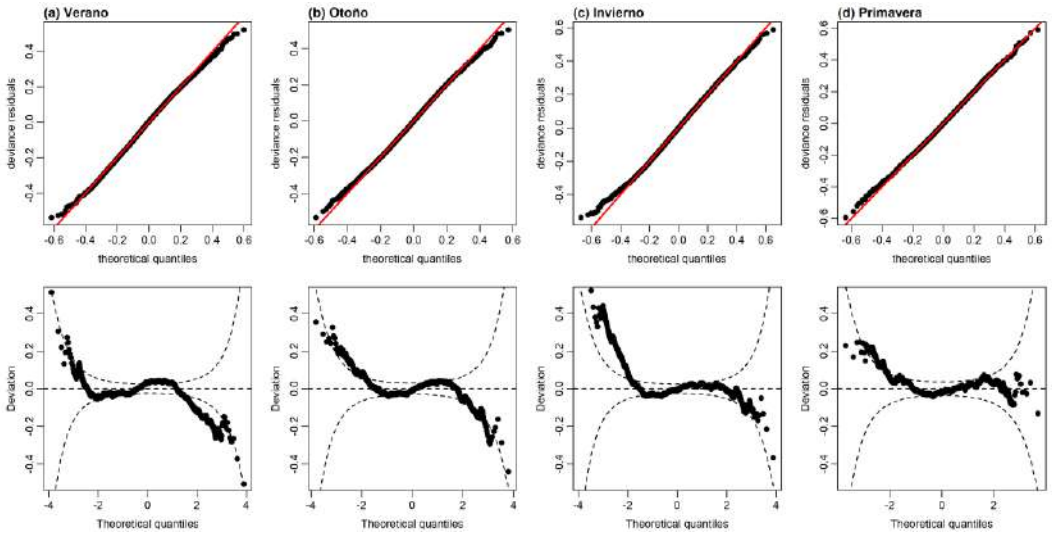
Nota. Elaboración propia.

4.1.2.2 Modelos aditivos generalizados (GAM)

En los diagramas de gusano los residuales exceden las bandas de confianza del 95%, representadas por las hipérbolas superior e inferior (línea discontinua), con un patrón en forma de S y el extremo izquierdo curvado hacia arriba, lo que indica que las colas de la distribución ajustada son demasiado pesadas (van Buuren & Fredriks, 2001), es decir, la variabilidad en las colas está sobreestimada y evidencia modelos incorrectos. El patrón es más pronunciado en verano y otoño, mientras que en invierno y primavera el exceso se concentra en el extremo inferior (Figura 12).

Figura 12

Diagramas cuantil cuantil normalizados y diagramas de gusano de los GAM-Gamma por estaciones



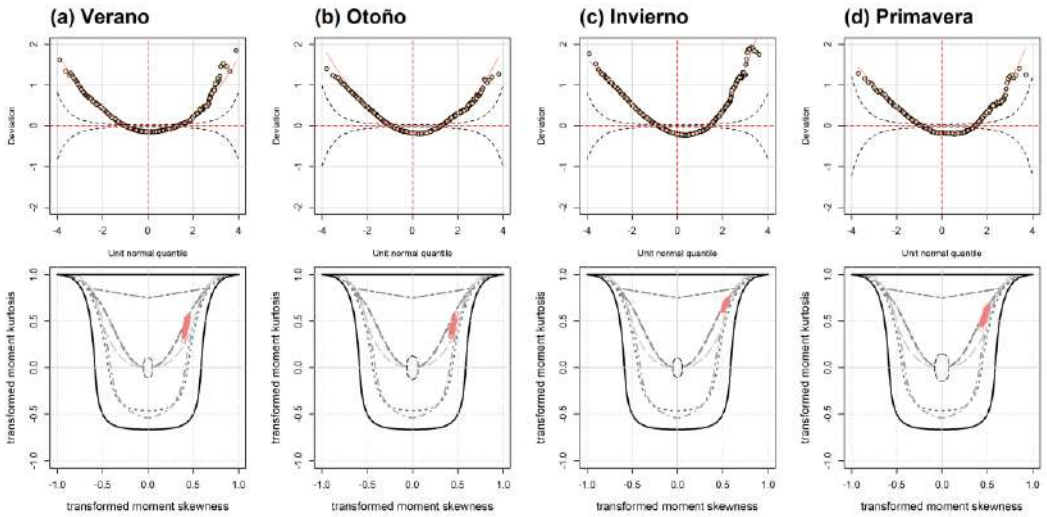
Nota. Elaboración propia.

4.1.2.3 Aproximación de Laplace integrada anidada (INLA)

Los diagramas de gusano de ambos modelos INLA con distribución Gamma (sin y con campo espacial latente continuo) muestran residuales con patrones sistemáticos cuadráticos, con formas en U, lo que evidencia que la distribución ajustada está demasiado sesgada hacia la izquierda (van Buuren & Fredriks, 2001); es decir, el modelo subestima tanto los valores bajos como los altos de la respuesta. En los diagramas de cubo, los puntos se ubican en el extremo superior derecho, fuera de la región elíptica en el centro del gráfico, lo que se asocia con exceso de asimetría y curtosis del momento, evidenciando un modelo ajustado incorrecto (Figuras 13 y 14).

Figura 13

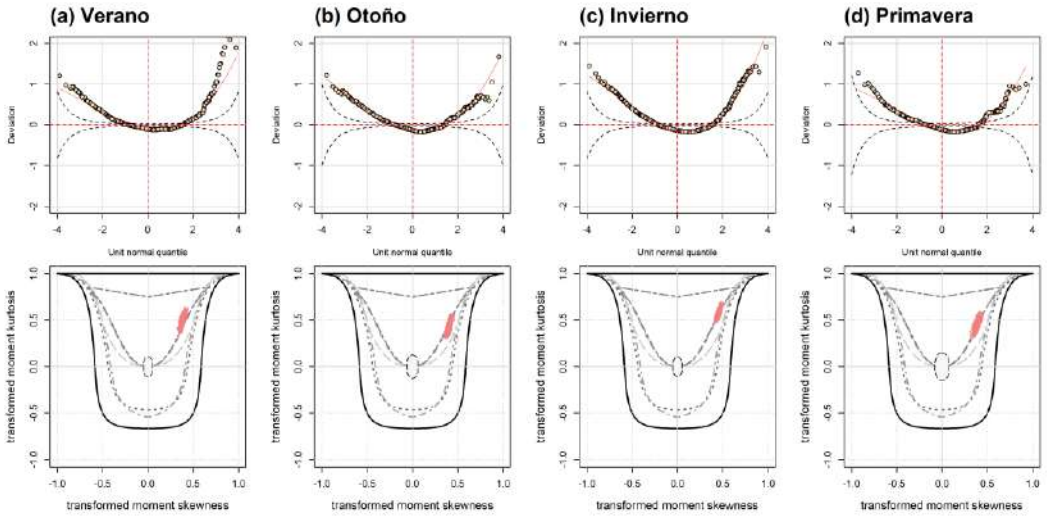
Diagramas gusano y cubo de los modelos INLA-Gamma sin campo espacial latente continuo por estaciones



Nota. Elaboración propia.

Figura 14

Diagramas gusano y cubo de los modelos INLA-Gamma con campo espacial latente continuo por estaciones



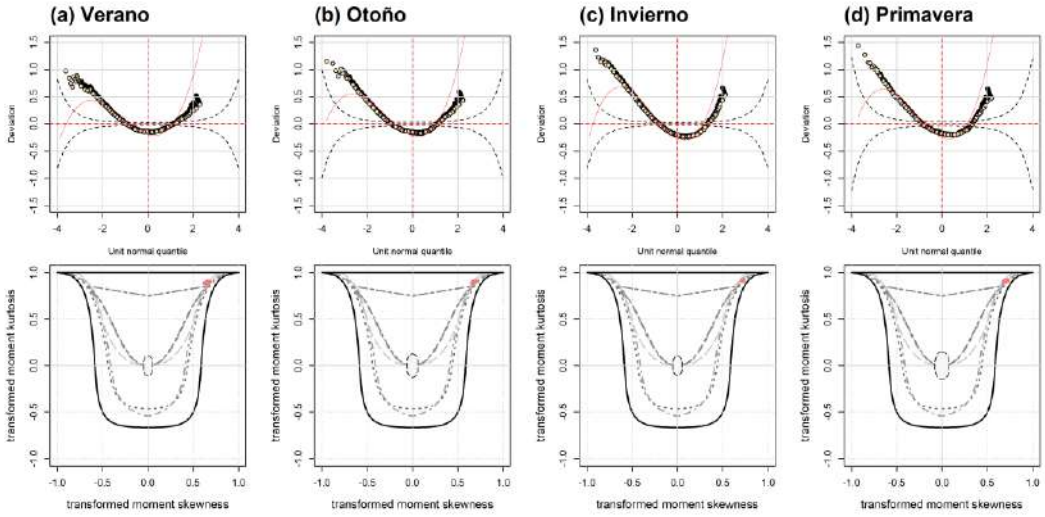
Nota. Elaboración propia.

4.1.3. Modelo espacio-temporal autorregresivo vectorial (VAST)

El modelo VAST con distribución Gamma y campo latente espacial de malla presentó una mala especificación residual evidente. En los diagramas de gusano, los residuos mostraron patrones sistemáticos cuadráticos, con forma en "U", lo que revela una distribución ajustada sesgada hacia la izquierda (van Buuren & Fredriks, 2001); es decir, el modelo subestima los cuantiles inferiores y superiores de la respuesta. En los diagramas de cubo, los residuos se ubicaron fuera de la región elíptica central, hacia el extremo superior derecho del gráfico, indicando la presencia de asimetría y curtosis del momento, lo que evidencia un modelo incorrecto (Figura 15).

Figura 15

Diagramas de gusano y cubo de los modelos VAST-Gamma con campo latente espacial de malla por estaciones



4.1.4. Enfoques estadísticos distribucionales

4.1.4.1 Distribución probabilística adecuada para la densidad de los cardúmenes de anchoveta

La distribución adecuada para el s_A de los cardúmenes de anchoveta en verano e invierno fue logSEP4, correspondiente a la transformación exponencial (logarítmica

inversa) de la distribución Skew Exponential Power type 4 SEP4 (Jones, 2005). La distribución adecuada para el s_A de los cardúmenes de anchoveta en otoño y primavera fue logSEP3, correspondiente a la transformación exponencial (logarítmica inversa) de la distribución Skew Exponential Power type 3 SEP3 (Fernandez et al., 1995).

Ambas son distribuciones continuas de cuatro parámetros definidas sobre la línea real:

$$-\infty < y < \infty$$

con parámetros

$$-\infty < \mu < \infty, \qquad \sigma > 0, \qquad \nu > 0, \qquad \tau > 0$$

La distribución SEP4 es denotada como:

$$Y \sim SEP4(\mu, \sigma, \nu, \tau)$$

La función de densidad de SEP4 puede escribirse como:

$$f_Y(y|\mu, \sigma, \nu, \tau) = \begin{cases} \frac{c}{\sigma} exp[-|z|^{\nu}], y < \mu \\ \frac{c}{\sigma} exp[-|z|^{\tau}], y \geq \mu \end{cases}$$

donde:

$$z = \frac{y - \mu}{\sigma}$$

y:

$$c = \left[\Gamma\left(1 + \frac{1}{\tau}\right) + \Gamma\left(1 + \frac{1}{\nu}\right)\right]^{-1}$$

$\Gamma(\cdot)$ corresponde a la función gamma. Esta forma coincide con la definición de la distribución SEP4, donde la densidad se expresa en función de $z = \frac{y-\mu}{\sigma}$, con constante de normalización $c = [\Gamma(1 + 1/\tau) + \Gamma(1 + 1/\nu)]^{-1}$, y con exponentes distintos para los lados izquierdo y derecho respecto de μ: ν para $y < \mu$ y τ para $y \geq \mu$. En esta parametrización, ν controla la pesadez de la cola izquierda y τ la pesadez de la cola derecha.

Para la distribución logSEP4, definida sobre la línea real positiva, se aplica el cambio de variable:

$$y = \log(x)$$

por lo que:

$$X = \exp(Y)\,, X \sim logSEP4(\mu, \sigma, \nu, \tau)$$

con soporte:

$$x > 0$$

La función de densidad de logSEP4 se obtiene evaluando la densidad SEP4 en log(x) e incorporando el jacobiano del cambio de variable $y = log(x)$, dado por $1/x$. Por tanto, la densidad puede escribirse como:

$$f_X(x|\mu, \sigma, \nu, \tau) = \begin{cases} \frac{c}{x\sigma}\exp(-|z_x|^{\nu})\,, \log(x) < \mu \\ \frac{c}{x\sigma}\exp(-|z_x|^{\tau})\,, \log(x) \geq \mu \end{cases}$$

donde:

$$z_x = \frac{\log(x) - \mu}{\sigma}$$

y:

$$c = \left[\Gamma\left(1 + \frac{1}{\tau}\right) + \Gamma\left(1 + \frac{1}{\nu}\right)\right]^{-1}$$

En logSEP4, el factor $1/x$ corresponde al jacobiano de la transformación $y = log(x)$. De esta manera, logSEP4 conserva la flexibilidad de SEP4 para representar diferente pesadez de colas a ambos lados de la moda, pero trasladada al soporte positivo requerido para modelar valores positivos de s_A. Esta propiedad resulta adecuada para datos acústicos de cardúmenes, caracterizados por alta asimetría, heterogeneidad y presencia de valores extremos.

La distribución SEP3 es denotada como:

$$Y \sim SEP3(\mu, \sigma, \nu, \tau)$$

La función de densidad de SEP3 puede escribirse como:

$$f_Y(y|\mu, \sigma, \nu, \tau) = \begin{cases} \frac{c}{\sigma} \exp\left[-\frac{1}{2}|\nu z|^{\tau}\right], y < \mu, \\ \frac{c}{\sigma} \exp\left[-\frac{1}{2}\left|\frac{z}{\nu}\right|^{\tau}\right], y \geq \mu, \end{cases}$$

Para la distribución logSEP3, definida sobre la línea real positiva, la función de densidad se obtiene mediante el cambio de variable $y = \log(x)$, por lo que:

$$X = \exp(Y), X \sim logSEP3(\mu, \sigma, \nu, \tau)$$

con soporte

$$-\infty < \mu < \infty$$

La función de densidad de logSEP3 puede escribirse como:

$$f_X(x|\mu, \sigma, \nu, \tau) = \begin{cases} \frac{c}{x\sigma} \exp\left[-\frac{1}{2}|\nu z_x|^{\tau}\right], \log(x) < \mu, \\ \frac{c}{x\sigma} \exp\left[-\frac{1}{2}\left|\frac{z_x}{\nu}\right|^{\tau}\right], \log(x) \geq \mu, \end{cases}$$

Donde

$$z = \frac{y - \mu}{\sigma}$$

$$z_x = \frac{\log(x) - \mu}{\sigma}$$

$$c = \frac{\nu\tau}{(1 + \nu^2)2^{1/\tau}\Gamma(1/\tau)}$$

y c corresponde a la constante de normalización. En logSEP3, el factor $1/x$ corresponde al jacobiano del cambio de variable $y = \log(x)$.

4.1.4.2 Modelos ajustados para los parámetros μ, σ, ν, τ de la densidad de los cardúmenes de anchoveta

En los modelos GAMLSS, BAMLSS, redes neuronales distribucionales (NN-GAMLSS) y bosques aleatorios distribucionales (RF-GAMLSS) se seleccionaron los

términos lineales y no lineales mediante el procedimiento "step-GAIC", aplicados de forma independiente para cada estación del año.

Todos los modelos distribucionales se ajustaron utilizando la familia Skew Exponential Power type 3 (logSEP3) y Skew Exponential Power type 4 (logSEP4), modelando los parámetros de locación (μ), escala (σ), asimetría (ν) y curtosis (τ).

Verano (logSEP4: $\boldsymbol{\mu, \sigma, \nu, \tau}$)

$$\hat{\mu} = s(Altura) + s(Latitud) + s(Profundidad) + s(Hora) + s(Longitud) + s(dc)$$
$$log(\hat{\sigma}) = s(Hora) + Altura + dc + s(Latitud) + s(Largo)$$
$$log(\hat{\nu}) = Profundidad + dc$$
$$log(\hat{\tau}) = Profundidad + dc$$

Otoño (logSEP3: $\boldsymbol{\mu, \sigma, \nu, \tau}$)

$$\hat{\mu} = s(Altura) + s(Longitud) + s(Hora) + s(dc) + s(Profundidad) + s(Volumen3D) + Latitud$$
$$log(\hat{\sigma}) = s(Hora) + s(Latitud) + s(Profundidad) + dc + Altura + Longitud$$
$$log(\hat{\nu}) = Altura + Área$$
$$log(\hat{\tau}) = 1$$

Invierno (logSEP4: $\boldsymbol{\mu, \sigma, \nu, \tau}$)

$$\hat{\mu} = s(Altura) + s(Hora) + s(Profundidad) + s(Latitud) + s(dc)$$
$$log(\hat{\sigma}) = s(Hora)$$
$$log(\hat{\nu}) = \text{Altura}$$
$$log(\hat{\tau}) = \text{Altura} + \text{Profundidad} + \text{s}(dc)$$

Primavera (logSEP3: $\boldsymbol{\mu, \sigma, \nu, \tau}$)

$$\hat{\mu} = s(Altura) + s(Latitud) + s(Longitud) + s(dc) + s(Profundidad) + s(Hora)$$
$$log(\hat{\sigma}) = s(Hora) + s(Latitud) + \text{Longitud} + \text{Volumen3D}$$
$$log(\hat{\nu}) = \text{Altura} + \text{Hora} + \text{Profundidad}$$
$$log(\hat{\tau}) = \text{Volumen3D} + \text{s}(Latitud)$$

4.1.4.2.1 GAMLSS

La Tabla 3 presenta la significancia aproximada de los términos incluidos en los modelos GAMLSS ajustados para la densidad acústica de los cardúmenes de anchoveta, evaluados de manera estacional y para los parámetros de la distribución de los datos: localización (μ), escala (σ), asimetría (ν) y curtosis (τ). Los valores p corresponden a pruebas tipo Wald y deben interpretarse como indicadores descriptivos de la contribución de los términos suavizados, dado que en presencia de efectos no lineales y estructuras aditivas complejas estos valores no constituyen pruebas de hipótesis.

Tabla 3

Resumen de los modelos GAMLSS por estaciones.

Parámetro	Término	Verano	Otoño	Invierno	Primavera
μ	Altura	$<2\times10^{-16}$	0.00318	$<2\times10^{-16}$	1.28×10^{-7}
	Latitud	0.00172	7.71×10^{-11}	$<2\times10^{-16}$	2.49×10^{-8}
	Longitud	2.05×10^{-12}	2.68×10^{-16}	–	3.43×10^{-5}
	Profundidad	$<2\times10^{-16}$	0.76439	3.20×10^{-8}	0.12314
	Hora	6.61×10^{-10}	0.51383	0.00299	$<2\times10^{-16}$
	dc	4.31×10^{-7}	3.96×10^{-8}	0.05662	0.12616
	Área	–	–	–	–
	Largo	–	–	–	–
	Volumen3D	–	0.7487	–	–
σ	Hora	2.63×10^{-6}	0.94849	8.92×10^{-5}	0.9585
	Altura	1.86×10^{-5}	0.00787	–	–
	dc	$<2\times10^{-16}$	0.00062	–	–
	Latitud	0.00549	0.01077	–	0.0863
	Largo	4.27×10^{-9}	–	–	–
	Profundidad	–	0.61611	–	–
	Longitud	–	0.17966	–	0.6483
	Volumen3D	–	–	–	0.0261
ν	Profundidad	0.01575	–	–	0.506
	dc	0.00935	–	–	–
	Altura	–	0.0159	1.10×10^{-5}	$<2\times10^{-16}$
	Área	–	0.1215	–	–
	Hora	–	–	–	1.28×10^{-8}
τ	Profundidad	$<2\times10^{-16}$	–	0.378	–
	dc	0.00978	–	0.575	–
	Altura	–	–	8.98×10^{-7}	–
	Volumen3D	–	–	–	$<2\times10^{-16}$
	Latitud	–	–	–	0.0481

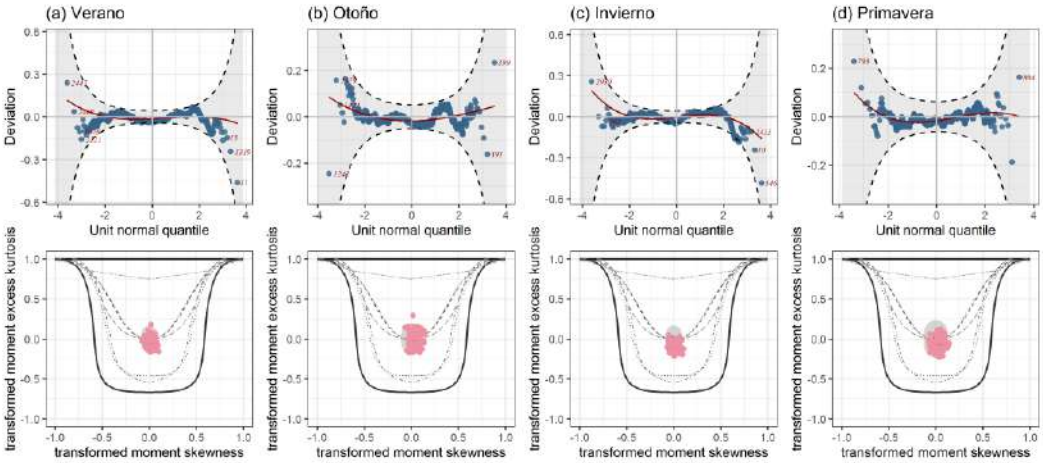
Nota. Elaboración propia.

En todos los modelos estacionales, los diagramas de gusano muestran un patrón residual aproximadamente horizontal, con alrededor del 95% de los puntos en el área sombreada entre las bandas de confianza aproximadas del 95%, representadas por las curvas

hiperbólicas superior e inferior. En los diagramas de cubo, los puntos se encuentran en la región elíptica sombreada en el centro del gráfico, que representa la región de confianza del 95 % para la prueba de Jarque-Bera, donde la asimetría y el exceso de curtosis del momento son iguales a cero, lo que demuestra que los modelos ajustados no son incorrectos (Figura 16).

Figura 16

Diagramas de gusano y cubo de los modelos GAMLSS por estaciones



Nota. Elaboración propia.

4.1.4.2.2 BAMLSS

La Tabla 4 resume las medias posteriores, los intervalos creíbles del 95% y las probabilidades de aceptación para los parámetros μ, σ, ν y τ de los modelos BAMLSS. Las altas probabilidades de aceptación indican un comportamiento estable y la amplitud de los intervalos de σ en otoño y primavera, evidencia mayor incertidumbre posterior.

Tabla 4

Resúmenes posteriores de los modelos BAMLSS por estaciones.

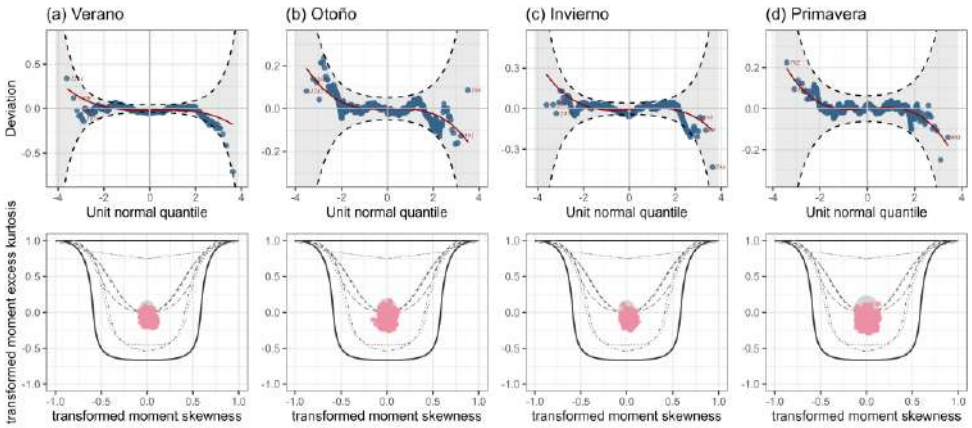
Estación	Parámetro	Media posterior	2.50%	50%	97.50%	Aceptación media	2.50%	50%	97.50%
Verano	μ	7.833	7.791	7.834	7.871	0.968	0.548	1	1
	σ	0.82	0.756	0.821	0.882	0.897	0.278	0.988	1
	ν	1.507	1.269	1.513	1.707	0.851	0.107	0.951	1
	τ	0.857	0.603	0.859	1.083	0.87	0.24	0.968	1
Otoño	μ	5.739	3.549	5.829	7.558	0.977	0.818	1	1
	σ	-1.549	-14.06	-0.648	5.84	0.89	0.43	0.972	1
	ν	0.43	0.295	0.429	0.569	0.888	0.234	0.984	1
	τ	1.21	1.088	1.209	1.335	0.984	0.871	1	1
Invierno	μ	7.377	7.327	7.377	7.431	0.989	0.908	1	1
	σ	0.715	0.675	0.716	0.756	0.982	0.831	1	1
	ν	1.58	1.389	1.583	1.754	0.868	0.15	0.971	1
	τ	0.442	0.29	0.442	0.61	0.892	0.368	0.972	1
Primavera	μ	7.092	6.828	7.095	7.369	0.753	0	0.996	1
	σ	1.49	-3.153	1.384	5.837	0.52	0	0.583	1
	ν	0.331	0.036	0.33	0.647	0.65	0	0.905	1
	τ	1.206	1.053	1.202	1.384	0.711	0	0.931	1

Nota. Elaboración propia.

En todas las estaciones, los diagramas de gusano mostraron un patrón residual aproximadamente horizontal, con la mayoría de los puntos dentro del área sombreada entre las bandas de confianza aproximadas del 95%, representadas por las curvas hiperbólicas superior e inferior. En los diagramas de cubo, los puntos se ubicaron en la región elíptica sombreada central, correspondiente a la región de confianza del 95% para la prueba de Jarque-Bera, donde la asimetría y el exceso de curtosis del momento son iguales a cero. El diagnóstico residual no evidenció modelos incorrectos (Figura 17).

Figura 17

Diagramas gusano y cubo para los modelos BAMLSS por estaciones



Nota. Elaboración propia.

4.1.4.2.3 Redes neuronales distribucionales

La Tabla 5 muestra la significancia de los interceptos de los modelos NN-GAMLSS ajustados por estación. Los valores de p corresponden a los interceptos de μ, σ, ν y τ, no a los componentes internos de la red neuronal, e indican la precisión con que se estimó el nivel base de cada parámetro.

Tabla 5

Resumen de los modelos NN-GAMLSS por estaciones

Estación	Parámetro	Intercepto	Error estándar	Valor p
Verano	μ	7.826	0.022	$<2\times10^{-16}$
	σ	0.756	0.015	$<2\times10^{-16}$
	ν	1.431	0.058	$<2\times10^{-16}$
	τ	1.002	0.052	$<2\times10^{-16}$
Otoño	μ	7.9	0.035	$<2\times10^{-16}$
	σ	2.348	0.032	$<2\times10^{-16}$
	ν	-0.094	0.021	7.99×10^{-6}
	τ	1.113	0.058	$<2\times10^{-16}$
Invierno	μ	7.547	0.026	$<2\times10^{-16}$
	σ	0.742	0.018	$<2\times10^{-16}$
	ν	1.416	0.064	$<2\times10^{-16}$
	τ	0.733	0.046	$<2\times10^{-16}$
Primavera	μ	7.341	0.033	$<2\times10^{-16}$
	σ	0.834	0.02	$<2\times10^{-16}$
	ν	1.633	0.109	$<2\times10^{-16}$
	τ	1.007	0.095	$<2\times10^{-16}$

Nota. Elaboración propia.

Los diagramas de gusano muestran residuales aproximadamente horizontales con alrededor del 95% entre las curvas hiperbólicas superior e inferior de las bandas de confianza aproximadas del 95% (área sombreada).

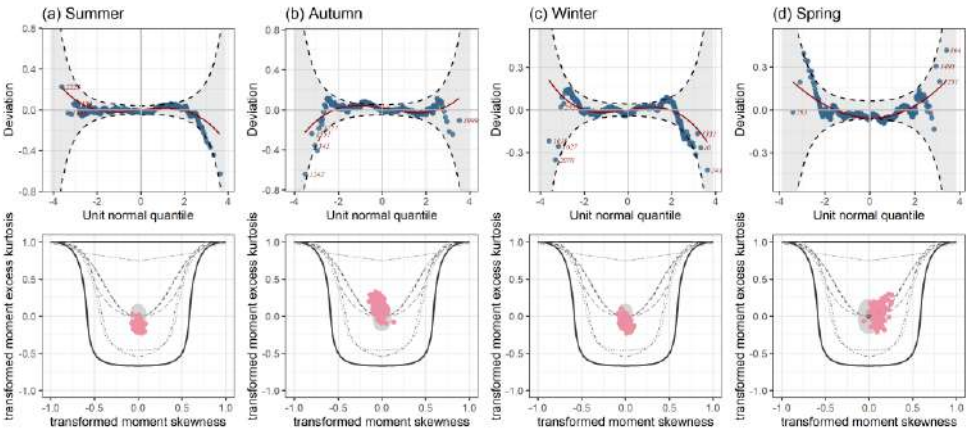
Los diagramas de cubo muestran en verano e invierno los residuales en la región elíptica sombreada en el centro del gráfico, que representa la región de confianza del 95 % para la prueba de Jarque-Bera, con asimetría y exceso de curtosis del momento iguales a cero. Sin embargo, en otoño y primavera la nube de puntos se desplazó parcialmente fuera de la elipse central (área sombreada), lo que indica presencia de asimetría y curtosis residual.

Según el diagnóstico residual, en verano e invierno no se evidenciaron modelos incorrectos, mientras que en otoño y primavera el desplazamiento fuera de la región central

del diagrama de cubo es compatible con limitaciones para capturar la asimetría y la curtosis (Figura 18).

Figura 18

Diagramas gusano y cubo de los modelos NN-GAMLSS por estaciones



Nota. Elaboración propia.

4.1.4.2.4 Bosques aleatorios distribucionales

La Tabla 6 muestra la significancia de los interceptos de los modelos RF-GAMLSS ajustados por estación. Los valores de p corresponden a los interceptos de μ, σ, ν y τ, no a los componentes internos del bosque aleatorio, e indican la precisión con que se estimó el nivel base de cada parámetro.

Tabla 6

Resumen de los modelos RF-GAMLSS por estaciones

Estación	Parámetro	Intercepto	Error estándar	Valor p
Verano	μ	7.823	0.063	$<2\times10^{-16}$
	σ	0.568	0.011	$<2\times10^{-16}$
	ν	1.169	0.055	$<2\times10^{-16}$
	τ	1.030	0.038	$<2\times10^{-16}$
Otoño	μ	7.408	0.049	$<2\times10^{-16}$
	σ	0.464	0.014	$<2\times10^{-16}$
	ν	0.145	0.019	1.36×10^{-14}
	τ	1.323	0.021	$<2\times10^{-16}$
Invierno	μ	7.306	0.056	$<2\times10^{-16}$
	σ	0.501	0.009	$<2\times10^{-16}$
	ν	1.077	0.022	$<2\times10^{-16}$
	τ	0.509	0.037	$<2\times10^{-16}$
Primavera	μ	7.462	0.074	$<2\times10^{-16}$
	σ	0.363	0.021	$<2\times10^{-16}$
	ν	0.117	0.028	4.02×10^{-5}
	τ	0.989	0.028	$<2\times10^{-16}$

Nota. Elaboración propia.

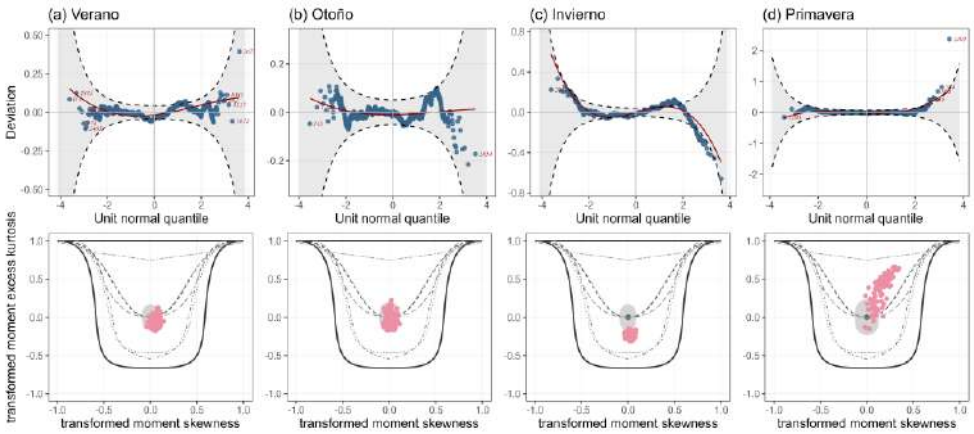
Los diagramas de gusano mostraron patrones residuales aproximadamente horizontales en las cuatro estaciones, con cerca del 95% de los puntos dentro de las bandas de confianza aproximadas del 95%. En invierno y primavera se observó una desviación leve en la cola derecha, donde algunos residuos extremos quedaron fuera de las bandas, consistente con la mayor dificultad de los bosques aleatorios para ajustar los valores extremos de la cola superior.

En los diagramas de cubo, los puntos de verano y otoño se ubicaron dentro de la región elíptica sombreada central, correspondiente a la región de confianza del 95% para la prueba de Jarque-Bera. En invierno y primavera, la nube de puntos se desplazó parcialmente fuera de la región central, indicando presencia de asimetría y curtosis residual.

El diagnóstico residual indica que en verano y otoño no se evidenciaron modelos incorrectos, mientras que en invierno y primavera las desviaciones en las colas y en la región central del diagrama de cubo son compatibles con modelos incorrectos (Figura 19).

Figura 19

Diagramas gusano y cubo de los bosques aleatorios distribucionales por estaciones



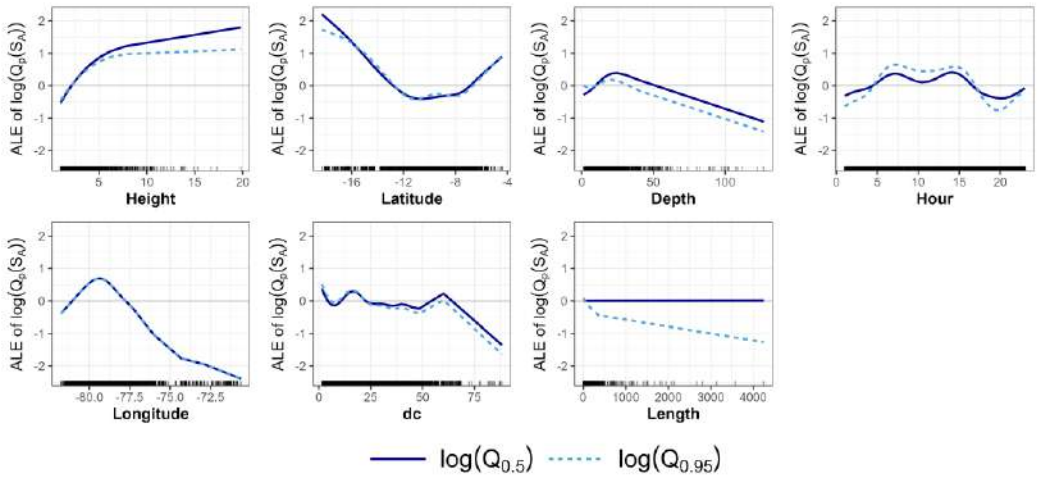
Nota. Elaboración propia.

4.1.5. Efectos de las variables explicativas sobre la densidad típica $Q_{0.50}$ y las densidades más altas $Q_{0.95}$ de los cardúmenes de anchoveta

En verano, sobre la densidad típica ($Q_{0.5}$) las variables espaciales latitud y longitud presentaron el mayor efecto: las altas densidades se localizaron al sur de los 12°S y al norte de los 8°S (rango ALE = 2.62), con valores máximos de longitud entre los 80.5°W y los 77.5°W (rango ALE = 3.08), dentro de las primeras 60 mn de la costa (rango ALE = 1.74) y entre los 10 y 50 m de profundidad (rango ALE = 1.5), lo que indica que las agregaciones más densas se ubicaron cerca de la superficie; la altura del cardumen al superar los 2 m mostró un aumento en la densidad acústica (rango ALE = 2.33), la hora evidenció un patrón no lineal con máximo diurno entre las 5 y 16 horas (rango ALE = 0.8), y el largo del cardumen no evidenció efecto (rango ALE = 0.01).

Figura 20

Efectos locales acumulados (ALE) sobre la densidad típica ($Q_{0.5}$) y las densidades extremas ($Q_{0.95}$) del s_A de los cardúmenes de anchoveta en verano.



Nota. Efecto sobre la densidad típica ($Q_{0.5}$) [línea azul continua] y las densidades extremas ($Q_{0.95}$) [línea azul discontinua] del s_A. Elaboración propia.

Sobre las densidades extremas ($Q_{0.95}$) las covariables espaciales tuvieron los mayores efectos: longitud (rango ALE = 3.08), distancia a la costa (rango ALE = 2.15), latitud (rango

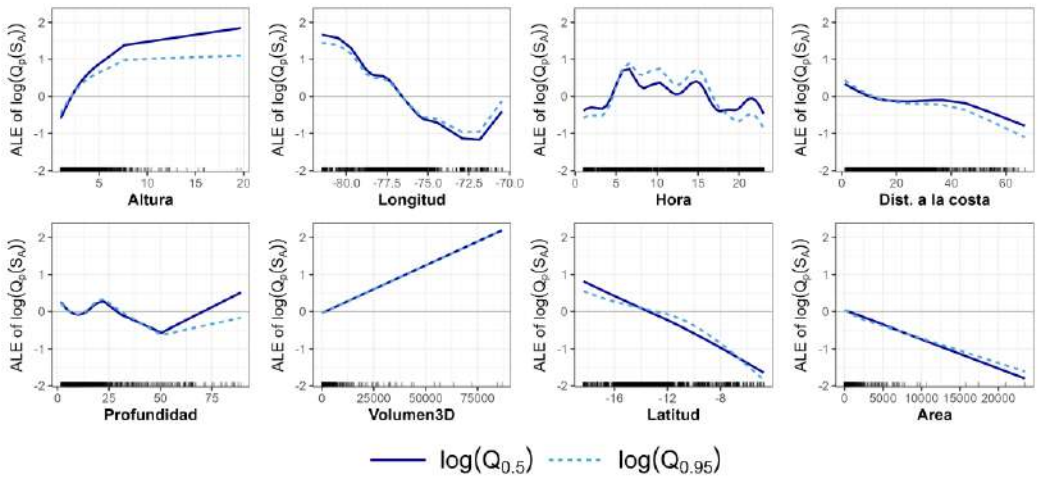
ALE = 2.14) y profundidad media (rango ALE = 1.59); la altura del cardumen fue el descriptor morfológico con mayor influencia (rango ALE = 1.58) y fueron menos influyentes la hora (rango ALE = 1.41) y el largo del cardumen (rango ALE = 1.35) (Figura 20, Tabla 7).

En otoño, sobre la densidad típica ($Q_{0.5}$) la longitud presentó el mayor efecto, con los valores más altos al oeste de los 76.5°W, máximo hacia los 81.5°W y decreciendo hacia el este (rango ALE = 2.84); le siguieron la latitud, con las mayores densidades al sur de los 13.5°S (rango ALE = 2.46), la altura del cardumen, que al superar los 2 m incrementó la densidad (rango ALE = 2.45), el volumen3D, que la incrementó al aumentar (rango ALE = 2.24), y el área, con un efecto decreciente con el tamaño del cardumen (rango ALE = 1.84). Los menores efectos correspondieron a la hora, con un patrón no lineal y densidades más altas entre las 5 y 16 horas (rango ALE = 1.22), la distancia a la costa, con las mayores densidades dentro de las 10 mn (rango ALE = 1.14), y la profundidad media, con incrementos dentro de las 25 mn y por fuera de las 75 mn (rango ALE = 1.08).

Sobre las densidades extremas ($Q_{0.95}$) los mayores efectos correspondieron a la longitud (rango ALE = 2.41), la latitud (rango ALE = 2.39) y el volumen3D (rango ALE = 2.24), seguidos de la hora (rango ALE = 1.76), el área (rango ALE = 1.66), la altura del cardumen (rango ALE = 1.57) y la distancia a la costa (rango ALE = 1.54); la profundidad media mostró el menor efecto (rango ALE = 0.96) (Figura 21, Tabla 7).

Figura 21

Efectos locales acumulados (ALE) sobre la densidad típica ($Q_{0.5}$) y las densidades extremas ($Q_{0.95}$) del s_A de los cardúmenes de anchoveta en otoño.



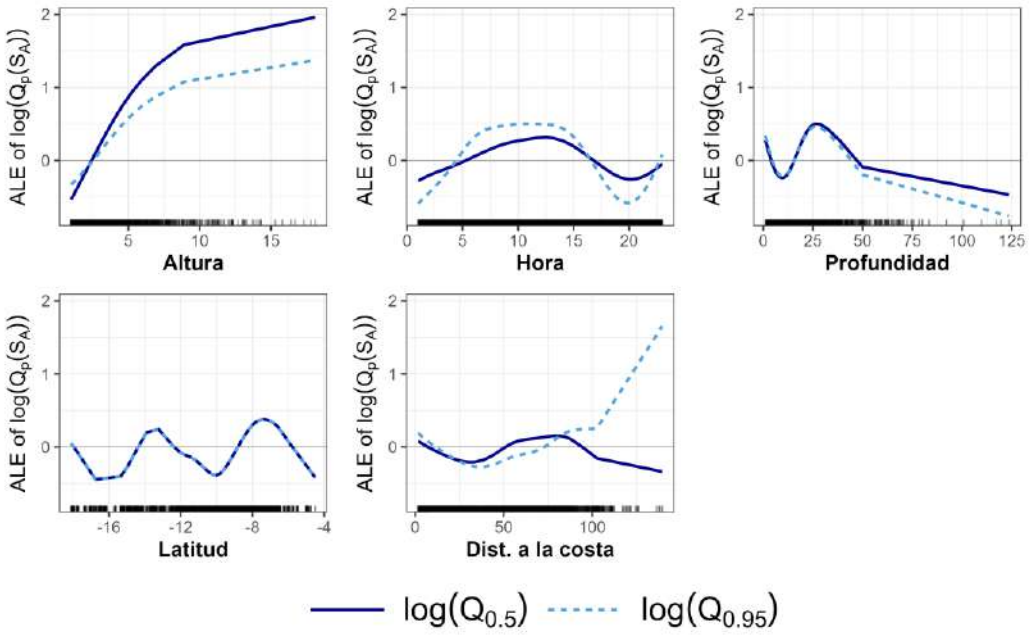
Nota. Efecto sobre la densidad típica ($Q_{0.5}$) [línea azul continua] y las densidades extremas ($Q_{0.95}$) [línea azul discontinua] del s_A. Elaboración propia.

En invierno, sobre la densidad típica ($Q_{0.5}$) la altura del cardumen presentó el mayor efecto, con un incremento marcado al superar los 2.5 m (rango ALE = 2.49); el resto de las variables mostró efectos menores: la profundidad media, con los valores más altos dentro de los 50 m (rango ALE = 0.97), la latitud, con un patrón no lineal y con incrementos entre 12°S-14°S y entre 9°S-5°S (rango ALE = 0.82), la hora, con máximos diurnos entre las 5 y 16 horas (rango ALE = 0.6), y la distancia a la costa (rango ALE = 0.49).

Sobre las densidades extremas ($Q_{0.95}$) el mayor efecto correspondió a la distancia a la costa, cuatro veces mayor que sobre la densidad típica (rango ALE = 1.93), con las mayores densidades hacia la zona oceánica por fuera de las 70 mn; le siguieron la altura del cardumen (rango ALE = 1.70), la profundidad media (rango ALE = 1.23), la hora (rango ALE = 1.09) y la latitud, que mantuvo el mismo efecto que sobre la densidad típica (rango ALE = 0.82) (Figura 22, Tabla 7).

Figura 22

Efectos locales acumulados (ALE) sobre la densidad típica ($Q_{0.5}$) y las densidades extremas ($Q_{0.95}$) del s_A de los cardúmenes de anchoveta en invierno.



Nota. Efecto sobre la densidad típica ($Q_{0.5}$) [línea azul continua] y las densidades extremas ($Q_{0.95}$) [línea azul discontinua] del s_A. Elaboración propia.

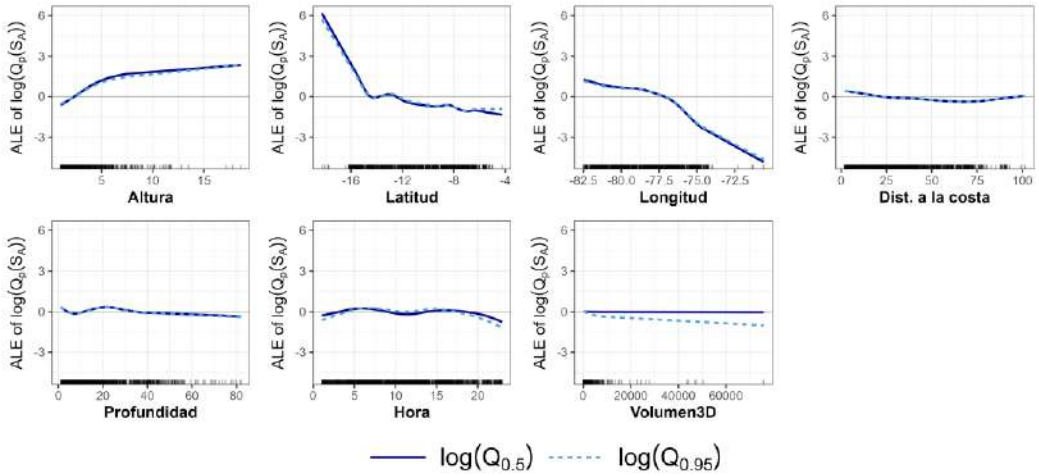
En primavera, sobre la densidad típica ($Q_{0.5}$) la latitud y la longitud duplicaron los efectos del resto de las covariables: la latitud, con incrementos al sur y disminución hacia el norte (rango ALE = 7.52), y la longitud, con los mayores valores al oeste de los 77°W y máximo hacia los 82.5°W (rango ALE = 6.12); le siguieron la altura del cardumen, que al superar los 2.3 m incrementó la densidad (rango ALE = 2.96), la hora, con un patrón no lineal y densidades más altas entre las 5 y 16 horas (rango ALE = 1.00), la distancia a la costa, con las mayores densidades dentro de las 22 mn (rango ALE = 0.80), y la profundidad media, con valores ligeramente mayores alrededor de los 20 m (rango ALE = 0.72); el volumen3D no evidenció efecto (rango ALE = 0.04).

Sobre las densidades extremas ($Q_{0.95}$) los mayores efectos correspondieron a la latitud (rango ALE = 6.77) y la longitud (rango ALE = 5.83), seguidas de la altura del

cardumen (rango ALE = 2.89), la hora (rango ALE = 1.44), el volumen3D (rango ALE = 1.07), la distancia a la costa (rango ALE = 0.80) y la profundidad media (rango ALE = 0.77) (Figura 23, Tabla 7).

Figura 23

Efectos locales acumulados (ALE) sobre la densidad típica ($Q_{0.5}$) y las densidades extremas ($Q_{0.95}$) del s_A de los cardúmenes de anchoveta en primavera.



Nota. Efecto sobre la densidad típica ($Q_{0.5}$) [línea azul continua] y las densidades extremas ($Q_{0.95}$) [línea azul discontinua] del s_A. Elaboración propia.

Tabla 7

Efectos de las covariables sobre la densidad típica ($Q_{0.5}$) y densidades extremas ($Q_{0.95}$) del s_A de los cardúmenes de anchoveta en cada estación

Cuantil	Verano	Rango ALE	Otoño	Rango ALE	Invierno	Rango ALE	Primavera	Rango ALE
Densidad típica ($Q_{0.5}$)	Longitud	3.08	Longitud	2.84	Altura	2.49	Latitud	7.52
	Latitud	2.62	Latitud	2.46	Profundidad	0.97	Longitud	6.12
	Altura	2.33	Altura	2.45	Latitud	0.82	Altura	2.96
	dc	1.74	Volumen3D	2.24	Hora	0.60	Hora	1.00
	Profundidad	1.50	Área	1.84	dc	0.49	dc	0.80
	Hora	0.80	Hora	1.22	—	—	Profundidad	0.72
	Largo	0.01	dc	1.14	—	—	Volumen3D	0.04
	—	—	Profundidad	1.08	—	—	—	—
Densidades extremas ($Q_{0.95}$)	Longitud	3.08	Longitud	2.41	dc	1.93	Latitud	6.77
	dc	2.15	Latitud	2.39	Altura	1.70	Longitud	5.83
	Latitud	2.14	Volumen3D	2.24	Profundidad	1.23	Altura	2.89
	Profundidad	1.59	Hora	1.76	Hora	1.09	Hora	1.44
	Altura	1.58	Área	1.66	Latitud	0.82	Volumen3D	1.07
	Hora	1.41	Altura	1.57	—	—	dc	0.80
	Largo	1.35	dc	1.54	—	—	Profundidad	0.77
	—	—	Profundidad	0.96	—	—	—	—

Nota. El mayor recorrido vertical (rango ALE) indica más efecto. Elaboración propia.

4.1.6. Función de densidad de probabilidad y excedencia de umbrales como índices biológico-pesqueros de la anchoveta

La función de densidad de probabilidad revela patrones estacionales en la forma de la distribución de los datos de la densidad acústica (s_A) de los cardúmenes de anchoveta, lo cual es consistente con los procesos que modulan el calentamiento o enfriamiento del ecosistema marino, que amplían o reducen el hábitat disponible y propician cambios en los patrones agregativos de la especie.

En verano, debido a las condiciones cálidas la anchoveta se encuentra confinada principalmente dentro de las 60 mn de la costa. La pdf refleja el comportamiento de la especie: presenta la menor masa de valores en el extremo inferior y se desplaza más hacia la derecha (con el pico de la curva aproximadamente en 500 m^2/mn^2), lo cual indica la presencia de colas más pesadas y una mayor probabilidad de observar densidades extremadamente altas, asociada al mayor valor de μ, moda de $\log(s_A) = 8.16$, y la mayor variabilidad ($\sigma = 2.33$). En la parametrización logSEP4, $\nu = 3.75$ y $\tau = 2.77$ gobiernan la cola inferior y superior, respectivamente; el menor valor de τ frente a ν indica que la cola superior decae más lentamente que la inferior.

En otoño, el enfriamiento del ecosistema y la ampliación de las aguas costeras frías promueven la migración de los cardúmenes hacia zonas más alejadas de la costa, disgregando las agregaciones densas y dispersando el recurso. La pdf refleja este comportamiento: desplaza el pico hacia densidades más bajas, aproximadamente 300 m^2/mn^2 respecto del verano, y concentra mayor masa en el extremo inferior, con una cola menos extendida, asociada a un valor intermedio de μ, moda de $\log(s_A) = 7.07$, y menor variabilidad ($\sigma = 1.95$). En logSEP3, $\nu = 1.39$ controla la asimetría y $\tau = 3.71$ es la potencia común a ambas colas. Esto evidencia agregaciones más dispersas y una distribución menos

desplazada hacia densidades extremas que en verano, con el predominio de densidades intermedias-bajas.

En invierno, con el enfriamiento máximo del ecosistema, la anchoveta se dispersa y desplaza hacia zonas más alejadas de la costa por fuera de las 100 mn. La pdf, respecto de otoño, concentra mayor masa en el extremo inferior y un pico más estrecho aproximadamente entre 200–250 m^2/mn^2, con μ, moda de $\log(s_A) = 7.02$, y una cola superior más ligera. En logSEP4, $\nu = 4.09$ y $\tau = 2.12$ corresponden a la cola inferior y superior; la diferencia entre ambos es la mayor de las dos estaciones ajustadas con esta familia. Esto indica un predominio de agregaciones con densidades bajas-intermedias, sin la alta probabilidad de observar las densidades extremas del verano.

En primavera, cuando se inicia la transición de condiciones frías a cálidas con la intromisión de las Aguas Subtropicales Superficiales (ASS), las agregaciones de anchoveta se encuentran ampliamente distribuidas hasta las 100 mn. La pdf evidencia el pico más alto y agudo (aproximadamente en 150 m^2/mn^2) de las cuatro estaciones, con una densidad en la moda cerca de cuatro veces mayor que en verano y casi el doble que en invierno, y el menor valor de μ, moda de $\log(s_A) = 6.23$. En logSEP3, $\nu = 1.34$ controla la asimetría y $\tau = 3.68$ es la potencia común a ambas colas. Esto evidencia la mayor concentración de masa en densidades bajas y una caída abrupta hacia densidades altas, lo que indica el predominio de agregaciones de baja densidad.

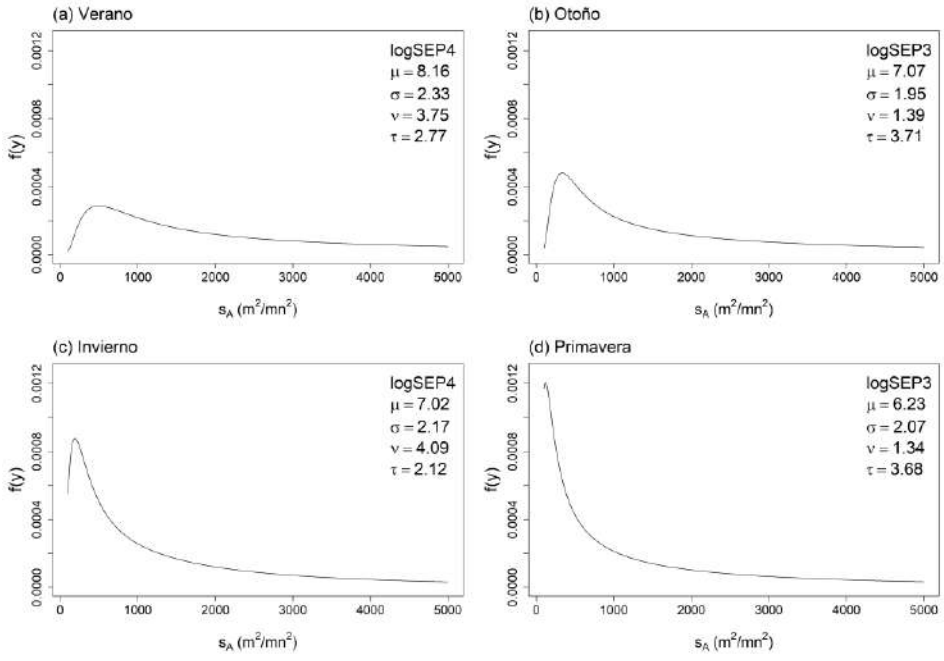
Por tanto, la variación estacional del s_A no se expresa solo como un cambio en la magnitud característica del parámetro μ, sino también como cambios simultáneos en escala (σ) y en los parámetros de forma (ν, τ), que en conjunto desplazan el pico de la distribución y la masa de probabilidad entre densidades bajas y altas según la estación, en respuesta a la variabilidad climática del ecosistema marino.

Cabe mencionar que los valores de ν y τ no son comparables únicamente entre las mismas familias. En logSEP4, ν y τ son las potencias exponenciales de la cola inferior y superior, respectivamente. En logSEP3, ν controla la asimetría y τ es la potencia común a ambas colas.

La pdf ajustada permite definir la forma completa de la distribución del s_A y convertir los parámetros μ, σ, ν y τ en índices biológico-pesqueros del comportamiento agregativo, cambios en la dispersión, predominio de densidades bajas y probabilidad de ocurrencia de agregaciones extremas. Un modelo centrado solo en la media no entregaría ninguno de estos índices ni los contrastes estacionales que revelan (Figura 24).

Figura 24

Función de densidad de probabilidad (pdf) de los modelos GAMLSS para el s_A de los cardúmenes de anchoveta por estación.



Nota. Elaboración propia.

La probabilidad de excedencia muestra diferencias estacionales en la ocurrencia de valores altos de s_A. Verano presentó las mayores probabilidades de superar umbrales elevados, con P(Y>2000 m²/mn²) = 55.5% y P(Y>5000 m²/mn²) = 32.1%, confirmando una mayor probabilidad de cardúmenes con alta densidad acústica. Otoño mostró probabilidades altas en umbrales bajos y medios, pero disminuyó a 28.7% para Y>5000 m²/mn², evidenciando una reducción progresiva del extremo superior. Primavera presentó P(Y>5000 m²/mn²) = 30.6%, superior a otoño pese a tener un valor menor del parámetro de localización (μ = 6.23 frente a 7.07), lo que indica que un desplazamiento hacia densidades bajas no implica un extremo superior más ligero. Invierno presentó las menores probabilidades de excedencia en todos los umbrales, con P(Y>5000 m²/mn²) = 23.4%, mostrando la menor ocurrencia de agregaciones acústicas extremas.

La excedencia muestral, próxima a la estimada en todos los umbrales y estaciones, confirma que el modelo reproduce adecuadamente la ocurrencia observada de densidades altas (Tabla 8).

Tabla 8

Probabilidad de excedencia ajustada y muestral de la densidad acústica (s_A) a umbrales fijos (250, 500, 1000, 2000 y 5000 m²/mn²*), por estación*

Estación	Distribución	Excedencia	P(Y>250)	P(Y>500)	P(Y>1000)	P(Y>2000)	P(Y>5000)
Verano	logSEP4	Ajustada	0.9525	0.8667	0.7279	0.5546	0.3214
		Muestral	0.9489	0.8529	0.7102	0.5287	0.3268
Otoño	logSEP3	Ajustada	0.9385	0.8274	0.6748	0.5040	0.2874
		Muestral	0.9330	0.8193	0.6648	0.4950	0.2873
Invierno	logSEP4	Ajustada	0.9062	0.7731	0.6054	0.4316	0.2344
		Muestral	0.8936	0.7583	0.5930	0.4290	0.2452
Primavera	logSEP3	Ajustada	0.9192	0.8086	0.6684	0.5119	0.3060
		Muestral	0.9066	0.7871	0.6466	0.4964	0.3089

Nota. Los valores de excedencia ajustada, $P(Y > y) = 1 - F(y)$ y muestral indican que el modelo reproduce la estructura del s_A. Elaboración propia.

El contraste entre primavera y otoño evidencia que la variación estacional del s_A no se explica únicamente por el parámetro de localización (μ), sino por la forma completa de la

distribución: el ordenamiento de las estaciones según μ no coincide con su ordenamiento según la probabilidad de agregaciones extremas. Esto permite usar la excedencia no solo para estimar la probabilidad de superar umbrales de detección, sino también como índice directo de ocurrencia de cardúmenes de alta densidad acústica.

4.1.7. Predicción y desempeño predictivo de los modelos GAMLSS, BAMLSS, NN-GAMLSS y RF-GAMLSS para el s_A de los cardúmenes de anchoveta

4.1.7.1 Predicción de la moda condicional

Los GLM, GAM, INLA y VAST no se emplearon en la predicción de la media condicional, ya que las gráficas de residuales demuestran que presentan patrones sistemáticos (Dunn y Smyth, 1996; van Buuren y Fredriks, 2001; Rigby y Stasinopoulos, 2005, 2007).

Los valores predichos por los modelos distribucionales corresponden a la moda condicional estimada del s_A, expresada como $\exp(\mu)$. En la parametrización de las familias logSEP3 y logSEP4, el parámetro μ representa el logaritmo de la moda, por lo que no se interpreta directamente en la escala original del s_A. Por ello, se aplicó la transformación exponencial para obtener la moda estimada en unidades de densidad acústica.

La moda condicional estimada $exp(\mu)$ presentó diferencias entre los modelos distribucionales según la estación. En verano, las modas medias fueron similares entre GAMLSS, BAMLSS y NN-GAMLSS (3160–3310 m^2/mn^2), con RF-GAMLSS ligeramente por debajo (2993.5 m^2/mn^2). En otoño, RF-GAMLSS presentó la moda más baja (1316.6 m^2/mn^2), mientras que los demás modelos se mantuvieron entre 2038 y 2308 m^2/mn^2. En invierno, NN-GAMLSS presentó la moda más alta (3275.5 m^2/mn^2) y RF-GAMLSS la más baja (1895.6 m^2/mn^2), con GAMLSS y BAMLSS en valores intermedios cercanos entre sí. En primavera, NN-GAMLSS presentó la mayor moda media (2626.2 m^2/mn^2) y RF-GAMLSS la menor (1616.3 m^2/mn^2) (Tabla 9).

Tabla 9

Moda condicional estimada $exp(\mu)$ del s_A de los cardúmenes de anchoveta por modelo distribucional y estación

Modelo	Verano Moda + DE	Otoño Moda + DE	Invierno Moda + DE	Primavera Moda + DE
GAMLSS	3310.2 ± 3089.0	2112.1 ± 4217.7	2257.1 ± 2545.4	2366.3 ± 4707.5
BAMLSS	3245.3 ± 2870.1	2037.7 ± 4482.0	2249.8 ± 2473.0	2216.2 ± 4372.8
NN-GAMLSS	3160.2 ± 2706.6	2308.1 ± 2972.8	3275.5 ± 5099.4	2626.2 ± 3510.6
RF-GAMLSS	2993.5 ± 2150.0	1316.6 ± 782.9	1895.6 ± 1408.0	1616.3 ± 1255.7

Nota**.** Valores expresados en m^2/mn^2, como media ± desviación estándar de $exp(\mu)$ sobre el conjunto de validación. Elaboración propia.

4.1.7.2 Media teórica de los modelos

En las familias logSEP3 y logSEP4, el parámetro μ es la moda de $\log(s_A)$. Por ello, $\exp(\mu)$ no corresponde a la media del s_A en la escala original y para comparar las predicciones de los modelos con el promedio observado, se estimó la media teórica de la distribución predictiva de cada modelo a partir de los parámetros μ, σ, ν y τ.

Todos los modelos generaron predicciones adecuadas a los valores observados del s_A, con diferencias entre estaciones en la aproximación de la media.

En verano, todos los modelos presentaron valores medios inferiores al promedio observado en los datos de validación (7438.4 m²/mn²). BAMLSS mostró la media más cercana (7010.0 m²/mn²), seguido por NN-GAMLSS (6925.3 m²/mn²), GAMLSS (6830.0 m²/mn²) y RF-GAMLSS (6324.7 m²/mn²). Durante otoño, los modelos también presentaron medias inferiores al valor observado en validación (7138.1 m²/mn²). BAMLSS fue el más cercano (6394.1 m²/mn²), seguido por GAMLSS (6385.5 m²/mn²), NN-GAMLSS (6339.1 m²/mn²) y RF-GAMLSS (6270.9 m²/mn²). En invierno, los modelos mostraron la mayor coherencia con la media observada (5424.2 m²/mn²). GAMLSS (5564.2 m²/mn²) y BAMLSS (5518.1 m²/mn²) generaron los valores más próximos al promedio de validación, seguidos por NN-GAMLSS (5178.6 m²/mn²) y RF-GAMLSS (4773.0 m²/mn²). En

primavera, NN-GAMLSS presentó la media más alta (7600.0 m²/mn²), por encima del valor observado en validación (7485.6 m²/mn²). Los demás modelos presentaron medias menores: BAMLSS (6674.5 m²/mn²), GAMLSS (6639.0 m²/mn²) y RF-GAMLSS (6635.1 m²/mn²).

En términos de variabilidad, la desviación estándar (DE) de las predicciones de la media teórica no es comparable con la de los datos de validación: la primera mide la variabilidad de las predicciones y la segunda la del proceso estocástico. La DE fue similar para todos los modelos en verano y otoño, con mayor dispersión en invierno y primavera (Tabla 10).

Tabla 10

Media y desviación estándar (DE) de los valores del conjunto de validación y la media teórica estimada por modelo distribucional y estación.

Modelo	Verano	Otoño	Invierno	Primavera
Datos de prueba	7438.4	7138.1	5424.2	7485.6
GAMLSS	6830.0	6385.5	5564.2	6639.0
BAMLSS	7010.0	6394.1	5518.1	6674.5
NN-GAMLSS	6925.3	6339.1	5178.6	7600.0
RF-GAMLSS	6324.7	6270.9	4773.0	6635.1

Nota. Valores expresados en m^2/mn^2, como media ± desviación estándar sobre el conjunto de validación. Elaboración propia.

4.1.7.3 Desempeño predictivo

En verano, GAMLSS presentó los menores valores puntuales de RMSE (12021.0 m^2/mn^2), MAE (6666.8 m^2/mn^2) y CRPS (4491.2 m^2/mn^2), aunque sus intervalos bootstrap se superpusieron ampliamente con los de BAMLSS, NN-GAMLSS y RF-GAMLSS en las tres métricas, por lo que esta ventaja numérica no es estadísticamente concluyente. Los cuatro modelos mostraron un desempeño muy similar en esta estación, con RMSE entre 12021.0 m^2/mn^2 y 12639.0 m^2/mn^2.

En otoño, NN-GAMLSS presentó el menor RMSE (11997.1 m^2/mn^2), MAE (6386.9 m^2/mn^2) y CRPS (4364.6 m^2/mn^2). Las diferencias entre los cuatro modelos fueron reducidas y sus intervalos bootstrap se superpusieron en las tres métricas, por lo que ninguna ventaja es concluyente (RMSE entre 11997.1 m^2/mn^2 y 12179.6 m^2/mn^2).

En invierno, ningún modelo dominó las tres métricas: BAMLSS presentó el menor RMSE (10426.5 m^2/mn^2), RF-GAMLSS el menor MAE (5275.6 m^2/mn^2), y GAMLSS el menor CRPS (3489.7 m^2/mn^2), seguido muy de cerca por BAMLSS (3522.6 m^2/mn^2) y RF-GAMLSS (3596.6 m^2/mn^2). Los intervalos de CRPS entre estos tres modelos se superpusieron ampliamente, indicando un desempeño probabilístico muy similar. NN-GAMLSS presentó los valores más altos en las tres métricas (RMSE = 13646.9 m^2/mn^2; MAE = 7347.6 m^2/mn^2; CRPS = 5290.8 m^2/mn^2), con intervalos que no se superpusieron con los de los demás modelos en MAE y CRPS.

En primavera, RF-GAMLSS presentó el menor RMSE (13181.9 m^2/mn^2) y GAMLSS el menor MAE (6838.1 m^2/mn^2) y CRPS (4734.6 m^2/mn^2), con intervalos superpuestos entre GAMLSS, BAMLSS y RF-GAMLSS. NN-GAMLSS mostró los valores más altos en RMSE (15279.7 m^2/mn^2) y MAE (7595.6 m^2/mn^2), con el intervalo de RMSE parcialmente separado del de los demás modelos.

GAMLSS y BAMLSS fueron los más estables, ubicándose consistentemente en el grupo de menor error, con intervalos bootstrap superpuestos entre sí y sin diferencias concluyentes. RF-GAMLSS mostró un desempeño competitivo, con los menores valores de MAE en invierno y de RMSE en primavera, aunque presentó los mayores valores de CRPS en verano (4757.0 m^2/mn^2) y otoño (4522.8 m^2/mn^2). NN-GAMLSS fue el menos estable, con los errores más altos en invierno y primavera y con intervalos que en algunos casos se separaron de los demás modelos.

Estos resultados confirman que la capacidad predictiva varía estacionalmente y que la evaluación debe considerar tanto los valores puntuales como los intervalos bootstrap al 95%, dado que el s_A de los cardúmenes de anchoveta presenta fuerte asimetría, heterogeneidad y valores extremos (Tabla 11).

Tabla 11

Métricas de desempeño predictivo de los modelos para el s_A de los cardúmenes de anchoveta, por estación, con intervalos Bootstrap al 95%.

Estación	Modelo	RMSE	IC95% RMSE	MAE	IC95% MAE	CRPS	IC95% CRPS
Verano	GAMLSS	12021.0	[11178.1–12793.3]	6666.8	[6333.9–7001.0]	4491.2	[4192.6–4775.4]
	BAMLSS	12178.6	[11353.4–12963.8]	6820.6	[6468.8–7178.0]	4559.4	[4255.6–4867.3]
	NN-GAMLSS	12251.5	[11482.2–12991.3]	6848.7	[6504.0–7182.5]	4587.9	[4290.7–4891.5]
	RF-GAMLSS	12639.0	[11794.5–13462.8]	6847.7	[6500.7–7225.4]	4757.0	[4437.9–5107.9]
Otoño	GAMLSS	12056.1	[11086.5–13015.5]	6481.2	[6078.4–6904.5]	4378.9	[4033.7–4752.2]
	BAMLSS	12179.6	[11180.6–13170.6]	6576.0	[6171.2–7012.4]	4426.3	[4060.5–4797.6]
	NN-GAMLSS	11997.1	[11049.5–12998.8]	6386.9	[5964.8–6823.3]	4364.6	[4003.6–4747.8]
	RF-GAMLSS	12106.6	[11102.1–13068.0]	6465.5	[6030.6–6880.0]	4522.8	[4146.1–4923.6]
Invierno	GAMLSS	11942.5	[9821.5–14812.9]	5517.8	[5192.3–5900.3]	3489.7	[3242.3–3771.4]
	BAMLSS	10426.5	[9634.4–11251.3]	5487.4	[5193.4–5788.4]	3522.6	[3264.6–3792.3]
	NN-GAMLSS	13646.9	[12795.7–14498.2]	7347.6	[6949.9–7747.9]	5290.8	[4941.8–5640.5]
	RF-GAMLSS	10577.0	[9691.2–11547.3]	5275.6	[4963.3–5605.6]	3596.6	[3311.1–3911.8]
Primavera	GAMLSS	13394.5	[11896.8–14696.0]	6838.1	[6244.2–7401.7]	4734.6	[4217.0–5233.5]
	BAMLSS	13806.0	[12345.2–15191.1]	7207.1	[6624.5–7777.5]	4876.2	[4344.7–5405.9]
	NN-GAMLSS	15279.7	[13428.3–17245.5]	7595.6	[6908.4–8290.7]	4907.8	[4398.5–5466.5]
	RF-GAMLSS	13181.9	[11758.9–14505.8]	6888.2	[6322.9–7455.6]	4787.7	[4281.6–5298.3]

Nota. RMSE = error cuadrático medio; MAE = error absoluto medio; CRPS = puntaje probabilístico continuo, todos expresados en m^2/mn^2. IC95% indican los límites inferior y superior del intervalo Bootstrap no paramétrico al 95%. Elaboración propia.

4.2 Contrastación de hipótesis

4.2.1. Hipótesis general

Los resultados demuestran que la regresión clásica centrada en la media no es equivalente a la regresión distribucional al modelar la densidad acústica de los cardúmenes de anchoveta. Se rechaza la hipótesis nula general y se acepta la hipótesis alternativa.

4.2.2. Hipótesis específica 1

Los resultados demuestran que los modelos clásicos centrados en la media no consiguieron representar adecuadamente el s_A ni en la escala logarítmica ni en la escala original, en ninguna estación. Se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

4.2.3. Hipótesis específica 2

Las variables explicativas presentaron efectos diferenciados en cada estación sobre la densidad típica ($Q_{0.5}$) y sobre las densidades extremas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta. Se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

4.2.4. Hipótesis específica 3

La regresión clásica presentó patrones sistemáticos en los residuos cuantílicos que evidencian modelos incorrectos. En contraste, en la regresión distribucional no se observaron patrones sistemáticos relevantes en los residuos cuantílicos que evidencien mala especificación. El desempeño predictivo fue casi indistinguible entre los modelos distribucionales. Se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

4.2.5. Hipótesis específica 4

Los modelos distribucionales permitieron modelar la densidad de los cardúmenes de anchoveta, obtener índices biológico-pesqueros y estimar probabilidades de excedencia relevantes para la gestión pesquera, dando uso cuantitativo a las ecosondas comerciales sin calibración masiva de la flota. Son aplicables como base metodológica de un servicio pesquero predictivo de s_A a escala de cardumen. Se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

CAPÍTULO V.

5. DISCUSIÓN

5.1. Discusión de resultados

5.1.1. Inconsistencia de la validación basada en comparar la variable observada Y con el parámetro de localización

En la estadística clásica, el empleo de métricas como RMSE y MAE como criterios de validación de un modelo se fundamenta en la comparación entre los valores observados (Y) y el parámetro de localización (μ). Dado que, en ningún modelo paramétrico o no paramétrico, se cumple la igualdad $Y = \mu$, el parámetro μ actúa como parámetro de localización de la distribución condicional $f(Y \mid X)$ y, según la parametrización de la familia asumida, puede ser la media, la mediana o la moda (Gneiting, 2011), funcionales que no son comparables directamente con los valores observados.

Además, el uso de dichas métricas implica asumir que caracterizar únicamente μ es suficiente para describir Y, lo cual resulta incorrecto en el caso de variables aleatorias continuas con asimetría y curtosis, donde la distribución también depende de parámetros de escala y forma, que deben ser modelados simultáneamente para representar adecuadamente el proceso (Rigby y Stasinopoulos, 2005; Gneiting y Raftery, 2007).

Esta limitación es particularmente importante porque RMSE y MAE son métricas de predicción puntual: resumen el error respecto a una localización estimada, pero no evalúan si el modelo captura adecuadamente todas las características de la distribución de la respuesta. En datos altamente agregados, como el s_A, un modelo puede presentar bajo error promedio y, al mismo tiempo, subestimar sistemáticamente los valores extremos o comprimir la variabilidad real del proceso. Por ello, la validación del modelo no debe basarse en los errores entre Y y μ, sino en herramientas avanzadas de diagnóstico que permitan comprobar la adecuación de la distribución asumida para la respuesta.

Durante las últimas tres décadas, la ciencia pesquera ha asumido que un modelo queda validado mediante métricas de error como RMSE y MAE, o mediante criterios de ajuste relativo como AIC, BIC, DIC o WAIC. Esta práctica se extiende a la densidad acústica (Murase et al., 2009; Quiroz et al., 2014; Yuffra, 2017; Castillo et al., 2018; Zhu et al., 2021; Blanluet et al., 2025), al CPUE (Barriga, 2017; Zhou et al., 2019; Pasache et al., 2023; Xu et al., 2025), a la biomasa (Roa-Ureta y Niklitschek, 2007; Fang et al., 2019; Thorson, 2019; Mesquita et al., 2021; Díaz-Astudillo et al., 2022; Citores et al., 2024) y a la distribución espacial (Swartzman et al., 2008; Alvarez-Trasviña et al., 2022; Cornejo et al., 2022), donde los estudios reportaron selección de modelos por criterios de información o partición de datos, omitiendo validar los supuestos distribucionales, lo que para Freckleton (2009) es uno de los siete pecados capitales del análisis estadístico comparativo.

5.1.2. Inconsistencia de pretransformar Y, la constante arbitraria +1 y el enlace logarítmico

La práctica extendida de pretransformar los datos pesqueros con presencia de ceros mediante transformaciones tipo logaritmo, como $log(Y + 1)$, presenta dos problemas independientes que comprometen su validez. El primero afecta la interpretación de los coeficientes. Chen & Roth (2024) demuestran que los efectos estimados sobre transformaciones log-like definidas en cero no son invariantes a la escala, de modo que su magnitud e incluso su comportamiento inferencial, puede depender de la unidad física usada para medir la respuesta, por ejemplo, kilogramos o toneladas. Esta dependencia aparece porque el cambio de cero a valores positivos corresponde al margen extensivo (el paso de ausencia a ocurrencia del fenómeno), donde un efecto porcentual individual no está matemáticamente definido. En consecuencia, la constante añadida en $log(Y + 1)$ no es un ajuste inocuo: asigna implícitamente un peso al paso de ausencia a ocurrencia y puede hacer

que el efecto estimado refleje más la escala de medición que una relación biológica interpretable.

El segundo problema, independiente del anterior, es un sesgo sistemático de retransformación debido a la desigualdad de Jensen. Por la naturaleza cóncava de la función logarítmica, el valor esperado del logaritmo es menor o igual que el logaritmo del valor esperado, $E[log(Y)] \leq log(E[Y])$, con igualdad solo cuando la varianza es nula; en presencia de variabilidad, la brecha se incrementa al aumentar la varianza (Mullahy, 1998; Manning & Mullahy, 2001). Por ello, un modelo ajustado sobre $log(Y)$ opera sobre $E[log(Y)]$, que corresponde al logaritmo de la media geométrica de Y y no de su media aritmética. Al retornar a la escala original, este desfase introduce una subestimación sistemática que no puede corregirse de forma insesgada cuando existe heterocedasticidad. Santos Silva y Tenreyro (2006) demuestran, en esta línea, que la estimación de modelos log-linealizados produce parámetros sesgados bajo heterocedasticidad, por lo que recomiendan estimar directamente el modelo multiplicativo en la escala natural de la respuesta.

Esta distinción es central porque, en los Modelos Lineales Generalizados, la función de enlace transforma la media esperada de la respuesta, no los valores individuales observados (McCullagh y Nelder, 1989). Bajo un enlace logarítmico, el modelo se formula como $g(\mu_i) = \eta_i$, con $\mu_i = E(Y_i)$, lo que implica modelar $log(E[Y])$ y no $E[log(Y)]$. La relación logarítmica opera así sobre la media condicional del proceso, lo que elimina la necesidad teórica de alterar la variable dependiente mediante transformaciones ad hoc como $log(Y + 1)$. La misma lógica se extiende a los GAM, donde el predictor incorpora funciones suaves pero el enlace continúa operando sobre la media condicional (Hastie & Tibshirani, 1986; Wood, 2017), y a INLA, donde el predictor latente se vincula con un parámetro de la distribución de la respuesta mediante una función de enlace, por lo que es innecesario pretransformar los datos observados (Rue et al., 2009; Blangiardo & Cameletti, 2015).

El problema se agrava al anidar ambos elementos. La práctica convencional en la literatura oceanográfica y de evaluación de stocks ha ajustado modelos GLM/GAM centrados en la media aplicando $log(Y + 1)$ sobre los valores positivos modelados con una distribución Gaussiana/Gamma y enlace logarítmico. Al imponer un vínculo logarítmico sobre una respuesta pretransformada al logaritmo, se duplica la función: el modelo no representa ni $log(E[Y])$ ni $E[Y]$, sino una operación sin interpretación coherente que rompe la correspondencia entre la verosimilitud optimizada, la escala física de la respuesta y el objeto inferencial. Además, la pretransformación y la constante +1 introducida por los ceros deforman la superficie de verosimilitud, corrompen las rutinas de optimización del gradiente y anulan la coherencia teórica del marco (Mullahy, 1998; Manning y Mullahy, 2001; Santos Silva y Tenreyro, 2006).

Además del sesgo de retransformación, bajo heterocedasticidad, la magnitud y el orden relativo de los efectos estimados en la escala logarítmica no corresponden a los efectos reales sobre la respuesta en su escala original (Manning, 1998). En aplicaciones pesqueras, esto distorsiona la identificación de los determinantes ambientales del recurso.

Frente a estas limitaciones, el modelamiento distribucional condicional en la escala original mediante familias de cuatro parámetros como logSEP3 y logSEP4 resuelve el problema: independiza el predictor de localización (μ) de los parámetros de escala (σ) y forma (ν, τ), estima la verosimilitud sin alterar la escala del proceso observado, y preserva la integridad de las colas pesadas, generando residuos cuantílicos normales e independientes.

5.1.3. Distribución apropiada para la densidad de los cardúmenes de anchoveta

El análisis exploratorio muestra que la densidad acústica (s_A) de los cardúmenes de anchoveta presenta una marcada asimetría positiva y la presencia recurrente de valores extremos, lo que se traduce en una cola pesada hacia la derecha. Esta estructura caracterizada por alta variabilidad, predominio de valores bajos y ocurrencia esporádica de eventos de

gran magnitud, es típica de procesos de agregación biológica y exige el uso de distribuciones flexibles capaces de representar adecuadamente colas largas, asimetría y alta curtosis (Rigby & Stasinopoulos, 2005; Stasinopoulos et al., 2017; Rigby et al., 2019; Stasinopoulos et al., 2024).

En este contexto, las distribuciones continuas estándar evaluadas, tanto con soporte en la línea real $(-\infty, +\infty)$ como con soporte en la línea real positiva $(0, +\infty)$, resultaron insuficientes para representar el $s_A>0$ de los cardúmenes de anchoveta. La asimetría extrema y las colas superpesadas requieren familias logarítmicas más flexibles, como logSEP3 y logSEP4, adaptadas de las distribuciones de Potencia Exponencial Sesgada tipo 3 y tipo 4 (Fernandez et al., 1995; Rigby & Stasinopoulos, 2005; Rigby et al., 2019; Stasinopoulos et al., 2024).

La familia logSEP3 es adecuada cuando la asimetría puede representarse mediante una dilatación diferencial de la escala a ambos lados de la moda. En esta parametrización, la asimetría no surge de un desplazamiento en la localización, sino de una modificación de la forma de la distribución alrededor de la moda. Esto permite describir distribuciones con predominio de valores bajos y una cola derecha extendida, patrón consistente con el comportamiento observado en la densidad acústica de los cardúmenes de anchoveta (Fernandez et al., 1995; Rigby et al., 2019; Stasinopoulos et al., 2024).

La familia logSEP4 proporciona una flexibilidad adicional para representar cambios más complejos en la forma de la distribución. Su uso resulta relevante cuando el s_A presenta no solo asimetría marcada, sino también colas derechas muy pesadas y variaciones importantes en la curtosis. En términos prácticos, logSEP4 permite capturar diferencias en la concentración central y en el peso relativo de los extremos, lo que es fundamental para modelar procesos acústicos dominados por agregaciones intensas y valores extremos (Rigby et al., 2019; Stasinopoulos et al., 2024).

Ambas familias comparten la capacidad de representar distribuciones unimodales fuertemente asimétricas a la derecha, lo que las hace apropiadas para la densidad acústica de anchoveta (s_A), caracterizado por un predominio de densidades bajas y eventos de agregación extrema en la cola derecha. La elección entre ellas depende de si la asimetría del s_A se ajusta mejor mediante una escala diferenciada (logSEP3) o mediante una pesadez de cola diferenciada (logSEP4).

Un aspecto importante de ambas familias es que el parámetro μ no representa la media del s_A, sino el logaritmo de la moda condicional. Por tanto, $\exp(\mu)$ debe interpretarse como la moda estimada en la escala original, es decir, como el valor más probable de la densidad acústica bajo las condiciones morfológicas y espacio-temporales consideradas. Esta propiedad explica por qué la comparación directa entre μ y los valores observados de s_A no es válida, y justifica la estimación de la media teórica de la distribución predictiva cuando se requiere comparar el nivel promedio predicho con los datos de validación (Rigby & Stasinopoulos, 2005; Rigby et al., 2019; Stasinopoulos et al., 2024).

En conjunto, la selección de logSEP3 y logSEP4 demuestra que la densidad acústica de los cardúmenes de anchoveta requiere familias distribucionales capaces de representar colas pesadas, asimetría marcada y alta curtosis en la escala original. Esto refuerza la limitación de los enfoques centrados en la media y de las transformaciones logarítmicas simples, ya que la estructura del s_A no se reduce a un problema de localización, sino que involucra cambios simultáneos en variabilidad, asimetría y comportamiento extremo (Rigby & Stasinopoulos, 2005; Stasinopoulos et al., 2017; Rigby et al., 2019; Stasinopoulos et al., 2024).

5.1.4. Comparabilidad de desempeño predictivo entre modelos de regresión distribucional e incertidumbre predictiva

La evaluación del desempeño de los modelos mediante estimaciones puntuales de RMSE, MAE y CRPS resulta incompleta si no se interpreta frente a la alta variabilidad, asimetría y presencia de valores extremos de la densidad acústica de la anchoveta. El análisis de los intervalos bootstrap al 95% trasciende la simple comparación de errores puntuales y revela una implicancia estadística fundamental: no existe un único enfoque que mantenga una superioridad predictiva absoluta y transversal en todos los escenarios estacionales.

Si bien las métricas puntuales sugieren ciertas ventajas estacionales (como el mejor desempeño numérico de GAMLSS en verano o de NN-GAMLSS en otoño), la superposición de los intervalos bootstrap indica que GAMLSS y BAMLSS presentan desempeños predictivos comparables dentro de la incertidumbre estimada. Esta superposición muestra que las diferencias numéricas entre ambos modelos no constituyen una superioridad algorítmica concluyente, sino variaciones dependientes de la estación, de la métrica evaluada y de la propia variabilidad de los datos acústicos. Los modelos basados en splines penalizados lograron capturar la heterogeneidad del s_A de manera robusta y estable en las cuatro estaciones, respaldando su pertinencia como base metodológica para un servicio pesquero predictivo.

Los algoritmos de aprendizaje automático mostraron mayor dificultad para ajustar los valores extremos de la cola pesada derecha del s_A, aunque por razones distintas. Los RF-GAMLSS predicen mediante el promedio de las observaciones dentro de cada partición del espacio de covariables; en la cola superior, donde las agregaciones densas son escasas, cada partición contiene pocas observaciones y el modelo no logra extrapolar más allá del rango observado, lo que produce desajustes en el extremo superior de los diagramas de gusano. Por su parte, NN-GAMLSS depende de la inicialización aleatoria de los pesos y de la

optimización de una superficie de error no convexa; en las regiones con pocos datos, como las colas pesadas del s_A, esta optimización resulta inestable y sensible a las condiciones iniciales, lo que se reflejó en intervalos bootstrap más amplios y errores más altos, particularmente en invierno y primavera, con una menor coherencia probabilística de sus predicciones, especialmente en el CRPS. En contraste, los modelos basados en splines penalizados (GAMLSS, BAMLSS) modelan directamente los parámetros de forma (ν y τ) de la distribución condicional, lo que les permite representar la asimetría y la cola pesada del s_A de manera continua y estable, sin depender de la densidad local de observaciones ni de la inicialización del ajuste.

El caso de RF-GAMLSS ilustra por qué las métricas puntuales no bastan para evaluar un modelo. En invierno y primavera, RF-GAMLSS alcanzó valores competitivos de RMSE, MAE y CRPS (incluso los menores en algunas métricas); sin embargo, su diagnóstico residual evidenció desviaciones en la cola derecha del diagrama de gusano y desplazamiento respecto al centro del diagrama de cubo. Un modelo con estructura residual sistemática puede predecir razonablemente la magnitud y aun así estar mal especificado, lo cual resulta engañoso si solo se observan las métricas. Por ello, RF-GAMLSS se consideró adecuado únicamente en verano y otoño, y en invierno y primavera no reprodujo consistentemente la estructura completa del s_A. Este resultado demuestra empíricamente que las métricas puntuales pueden inducir a seleccionar un modelo mal especificado si no se acompaña del diagnóstico residual.

En síntesis, estos hallazgos advierten sobre los riesgos metodológicos de seleccionar modelos basándose exclusivamente en la minimización de errores puntuales. El caso de RF-GAMLSS, con métricas competitivas pero residuos sistemáticos, muestra que un buen valor de error puede enmascarar una mala especificación del modelo. Frente a variables ecológicas complejas, altamente asimétricas y agregadas, la selección metodológica debe

fundamentarse primero en la validez del diagnóstico residual (que determina si el modelo está correctamente especificado) y solo después en la precisión puntual, la amplitud de los intervalos bootstrap y la estabilidad estacional. La interpretación conjunta del diagnóstico distribucional y de las métricas de error, y no estas últimas de forma aislada, proporciona una base sólida para evaluar modelos predictivos aplicados al s_A de los cardúmenes de anchoveta.

5.1.5. Diagnóstico residual engañoso en los diagramas Q-Q

En los resultados de esta tesis, la aparente normalidad residual observada en los Q-Q plots de los modelos GLM-Gamma y GAM-Gamma constituye un diagnóstico engañoso. Aunque el Q-Q plot es una herramienta clásica, su dependencia de una perspectiva global provoca que los cuantiles centrales (que concentran la mayor masa de datos) dominen la escala visual y compriman los residuales hacia la línea de tendencia, enmascarando las deficiencias del modelo (Wilk y Gnanadesikan, 1968; Atkinson, 1985; Loy et al., 2016). Al promediar y diluir las discrepancias a lo largo de todo el dominio, el Q-Q plot genera una falsa percepción de bondad de ajuste y oculta los errores críticos en las colas de la distribución.

Para superar esta limitación de escala, los diagramas de gusano (van Buuren y Fredriks, 2001) eliminan la tendencia diagonal mediante un reescalado por diferencias, amplificando las desviaciones locales y transformándolas en patrones sistemáticos inequívocos del desajuste de un modelo. En consecuencia, los Q-Q plots no validan la distribución asumida; por el contrario, pueden encubrir fallas relevantes en un modelo.

5.1.6. Insuficiencia del Campo Latente y mallas espaciales

Todo modelo probabilístico debe demostrar que la distribución condicional asumida representa adecuadamente la variabilidad de los datos observados. Este requisito también se aplica a los modelos ajustados mediante la Aproximación de Laplace Integrada Anidada

(INLA), debido a que INLA es un procedimiento computacional para aproximar distribuciones posteriores bajo una función de verosimilitud previamente especificada, pero no determina si dicha verosimilitud representa correctamente el proceso generador de los datos (Rue et al., 2009). Así, la incorporación de un campo espacial latente mediante SPDE puede representar la dependencia espacial residual, pero no corregir una especificación distribucional inadecuada; la misma distinción se aplica a la malla espacial de VAST. Por tanto, valores favorables de DIC, WAIC o LPML, la convergencia numérica, la obtención de campos espaciales suavizados o la ausencia de autocorrelación espacial residual no demuestran adecuación distribucional. Esta debe verificarse mediante diagnósticos capaces de evaluar la distribución condicional completa, como los residuos cuantílicos y los diagramas de gusano, que permiten identificar desviaciones en localización, dispersión, asimetría y curtosis (Dunn & Smyth, 1996; van Buuren & Fredriks, 2001; Held et al., 2010).

Dado que los modelos INLA-Gamma son en estructura GAM bayesianos y el VAST-Gamma un GLM con campo latente, la comparación aísla el efecto del método de inferencia y de la estructura espacial: todos comparten la distribución Gamma centrada en la media condicional y todos exhiben los mismos patrones residuales sistemáticos. La mala especificación no proviene, por tanto, del paradigma inferencial ni de la ausencia de dependencia espacial, sino de la familia asumida para la respuesta.

La evidencia de este colapso inferencial se manifiesta en la transición de las métricas de resumen global al diagnóstico de los residuos del modelo INLA-Gamma con campo espacial latente continuo SPDE y VAST-Gamma con campo latente espacial de malla, que exhiben una distorsión sistemática en todas las estaciones del año (verano, otoño, invierno y primavera). En los diagramas de gusano, se observa un patrón parabólico en forma de "U" que evade masivamente las hipérbolas de confianza del 95%. Este comportamiento demuestra que, a pesar de la sofisticación de la malla espacial, el modelo presenta un sesgo

en la media y la dispersión, fallando en reproducir el comportamiento de las colas pesadas y subestimando de manera sistemática los picos críticos locales donde reside la señal biológica de las grandes agregaciones de anchoveta. Asimismo, los diagramas de cubo corroboran este desajuste al situar la totalidad de las observaciones fuera del elipsoide central, equivalente a la prueba de Jarque-Bera, lo que comprueba la presencia de asimetría y curtosis del momento, rechazando el supuesto de normalidad de los residuos cuantílicos y demostrando que la complejidad espacial es analíticamente engañosa si la distribución asumida es restrictiva.

En conjunto, los cuatro modelos centrados en la media recorren el espectro de complejidad estructural e inferencial disponible: del modelo lineal al aditivo, del ajuste frecuentista al bayesiano, y de la ausencia de estructura espacial al campo latente continuo y al campo sobre malla. Ninguna de esas variaciones modificó la distribución asumida para la respuesta, y ninguna corrigió los patrones residuales sistemáticos.

5.1.7. Limitaciones de la regresión clásica centrada en la media frente a la regresión distribucional

Estos resultados demuestran que las distribuciones SEP3 y SEP4 adaptadas a exponenciales (logSEP3 y logSEP4) representan con mayor fidelidad la variabilidad, asimetría y colas superpesadas de los valores originales de s_A. Al modelar simultáneamente la localización, la escala y la forma es posible interpretar directamente la densidad acústica en escala original, sin depender de pretransformaciones que modifican la lectura estadística y ecológica de los cardúmenes densos. Además, los modelos clásicos centrados en la media condicional son insuficientes para modelar la complejidad de la densidad acústica de anchoveta: los residuos presentan patrones sistemáticos en forma de S invertida y cuadráticos que evidencian modelos incorrectos.

Los antecedentes previos evidencian, en los diagramas Q-Q, patrones residuales sistemáticos. Barriga (2017) presenta una distribución ajustada demasiado sesgada hacia la izquierda en el logaritmo de la captura por unidad de esfuerzo (CPUE) de merluza peruana (*Merluccius gayi peruanus*); Yuffra (2017) presenta colas demasiado ligeras en modelos gaussianos aplicados al logaritmo de la densidad acústica de anchoveta peruana (*Engraulis ringens*), múnida (*Grimothea monodon*) y macrozooplancton fluid-like; Alegre (2011) exhibe una distribución ajustada demasiado sesgada hacia la derecha en modelos Tweedie aplicados al índice de llenura estomacal e índice de diversidad de Margalef para la pota (*Dosidicus gigas*). En la misma línea, modelos Tweedie aplicados por Wood (2020) a conteos de huevos de jurel (*Trachurus sp.*), Lima et al. (2023) a la CPUE de pez espada (*Xiphias gladius*) y Blanluet et al. (2025) a la biomasa acústica de atunes oceánicos (*Thunnus spp.*) muestran una distribución ajustada demasiado sesgada hacia la derecha, compatible con residuos positivos grandes y subestimación de valores altos.

En el caso de la distribución espacial de la densidad acústica (s_A) de anchoveta peruana, los datos presentan una estructura semicontinua, asimetría positiva extrema, colas pesadas y más del 50 % de valores iguales a cero. Las transformaciones logarítmicas no consiguen normalidad; por el contrario, reducen la variabilidad extrema y enmascaran señales ecológicas críticas asociadas a las áreas de alta agregación de la especie (Flores et al., 2026).

La evidencia diagnóstica confirma que los GAM ajustados bajo distribuciones gaussianas y Tweedie sobre datos transformados exhiben residuos cuantílicos normalizados con desviaciones sistemáticas, curvaturas en ambas colas de los Q-Q plots y patrones fuera de las bandas de confianza del 95 % en los diagramas de gusano, lo que evidencia un modelo incorrecto (Stasinopoulos et al., 2024; Flores et al., 2026).

La misma limitación se extiende a los modelos INLA y VAST, al combinar una distribución Gamma para el componente positivo con inferencia centrada en la media condicional. Quiroz et al. (2014) reconocen que la biomasa acústica de anchoveta presenta tres rasgos críticos: alta proporción de ceros, dependencia espacial y distribución fuertemente asimétrica hacia la derecha. Sin embargo, el modelo final mantiene el componente positivo bajo una distribución Gamma y evalúa su desempeño mediante criterios globales como el Criterio de Información de Desviación (DIC), la log pseudo-verosimilitud marginal (LPML), la tasa de acierto y errores predictivos, sin presentar diagnósticos residuales que demuestren que la distribución asumida reproduce la forma empírica de los datos. A ello se suma que el principal mapa de biomasa posterior no se presenta en la escala original del s_A, sino en escala logarítmica: aunque los autores indican que la biomasa estimada se computa mediante $\exp(\eta_{ij})$, la evidencia cartográfica usada para sostener el ajuste del modelo se muestra comprimida.

Castillo (2017) reproduce la misma lógica metodológica. El análisis exploratorio presenta valores NASC transformados como *log(x+1)*, y en la predicción espacial utiliza un modelo Hurdle Bayesiano con INLA, ecuaciones diferenciales parciales estocásticas (SPDE) y estructura espacio-temporal para modelar ocurrencia y valores positivos de s_A. La mayor cobertura temporal y la incorporación de campos espaciales latentes no corrigen el problema principal: la distribución Gamma sigue siendo insuficiente para representar el componente positivo de una variable acústica extremadamente asimétrica y con colas pesadas.

En consecuencia, los mapas derivados de estos modelos no deben interpretarse como validación de la biomasa acústica en escala original. Un mapa en escala logarítmica puede parecer espacialmente razonable porque comprime los valores extremos, pero esa compresión oculta precisamente la estructura que debe modelarse: la agregación extrema de

los cardúmenes. Si al volver a la escala original aparecen valores desproporcionados o colas no reproducidas por el modelo, el problema no es cartográfico, sino distribucional.

Los resultados de la presente tesis demuestran consistentemente que los modelos INLA-Gamma ajustados al s_A de los cardúmenes de anchoveta, tanto con campo espacial latente gaussiano mediante SPDE como sin él, y los modelos VAST-Gamma con campo latente espacial definido sobre una malla, no generaron residuos cuantílicos aleatorizados compatibles con una distribución normal *N(0,1)*. Los diagramas de gusano mostraron patrones sistemáticos cuadráticos y los diagramas de cubo presencia de asimetría y curtosis del momento, evidenciando modelos incorrectos. Por tanto, la complejidad espacial, la malla, el SPDE y la estructura jerárquica son insuficientes para describir la densidad acústica de anchoveta si la distribución del componente positivo está mal representada. En ese caso, el modelo puede producir mapas visualmente continuos, pero ser estadísticamente incorrecto.

En contraste, GAMLSS (Rigby & Stasinopoulos, 2005) y BAMLSS (Umlauf et al., 2018) disponen de familias continuas flexibles, como Box-Cox Power Exponential (Rigby & Stasinopoulos, 2004), adecuada para variables continuas con asimetría y curtosis (Flores et al., 2025), así como de distribuciones ajustadas a cero que combinan, para el componente continuo ($s_A > 0$), distribuciones como Box-Cox t (Rigby & Stasinopoulos, 2006) y Gaussiana Inversa Generalizada (Jørgensen, 1982; Rigby et al., 2019), junto con una distribución binomial para la ausencia-presencia ($s_A = 0$). Esto permite representar simultáneamente heterogeneidad estructural, asimetría, curtosis y exceso de ceros. En estos modelos, los residuos no muestran patrones sistemáticos en los diagramas de gusano y permanecen dentro de la región elíptica central de los diagramas de cubo, por lo que no evidencian un modelo incorrecto (Flores et al., 2025; Flores et al., 2026).

El modelado en la escala original de la densidad acústica (s_A) permite recuperar dinámicas ecológicas que permanecen ocultas bajo transformaciones, tales como umbrales ambientales no lineales, variabilidad nictemeral y fluctuaciones espaciales extremas (Flores et al., 2026). La capacidad de modelar conjuntamente la localización (μ), la escala (σ) y los parámetros de forma (ν y τ), integrando además la probabilidad de ausencia (π_0), proporciona una representación matemáticamente coherente de la naturaleza discontinua y altamente agregada del recurso. La transformación logarítmica no solo reescala la respuesta; la comparación de modelos sobre la escala original mediante GAIC corregido por el Jacobiano demuestra que introduce una pérdida relativa de información y puede conducir a modelos sin soporte estadístico (Flores et al., 2026). El tratamiento de los valores extremos es decisivo: la regresión distribucional aplicada en esta tesis preserva la estructura real del s_A, incluyendo los valores altos asociados a cardúmenes densos de interés biológico y pesquero.

En este contexto, la validación en ciencia pesquera no puede basarse en transformaciones logarítmicas, distribuciones rígidas de la familia exponencial ni métricas relativas para la selección de modelos. La evaluación debe fundamentarse en el análisis de residuos cuantílicos normalizados *N(0,1)* y en inspecciones visuales rigurosas que verifiquen la coherencia del modelo en todo el dominio de la variable.

5.1.8. La superioridad de un modelo simple frente a uno complejo

Izquierdo (2026) sostiene que incrementar la complejidad no necesariamente mejora el desempeño en los modelos de evaluación de stocks. Sin embargo, afirmar que un modelo simple correctamente especificado puede superar a uno complejo requiere demostrar previamente esa correcta especificación, lo cual no puede establecerse solo mediante métricas de desempeño, estabilidad estructural o comparación relativa entre modelos. Esta conclusión debe interpretarse, además, dentro del marco estadístico utilizado, que incluye modelos bayesianos implementados con INLA aplicados a la merluza europea (*Merluccius*

merluccius) y al atún aleta amarilla (*Thunnus albacares*), con la complejidad introducida mediante modelos delta jerárquicos con dependencia espaciotemporal y configuraciones estructurales en Stock Synthesis 3 (SS3).

Sin embargo, Izquierdo et al. (2021), en el reclutamiento de merluza europea (*Merluccius merluccius*), aplican mediante INLA un modelo bayesiano en dos partes con la batimetría como única covariable, campo geoestadístico Matérn y estructura temporal autorregresiva, con una distribución Gamma para el componente positivo, y seleccionan modelos con WAIC y LCPO. En este sentido, comparten las mismas limitaciones del GLM bayesiano de Quiroz et al. (2014) y del GAM bayesiano de Castillo (2017), de los GLM (Barriga, 2017) y de los GAM clásicos (Yuffra, 2017; Alegre, 2011; Castillo et al., 2018): la inferencia se concentra en la media bajo distribuciones de la familia exponencial. Si no se incorporan diagnósticos residuales capaces de evaluar patrones sistemáticos en la respuesta, no es posible verificar que el modelo, complejo o simple, esté correctamente especificado en sentido probabilístico.

En INLA, la validación mediante DIC, WAIC y CPO permite evaluar el ajuste relativo entre modelos, mientras que herramientas como el PIT o los Q-Q plots, empleadas de manera aislada, no bastan para verificar la adecuación distribucional ni para detectar patrones sistemáticos en los residuos. En consecuencia, un modelo puede presentar un ajuste global aparentemente óptimo y, aun así, estar estructuralmente mal especificado. La incorporación de complejidad espacial, temporal o poblacional no modifica por sí misma la forma de la distribución asumida para la respuesta; por tanto, una mala especificación puede persistir independientemente del nivel de detalle del modelo.

Esta tesis ha demostrado que la modelación del componente positivo mediante una distribución simple como la Gamma resulta insuficiente para reproducir asimetrías severas

y colas pesadas, un problema ampliamente documentado en procesos con estas características (Stasinopoulos & Rigby, 2007; Rigby et al., 2019; Stasinopoulos et al., 2024).

La capacidad de un modelo para incorporar estructura espacial, temporal o jerárquica no implica que pueda emplearse una única distribución para todos los procesos, especies, estaciones o escalas de observación. En GAMLSS, las variables explicativas pueden incorporarse mediante splines penalizados (P-splines), splines cíclicos para procesos periódicos, suavizadores de campo aleatorio de Markov para estructuras espaciales discretas, interacciones factor-suavizador para efectos diferenciados por estación, año, zona o grupo, y suavizadores de producto tensorial espacio-temporales para representar interacciones suaves entre coordenadas espaciales y variables temporales. Sin embargo, la incorporación de estas estructuras no sustituye la selección de una familia de distribución adecuada para la respuesta: la validez inferencial sigue dependiendo de que la distribución asumida reproduzca la forma condicional de los datos, incluyendo asimetría, heterocedasticidad, curtosis y comportamiento de los extremos. Por ello, aun dentro de un marco flexible como GAMLSS, el modelado de la dependencia o la estructura latente debe complementarse con la validación de la distribución mediante residuos cuantílicos, diagramas de gusano y diagramas de cubo.

5.1.9. Ausencia de datos de ecosondas comerciales calibradas en el entrenamiento y validación de los modelos

La estimación cuantitativa de la densidad acústica requiere registros obtenidos bajo estricto control metrológico: calibración, diagnóstico del sistema acústico y una adecuada relación señal/ruido. Por ello, los datos de ecosondas comerciales, incluso calibradas, no constituyen un patrón cuantitativo equivalente al de las ecosondas científicas calibradas (Foote et al., 1987; Simmonds & MacLennan, 2005; Karp, 2007; Demer et al., 2015).

La literatura reconoce que las embarcaciones pesqueras pueden funcionar como plataformas de observación acústica, pero su uso cuantitativo exige protocolos estrictos de calibración, configuración instrumental, reducción de interferencias y pruebas de ruido propio (Karp, 2007; Demer et al., 2015; Barbeaux et al., 2017). Las embarcaciones comerciales no son plataformas acústicas silenciosas: el ruido del motor y la hélice, la vibración, la cavitación, las burbujas bajo el transductor y la interferencia eléctrica y de otros equipos degradan la relación señal/ruido (Mitson, 1995; Simmonds & MacLennan, 2005; Karp, 2007; Demer et al., 2015), y el ruido radiado puede inducir evitamiento de los peces y sesgar las mediciones (Mitson, 1995; De Robertis & Handegard, 2013).

La evidencia peruana confirma estas limitaciones. Gutiérrez et al. (2016) reportaron ruido asociado al oleaje, el viento, fuentes biológicas y la propia embarcación, además del error secuencial de onda triangular (TWES) de los sistemas SIMRAD ES60 y ES70, con amplitud cercana a 1 dB pico-pico, que debe removerse antes de cualquier análisis cuantitativo. Ostos (2024) documentó fuentes de incertidumbre instrumental y operativa como la baja frecuencia de calibración, la configuración inadecuada por la tripulación, la degradación del transductor y fallas en sus sectores, y mostró que un RMS menor de 0.4 dB no basta para validar una calibración si no se verifican la ganancia, las aberturas del haz y los ángulos de compensación. Santivañez (2025), en mediciones in situ de TS de anchoveta con una SIMRAD ES60 en una embarcación cerquera, debió remover ruido impulsivo y de fondo, corregir el TWES y aplicar un filtro angular que eliminó entre 40% y 70% de los blancos.

Utilizar datos de ecosondas comerciales como patrón de validación incorporaría todas estas fuentes de error y sesgaría el ajuste y las predicciones del modelo. La validación estadística debe sustentarse en datos científicos calibrados, donde el sA es una respuesta cuantitativa confiable.

Los descriptores morfológicos, en cambio, son menos sensibles a estas fuentes de error que el s_A. El ruido y los errores de ganancia agregan energía a toda la agregación y sesgan directamente la integración, mientras que los límites geométricos del cardumen solo se desplazan en el borde, donde el S_v decae abruptamente. Por ello, cardúmenes con la misma geometría pueden presentar valores de s_A muy distintos por efecto del ruido. La utilidad de las ecosondas comerciales es, por tanto, aportar descriptores morfológicos, siempre que generen datos RAW con control de calidad que permitan su extracción estandarizada.

La tesis no busca demostrar equivalencia instrumental entre ecosondas científicas y comerciales ni validar metrológicamente los sistemas de la flota. Su contribución es una herramienta estadística para estimar la densidad acústica de los cardúmenes, en m^2/mn^2, a partir de descriptores morfológicos. La transferencia a ecosondas comerciales es una aplicación condicionada al control de calidad de los registros, la medición de ruido, la correcta configuración del equipo y la extracción estandarizada de los descriptores.

5.1.10. Ventajas de la regresión distribucional en el estudio de la ecología de la anchoveta

El comportamiento agregativo estacional de la anchoveta no se limita a variaciones en la densidad media: se manifiesta en cambios de la forma completa de la distribución del sA. El modelado distribucional en escala original permitió analizarlo con tres herramientas complementarias: los efectos locales acumulados (ALE), la función de densidad de probabilidad y la probabilidad de excedencia.

Las curvas ALE, al evaluar por separado la densidad típica ($Q_{0.5}$) y las densidades extremas ($Q_{0.95}$), diferenciaron los factores que desplazan la mediana de los que modulan las hiperagregaciones. Los descriptores morfológicos del cardumen (largo en verano y volumen3D en primavera) no alteraron la densidad típica, pero redujeron las densidades

extremas a medida que aumentó el tamaño del cardumen. Los factores espaciales actuaron de forma marcadamente estacional; la distancia a la costa reforzó su efecto sobre las densidades extremas ($Q_{0.95}$) en invierno, reflejando el desplazamiento mar afuera de los núcleos más densos. Un modelo convencional, al estimar un solo efecto sobre la media, no habría podido separar ambos comportamientos.

La función de densidad describió la transición estacional de la masa de probabilidad. En verano, el pico se desplazó hacia densidades altas, consistente con un régimen de fuerte agregación; en primavera, la masa se concentró cerca del origen, con el pico más alto y agudo de las cuatro estaciones. Este desplazamiento de la moda no implicó una reducción proporcional del extremo superior.

La probabilidad de excedencia confirmó ese desacople: primavera, con el menor parámetro de localización ($\mu = 6.23$, frente a 7.07 en otoño), mostró mayor probabilidad de superar los 5000 m^2/mn^2 (30.6% frente a 28.7%). El orden estacional según μ no coincide con el orden según la probabilidad de agregaciones extremas, porque esta depende de la forma completa de la distribución. La concordancia entre la excedencia ajustada y la empírica confirmó, además, que el modelo reproduce la cola derecha observada.

La agregación de la anchoveta es, por tanto, un proceso estacional expresado en la estructura distribucional completa del s_A. Un enfoque centrado solo en el parámetro de localización habría situado a primavera por debajo de otoño, omitiendo la estación con mayor probabilidad de agregaciones densas entre ambas e invirtiendo la señal relevante para la evaluación y la gestión pesquera.

5.1.11. Implicaciones para las evaluaciones acústicas y la gestión pesquera

Los resultados de este estudio demuestran que es posible estimar la densidad acústica a partir de indicadores morfológicos de los cardúmenes, lo que abre la posibilidad de aprovechar la información generada por la flota pesquera comercial, incluso en ausencia de

calibraciones acústicas. Tradicionalmente, los análisis cuantitativos se han restringido a ecosondas científicas calibradas, lo que limita el alcance espacial y temporal de la observación. Este enfoque permite transformar a la flota en una plataforma de observación complementaria a los cruceros científicos.

Para la gestión pesquera, una mayor disponibilidad de información sobre la densidad y agregación de los cardúmenes podría contribuir a delimitar mejor las zonas de pesca, optimizar el esfuerzo de la flota y detectar con mayor rapidez cambios en la distribución de la anchoveta asociados a la variabilidad del Sistema de la Corriente de Humboldt. Para el IMARPE, esto significa ampliar su capacidad de monitoreo más allá de los cruceros científicos y fortalecer el soporte para las decisiones de manejo.

La mayor contribución metodológica de esta tesis radica en comprobar las limitaciones de la regresión clásica centrada en la media al modelar la complejidad de la densidad acústica de la anchoveta, y proponer como solución una familia ultraflexible de cuatro parámetros con soporte en la línea real transformada a exponencial, capaz de capturar la asimetría, la heterogeneidad y las colas pesadas, en lugar de distribuciones como la lognormal o la Gamma. En una etapa posterior, este enfoque puede extenderse hacia mapas de densidad acústica condicionados por el hábitat, integrando las predicciones del modelo para los valores positivos y un modelo binario para la probabilidad de ausencia de la especie en ubicaciones no muestreadas entre transectos. Esta representación espacial puede escalar hacia estimaciones de abundancia o biomasa, siempre que se incorporen ecuaciones de fuerza de blanco (TS), estructura de tallas y relación longitud-peso.

CAPÍTULO VI.

6. CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

Las conclusiones de la presente tesis se derivan del cumplimiento del objetivo general y de los objetivos específicos orientados al desarrollo de una base metodológica para un servicio pesquero de predicción de la densidad acústica de los cardúmenes de anchoveta en el mar peruano.

- La regresión distribucional es superior a la regresión clásica centrada en la media para modelar la densidad acústica (s_A) de los cardúmenes de anchoveta, y sienta las bases operativas para un servicio pesquero predictivo aplicable a ecosondas comerciales sin calibración previa.
- El diagnóstico residual de la regresión clásica mediante GLM, GAM, INLA y VAST con la distribución Gamma reveló patrones sistemáticos propios de modelos incorrectos. GAMLSS, BAMLSS y NN-GAMLSS con las familias logSEP3 y logSEP4 no mostraron evidencia de mala especificación; RF-GAMLSS presentó desviaciones en el extremo superior en invierno y primavera.
- La longitud, la latitud, la distancia a la costa y la altura del cardumen ejercieron los mayores efectos sobre la densidad típica ($Q_{0.50}$) y las densidades más altas ($Q_{0.95}$), con patrones estacionales marcados.
- Los modelos distribucionales mostraron un desempeño predictivo indistinguible en casi todas las estaciones, a excepción de NN-GAMLSS en invierno, con el peor desempeño.
- GAMLSS y BAMLSS dan viabilidad metodológica a un servicio pesquero predictivo que aproveche la información de la flota industrial con ecosondas

comerciales no calibradas, sin los costos ni las restricciones operativas de calibrar cada embarcación.

6.2. Recomendaciones

- Modelar la densidad acústica (s_A) de los cardúmenes de anchoveta en su escala original mediante marcos distribucionales, evitando que los sesgos de la regresión clásica se trasladen a la toma de decisiones en la gestión pesquera.
- Emplear distribuciones flexibles como logSEP3 y logSEP4, capaces de representar las agregaciones hiperdensas que caracterizan la densidad acústica de los cardúmenes de anchoveta.
- Fundamentar la validación en residuos cuantílicos normalizados mediante diagramas de gusano y diagramas de cubo, verificar la autocorrelación con las funciones ACF y PACF, y emplear de manera complementaria el CRPS.
- Implementar el servicio pesquero predictivo como herramienta complementaria a las evaluaciones hidroacústicas destinadas a estimar biomasa, orientado a ampliar la capacidad de monitoreo y fortalecer la toma de decisiones bajo incertidumbre.
- Actualizar y reentrenar periódicamente los modelos incorporando nuevos datos hidroacústicos de cruceros científicos calibrados, a fin de ajustarlos a los cambios temporales, estacionales y estructurales en los cardúmenes.
- Extender, en una etapa posterior, el marco distribucional hacia la generación de mapas de densidad acústica condicionados por variables del hábitat marino, sentando las bases para mejorar la representación espacial de la densidad relativa y permitir su escalamiento hacia estimaciones de abundancia o biomasa.

CAPÍTULO VII.

7. REFERENCIAS BIBLIOGRÁFICAS

Akaike, H. (1983) Information measures and model selection. Int Stat Inst, 44, pp. 277–291.

Alegre, A. (2011). *Relaciones ontogénicas y espacio-temporales en la dieta del calamar gigante (Dosidicus gigas) en Perú, utilizando un Modelo Aditivo Generalizado (GAM)* [Tesis de maestría, Universidad Nacional Mayor de San Marcos]. Repositorio Institucional - UNMSM. https://hdl.handle.net/20.500.12958/2343

Alfatinah, A., Chu, H., Tatas, T., & Patra, S. (2023). Fishing area prediction using scene-based ensemble models. *Journal of Marine Science and Engineering, 11*, 1398. https://doi.org/10.3390/jmse11071398

Alvarez-Trasviña, E., Arizmendi-Rodríguez, D. I., Marín-Enríquez, E., Salcido-Guevara, L. A., Sánchez-Cárdenas, R., & Rodríguez-Domínguez, G. (2022). Distribution and abundance of the Pacific hake, *Merluccius productus*, and relationship with the environment in the Gulf of California, Mexico. *Ciencias Marinas, 48*(1). https://doi.org/10.7773/cm.y2022.3256

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Armas, E. (2024). *Pronóstico de corto plazo de zonas de pesca de anchoveta (Engraulis ringens) en el norte de Chile* (Tesis doctoral, Universidad de Concepción). Repositorio Institucional de la Universidad de Concepción. https://repositorio.udec.cl/items/62190e81-1e67-4490-ae7a-19ab1215de49

Atkinson, A. C. (1985). *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*. Oxford University Press.

Baptist, Martin, Van Bemmelen, R., Leopold, M., de Haan, D., Flores, H., Couperus, B., Fässler, S., & Geelhoed, S. (2018). *Self-foraging vs facilitated foraging by Lesser Black-backed Gull (Larus fuscus) at the Frisian Front, the Netherlands*. Bulletin of Marine Science -Miami-. 95. 10.5343/bms.2017.1179.

Barange, M. (1994). Acoustic identification, classification and structure of biological patchiness on the edge of the Agulhas Bank and its relation to frontal features. *South African Journal of Marine Science, 14*, 333–347. https://doi.org/10.2989/025776194784286969

Barbeaux, S. J., Fraser, D., Fritz, L. W., & Logerwell, E. A. (2017). *Cooperative multispecies acoustic surveys in the Aleutian Islands* (NOAA Technical Memorandum NMFS-AFSC-347). Seattle, WA: U.S. Department of Commerce, NOAA, National Marine Fisheries Service. https://doi.org/10.7289/V5/TM-AFSC-347

Barriga, E. (2017). *Variación en los patrones de distribución y abundancia de merluza (Merluccius gayi peruanus Ginsburg, 1954) frente a la costa norte de Perú (2004-2014)* [Tesis de maestría, Universidad Nacional Mayor de San Marcos]. Repositorio Institucional del Instituto del Mar del Perú. https://hdl.handle.net/20.500.12958/3233

Bertrand, A., Chaigneau, A., Peraltilla, S., Ledesma, J., Graco, M., Monetti, F., & Chavez, F. P. (2011). Oxygen: A fundamental property regulating pelagic ecosystem structure in the coastal southeastern tropical Pacific. *PLoS ONE, 6*(12), e29558. https://doi.org/10.1371/journal.pone.0029558

Bertrand, A., Segura, M., Gutiérrez, M., & Vásquez, L. (2004). From small-scale habitat loopholes to decadal cycles: A habitat-based hypothesis explaining fluctuations in pelagic fish populations off Peru. *Fish and Fisheries, 5*(4), 296–316. https://doi.org/10.1111/j.1467-2979.2004.00165.x

Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blanluet, A., Game, E., Pollock, K., Wolff, N., Everett, J., Neubert, S., Dunn, D., & Richardson, A. (2025). Drifting fish aggregation devices as a tool to study oceanic marine protected areas. *Fisheries Research, 289*, 107474. https://doi.org/10.1016/j.fishres.2025.107474

Breiman, L. (2001). *Random forests*. Machine Learning*, 45*, 5–32. https://doi.org/10.1023/A:1010950718922

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

Castillo, P. R. (2017). *Distribuição espaço-temporal da anchoveta peruana (Engraulis ringens) com base em dados acústicos e ambientais* (Tesis doctoral). Universidade Federal do Rio Grande (FURG), Brasil.

Castillo, R., Dalla, L., García, W., Madureira, L., Gutiérrez, M., Vásquez, L., & Koppelmann, R. (2018). Anchovy distribution off Peru in relation to abiotic parameters: A 32-year time series from 1985 to 2017. *Fisheries Oceanography, 28*(1), 1–13. https://doi.org/10.1111/fog.12419

Chavez, F. P., Bertrand, A., Guevara-Carrasco, R., Soler, P., & Csirke, J. (2008). The northern Humboldt Current System: Brief history, present status and a view towards the future. *Progress in Oceanography, 79*(2–4), 95–105. https://doi.org/10.1016/j.pocean.2008.10.012

Chávez, F. P., Ryan, J., Lluch-Cota, S. E., & Ñiquen, M. (2003). From anchovies to sardines and back: Multidecadal change in the Pacific Ocean. *Science, 299*(5604), 217–221. https://doi.org/10.1126/science.1075880

Chen, J., & Roth, J. (2024). Logs with zeros? Some problems and solutions. *The Quarterly Journal of Economics, 139*(2), 891–936. doi:10.1093/qje/qjad054

Chen, X., Chen, Y., Yu, H., & Liu, Q. (2024). Deep learning-based fishing ground prediction with multiple environmental factors at different temporal scales. *Applied Sciences, 11*(6), 11602920. https://doi.org/10.3390/app110611602

Citores, L., Ibaibarriaga, L., Santos, M., & Uriarte, A. (2024). A Bayesian spatially explicit estimation of daily egg production: Application to anchovy in the Bay of Biscay. *Canadian Journal of Fisheries and Aquatic Sciences, 81*, 1013-1028. https://doi.org/10.1139/cjfas-2023-0126

Clark, J. S. (2007). *Models for ecological data: An introduction*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691121789

Coetzee, J. (2000). Use of a shoal analysis and patch estimation system (SHAPES) to characterize sardine schools. *Aquatic Living Resources, 13*(1), 1–10. https://doi.org/10.1016/S0990-7440(00)00137-0

Cornejo, R., La Cruz, L., & Castillo, R. (2022). Distribución y biomasa de anchoveta y múnida en la Reserva Nacional de Paracas. *Boletín de Investigaciones Marinas y Costeras, 51*, 99–116. https://doi.org/10.25268/bimc.invemar.2022.51.1.947

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE. https://us.sagepub.com/en-us/nam/research-design/book237357

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems, 2(4), 303–314.

De Bastiani, F., Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., & Silva, L. A. (2022). Bucket plot: A visual tool for skewness and kurtosis comparisons. *Brazilian Journal of Probability and Statistics, 36*, 421–440. https://doi.org/10.1214/21-BJPS528

De Robertis, A., & Handegard, N. O. (2013). Fish avoidance of research vessels and the efficacy of noise-reduced vessels: A review. *ICES Journal of Marine Science, 70*(1), 34–45. https://doi.org/10.1093/icesjms/fss155

Demer, D. A., Berger, L., Bernasconi, M., Bethke, E., Boswell, K. M., Chu, D., Domokos, R., Dunford, A., Fässler, S., Gauthier, S., Hufnagle, L. T., Jech, J. M., Bouffant, N., Lebourges-Dhaussy, A., Lurton, X., Macaulay, G. J., Perrot, Y., Ryan, T., Parker-Stetter, S. L., Stienessen, S., Weber, T. C., & Williamson, N. (2015). *Calibration of acoustic instruments* (ICES Cooperative Research Report No. 326). Copenhagen, Denmark: International Council for the Exploration of the Sea. https://doi.org/10.17895/ices.pub.5494

Díaz-Astudillo, M., Riquelme-Bugueño, R., Bernard, K. S., Saldías, G. S., Rivera, R., & Letelier, J. (2022). Disentangling species-specific krill responses to local oceanography and predator's biomass: The case of the Humboldt krill and the Peruvian anchovy. *Frontiers in Marine Science, 9*, 979984. https://doi.org/10.3389/fmars.2022.979984

Diner, N. (2001). Correction on school geometry and density: Approach based on acoustic image simulation. *Aquatic Living Resources, 14*, 211–222. https://doi.org/10.1016/S0990-7440(01)01121-4

Doray, M., Masse, J., & Petitgas, P. (2010). Pelagic Fish Stock Assessment by Acoustic Methods at IFREMER.

Dunn, P. K., & Smyth, G. K. (1996). Randomised quantile residuals. *Journal of Computational and Graphical Statistics, 5*(3), 236–244. https://doi.org/10.1080/10618600.1996.10474708

Echoview Software Pty Ltd. (2025). Schools detection module algorithms. Hobart, Australia: Echoview Software Pty Ltd. Recuperado de

https://support.echoview.com/WebHelp/Reference/Algorithms/Schools_Detection_Module/Schools_detection_module_algorithms.htm

Efron, B. & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall, New York.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89–102. https://doi.org/10.1214/ss/1038425655

Eilers, P.H.C. & Marx, B.D. (2010). Splines, knots and penalties. Wiley Interdisciplinary Reviews: Computational Statistics, 2, 637–653.

Eilers, P.H.C. & Marx, B.D. (2021). Practical smoothing. The Joys of P-splines. Cambridge University Press.

Espinoza, P., & Bertrand, A. (2008). Revisiting Peruvian anchoveta (*Engraulis ringens*) trophodynamics provides a new vision of the Humboldt Current system. *Progress in Oceanography, 79*(2-4), 215–227. https://doi.org/10.1016/j.pocean.2008.10.022

Fang, S., Del Giudice, D., Scavia, D., Binding, C. E., Bridgeman, T. B., Chaffin, J. D., Evans, M. A., Guinness, J., Johengen, T. H., & Obenour, D. R. (2019). A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent. *Science of The Total Environment, 695*, 133776. https://doi.org/10.1016/j.scitotenv.2019.133776

Fernandez, C., Osiewalski, J., Steel, M.J.F., 1995. Modeling and inference with v-spherical distributions. Journal of the American Statistical Association 90, 1331–1340.

Fisher, L. & Navarro, V. (1994). *Introducción a la investigación de mercado (3ª ed.)*. México: McGraw-Hill Interamericana S.A. de C.V.

Flores, O., Rigby, R., Segura, M., Ledesma, J., Paulino, C., Chacón, G., & Valles, J. (2026). Effect of oceanographic variables and space-time factors on the density of Peruvian anchovy. *Progress in Oceanography*. https://doi.org/10.1016/j.pocean.2026.103688

Flores, O., Segura, M., Chacón, G., & Paulino, C. (2025). Predicting Peruvian anchovy school density with GAMLSS and BAMLSS using morphological descriptors. *Ocean Science Journal, 60*, 16. https://doi.org/10.1007/s12601-025-00211-3

Foote, K. G., Knudsen, H. P., Vestnes, G., MacLennan, D. N., & Simmonds, E. J. (1987). *Calibration of acoustic instruments for fish density estimation: A practical guide* (ICES Cooperative Research Report No. 144). Copenhagen, Denmark: International Council for the Exploration of the Sea. https://doi.org/10.17895/ices.pub.8265

Freckleton, R. (2009). The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology, 22*, 1367–1375. https://doi.org/10.1111/j.1420-9101.2009.01757.x

Fréon, P., & Misund, O. A. (1999). *Dynamics of pelagic fish distribution and behaviour: Effects on fisheries and stock assessment*. Fishing News Books.

Furusawa, M. (1991). Effects of transducer motion on quantifying single fish echoes. *Bulletin of the Japanese Society of Scientific Fisheries (Japan)*. ISSN: 0021-5392.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application, 1*, 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. doi:10.1198/016214506000001437

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gosset, W. S. (1908). *The probable error of a mean*. Biometrika, 6(1), 1-25.

Grüss, A., & Thorson, J. T. (2019). Developing spatio-temporal models using multiple data types for evaluating population trends and habitat usage. *ICES Journal of Marine Science, 76*(6), 1748–1761. https://doi.org/10.1093/icesjms/fsz075

Gutiérrez, M., Ramírez, A., Bertrand, S., Morón, O., & Bertrand, A. (2008). Ecological niches and areas of overlap of the squat lobster "munida" (*Pleuroncodes monodon*) and anchoveta (*Engraulis ringens*) off Peru. *Progress in Oceanography, 79*, 256–263. doi:10.1016/j.pocean.2008.10.019

Gutiérrez, M., Vásquez, C., Peraltilla, S., Aliaga, A., Zuzunaga, A., Méndez, E., Yarlequé, E., Habasque, J., & Munaylla, U. (2016). *Notes on the Peruvian experience on acoustic data collection and quantitative analysis of fish and macro zooplankton habitat using industry vessels* (SC-04-26). 4th Meeting of the Scientific Committee, South Pacific Regional Fisheries Management Organisation.

Halsey, L. G. (2019). The reign of the p-value is over. *Biology Letters, 15*(2). https://doi.org/10.1098/rsbl.2018.0814

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford University Press.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–310.

Haykin, S. (2009). *Neural networks and learning machines (3rd ed.)*. Prentice Hall.

Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In T. Kneib & G. Tutz (Eds.), *Statistical modelling and regression structures: Festschrift in honour of Ludwig Fahrmeir* (pp. 91–110). Heidelberg, Germany: Physica-Verlag. doi:10.1007/978-3-7908-2413-1_6

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación (6.ª ed.)*. McGraw-Hill.

Hilborn, R., & Mangel, M. (1997). *The ecological detective: Confronting models with data*. Princeton University Press.

Hilborn, R., & Walters, C. J. (1992). *Quantitative fisheries stock assessment: Choice, dynamics and uncertainty*. Chapman & Hall.

Hornik, K. (1991). *Approximation capabilities of multilayer feedforward networks*. Neural Networks, 4(2), 251–257.

Ihaka, R. & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3), 299–314. https://doi.org/10.2307/1390807

Izquierdo, F., Paradinas, I., Cerviño, S., Conesa, D., Alonso-Fernández, A., Velasco, F., Preciado, I., Punzón, A., Saborido-Rey, F., & Pennino, M. G. (2021). Spatio-temporal assessment of the European hake (*Merluccius merluccius*) recruits in the northern Iberian Peninsula. *Frontiers in Marine Science, 8*, 614675. https://doi.org/10.3389/fmars.2021.614675

Izquierdo, F. (2026). *Accounting for current challenges in commercial fish stock assessment models: Is it worth increasing the complexity?* [Tesis doctoral, Universidad de Vigo]. Repositorio del Instituto Español de Oceanografía (IEO-CSIC).

Jones, M. C. (2005). In discussion of Rigby, R. A., & Stasinopoulos, D. M. (2005), Generalized additive models for location, scale and shape. *Applied Statistics, 54*, 507–554.

Joo, R. (2008). *Identificación de puntos de pesca en los viajes pesqueros de anchoveta monitoreados por satélite mediante redes neuronales artificiales* [Tesis de pregrado]. Universidad Nacional de Ingeniería. Repositorio Digital IMARPE. https://hdl.handle.net/20.500.12958/2336

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution* (Lecture Notes in Statistics, Vol. 9). Springer. https://doi.org/10.1007/978-1-4612-5698-4

Karp, W. A. (Ed.). (2007). *Collection of acoustic data from fishing vessels* (ICES Cooperative Research Report No. 287). Copenhagen, Denmark: International Council for the Exploration of the Sea. https://doi.org/10.17895/ices.pub.5452

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Klein, N., Kneib, T. & Lang, S. (2015). Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data. *Journal of the American Statistical Association*, 110(509), 405–419. https://doi.org/10.1080/01621459.2014.912955

Korneliussen, R. J., Berger, L., Campanella, F., Chu, D., Demer, D., De Robertis, A., Domokos, R., Doray, M., Fielding, S., Fässler, S., Gauthier, S., Gastauer, S., Horne, J., Hutton, B., Iriarte, F., Jech, J. M., Kloser, R., Lawson, G., Lebourges-Dhaussy, A., & Thompson, C. (2018). *Acoustic target classification* (ICES Cooperative Research Report No. 344). International Council for the Exploration of the Sea. https://doi.org/10.17895/ices.pub.4567

Kotler, P. (1997). *Mercadotecnia. México*: Prentice-Hall.

Lawson, G. L., Barange, M., & Fréon, P. (2001). Species identification of pelagic fish schools on the South African continental shelf using acoustic descriptors and ancillary information. *ICES Journal of Marine Science*, 58, 275–287. https://doi.org/10.1006/jmsc.2000.1009

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Lima, A., Hazin, H., Silva, G., & Bezerra, M. (2023). Spatio-temporal distribution of the swordfish *Xiphias gladius* (Linnaeus, 1759) caught by the Brazilian longline fleet. *Ocean and Coastal Research*, 71, e23019. https://doi.org/10.1590/2675-2824071.20019adfl

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of Q-Q plots: The power of our eyes! *The American Statistician, 70*(2), 202–214. doi:10.1080/00031305.2015.1077728

MacLennan, D. N., Fernandes, P. G., & Dalen, J. (2002). A consistent approach to definitions and symbols in fisheries acoustics. *ICES Journal of Marine Science, 59*, 365–369. doi:10.1006/jmsc.2001.1158

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, 17(3), 283–295. https://doi.org/10.1016/S0167-6296(98)00025-3

Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics, 20*(4), 461–494. doi:10.1016/S0167-6296(01)00086-8

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis, 67*, 68–83. https://doi.org/10.1016/j.csda.2013.04.014

Maynard Smith, J. (1978). Optimization theory in evolution. *Annual Review of Ecology and Systematics, 9*, 31–56.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

Mesquita, C., Dobby, H., Pierce, G. J., Jones, C. S., & Fernandes, P. G. (2021). Abundance and spatial distribution of brown crab (*Cancer pagurus*) from fishery-independent

dredge and trawl surveys in the North Sea. *ICES Journal of Marine Science, 78*(2), 597-610. https://doi.org/10.1093/icesjms/fsaa105

Mitson, R. B. (Ed.). (1995). *Underwater noise of research vessels: Review and recommendations* (ICES Cooperative Research Report No. 209). Copenhagen, Denmark: International Council for the Exploration of the Sea.

Morane, C., Piroddi, C., Quattrocchi, F., Macías, D. & Christensen V. (2020). Spatial Distribution and Abundance of Mesopelagic Fish Biomass in the Mediterranean Sea. *Frontiers in Marine Science*. 7. 10.3389/fmars.2020.573986.

Mullahy, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics, 17*(3), 247–281. doi:10.1016/S0167-6296(98)00030-7

Murakami, D., Peters, G., Septier, F. & Matsui, T. (2023). Generalised hyperbolic state space models with application to spatio-temporal heat wave prediction. *Spatial Statistics*. 58. 100778. 10.1016/j.spasta.2023.100778.

Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., & Kitakado, T. (2009). Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: A case study in Sendai Bay, Japan. *ICES Journal of Marine Science, 66*(6), 1417-1424. https://doi.org/10.1093/icesjms/fsp105

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A, 135*, 370–384. https://doi.org/10.2307/2344614

Nielsen, M. (2015). *Neural networks and deep learning*. Determination Press.

Ñiquen, M., & Bouchon, M. (2004). Impact of ENSO on Peruvian anchovy. *Fisheries Oceanography, 13*, 159–165.

Nottestad, L., Anthonypillai, V., Høines, Å., Salthaug, A., Ólafsdóttir, A., Kennedy, J., Jacobsen, J., Smith, L., Eliasen, S., Jansen, T., Wieland, K. & dos Santos Schmidt, T.C. (2021). *International Ecosystem Summer Survey in Nordic Seas cruise report 2021*.

Ona, E. & Barange, M. (1999). Single target recognition. In: Ona, E. (Ed.), Methodology for Target Strength Measurements, ICES Cooperative Research Reports No. 235, pp.

Ospina-Alvarez, A., Bernal, M., Catalán, I., Roos, D., Bigot, J. & Palomera, I. (2013). Modeling Fish Egg Production and Spatial Distribution from Acoustic Data: A Step Forward into the Analysis of Recruitment. *PloS ONE*. 8. e73687. 10.1371/journal.pone.0073687.

Ostos, R. F. (2024). *Calibración y diagnóstico del funcionamiento de las ecosondas comerciales Simrad* (Tesis de pregrado). Universidad Nacional José Faustino Sánchez Carrión, Huacho, Perú.

Pantraki, E. K., Dimou, M. & Tekidou, E. (2024). Using predictive analytics to forecast fisheries trends and outcomes. *OMICS Online*.

Partridge, B.L. (1982). Structure and functions of fish schools. *Scientific American*, 245: 114-123

Pasache Medina, G. A., Tam Málaga, J., Puitiza, A. L., Yábar Torres, G., & Cahua, I. A. (2023). Nicho ecológico de la *Anchoa nasus* "samasa" en la bahía de Pisco (Perú). *Ecología Aplicada, 22*(2), 79-90. https://doi.org/10.21704/rea.v22i2.2084

Pauly, D., & Tsukayama, I. (Eds.). (1987). *The Peruvian anchoveta and its upwelling ecosystem: Three decades of change* (ICLARM Studies and Reviews No. 15). Manila, Philippines: International Center for Living Aquatic Resources Management.

Punt, A. E., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A. A., & Haddon, M. (2016). Management strategy evaluation: Best practices. *Fish and Fisheries, 17*(2), 303–334. doi:10.1111/faf.12104

Quijas, S. & Balvanera, P. (2013). *Biodiversity and Ecosystem Services*. 10.1016/B978-0-12-384719-5.00349-X.

Quiroz, Z., Prates, M. & Rue, H. (2014). A Bayesian Approach to Estimate the Biomass of Anchovies Off the Coast of Peru. *Biometrics*, Volume 71, Issue 1, Pages 209–217, https://doi.org/10.1111/biom.12227

Raftery A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83:251{266.

Raftery A. E. (1999). Bayes Factors and BIC, comment on: A critique of the Bayesian Information Criterion for Model Selection. *Sociological Methods & Research*, 27:411{427.

Reid, D. (2000). *Report on echo trace classification*. ICES Cooperative Research Report No. 238. 115 pp.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, 54(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Rigby, R. A., & Stasinopoulos, D. M. (2006). Using the Box–Cox t distribution in GAMLSS. *Statistical Modelling, 6*(3), 209–229. https://doi.org/10.1191/1471082X06st122oa

Rigby, R.A. & Stasinopoulos, D.M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statistics in Medicine*, 23: 3053-3076.

Rigby, R.A., Stasinopoulos, D.M., Heller, G.Z. & De Bastiani, F. (2019) *Distributions for modeling location, scale, and shape: Using GAMLSS in R*, Chapman and Hall/CRC.

Ripley, B. D. (2008). *Pattern recognition and neural networks*. Cambridge University Press.

Roa-Ureta, R. H., & Niklitschek, E. J. (2007). Biomass estimation from surveys with likelihood-based geostatistics. *ICES Journal of Marine Science, 64*(9), 1723-1734. https://doi.org/10.1093/icesjms/fsm149

Robotham, H., Bosch, P., Castillo, J. & Tapia, I. (2012). Clasificación acústica de anchoveta (*Engraulis ringens*) y sardina común (*Strangomera bentincki*) mediante máquinas de vectores soporte en la zona centro-sur de Chile: efecto de la calibración de los parámetros en la matriz de confusión. *Latin american journal of aquatic research*, 40(1), 90-101.

Rogers, A., Appeltans, W., Assis, J., Ballance, L., Cury, P., Duarte, C., Favoretto, F., Hynes, L., Kumagai, J., Lovelock, C., Miloslavich, P., Niamir, A., Obura, D., O'Leary, B., Ramirez-Llodra, E., Reygondeau, G., Roberts, C., Sadovy, Y., Steeds, O., & Aburto-Oropeza, O. (2022). Discovering marine biodiversity in the 21st century. In C. Sheppard (Ed.), *Advances in marine biology* (Vol. 93, pp. 23–115). Academic Press. https://doi.org/10.1016/bs.amb.2022.09.002

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.

Santivañez, M. A. (2025). *Mediciones in situ de fuerza de blanco de anchoveta (Engraulis ringens) a partir de datos colectados con una ecosonda comercial durante operaciones de pesca industriales en el stock norte-centro peruano* (Tesis de magíster). Universidad de Concepción, Concepción, Chile.

Santos Silva, J. M. C., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics, 88*(4), 641–658. doi:10.1162/rest.88.4.641

Schwarz, G., 1978. Estimating the dimension of a model. *Annals of Statistics* 6 (2): 461-464.

Simmonds, J., & MacLennan, D. N. (2005). *Fisheries acoustics: Theory and practice* (2nd ed.). Oxford, England: Blackwell Science.

Stasinopoulos, D. & Rigby, R. (2007). Generalised additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, Vol. 23, Issue 7, Dec 2007, https://www.jstatsoft.org/v23/i07/.

Stasinopoulos, D. M., Rigby, R. A., Heller, G., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R*. Chapman & Hall/CRC.

Stasinopoulos, M. & Rigby, R. (2024). gamlss.add: Extra Additive Terms for Generalized Additive Models for Location Scale and Shape [R package version 5.1-13]. CRAN. https://CRAN.R-project.org/package=gamlss.add

Stasinopoulos, M. D., Kneib, T., Klein, N., Mayr, A., & Heller, G. Z. (2024). Generalized additive models for location, scale and shape: A distributional regression approach, with applications. Cambridge University Press.

Tadikamalla P. & Durcikova A. (2003). Statistical Models (Non-Optimization Models). Encyclopedia of Information Systems. Elsevier, 2003. Pages 227-241. https://doi.org/10.1016/B0-12-227240-4/00169-6.

Thorson, J. T. (2019). Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research, 210*, 143-161. https://doi.org/10.1016/j.fishres.2018.10.013

Thorson, J. T., & Barnett, L. A. K. (2017). Comparing estimates of abundance trends and distribution shifts using single- and multispecies models of fishes and biogenic habitat. *ICES Journal of Marine Science, 74*(5), 1311–1321. https://doi.org/10.1093/icesjms/fsw193

Umlauf, N., Klein, N., & Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale and shape. *Journal of Computational and Graphical Statistics, 27*(3), 612–627. https://doi.org/10.1080/10618600.2017.1407325

van Buuren, S.V. & Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20(8), pp. 1259-1277. https://doi. org/10.1002/sim.746. https://pubmed.ncbi.nlm.nih.gov/11304741/

van Buuren, S.V. (2007). Worm plot to diagnosis fit in quantile regression. Stat. Model.,7, pp. 363-376. https://journals.sagepub.com/doi/10.1177/1471082X0700700406

Vásquez-Quispesivana, W., Inga, M. & Betalleluz-Pallardel, I. (2022). Inteligencia artificial en acuicultura: fundamentos, aplicaciones y perspectivas futuras. *Scientia Agropecuaria*, 13(1), 79-96. Epub 05 de enero de 2022.https://doi.org/10.17268/sci.agropecu.2022.008

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55*(1), 1–17. doi:10.1093/biomet/55.1.1

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC. https://doi.org/10.1201/9781420010404

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.

Wood, S. N. (2020). Inference and computation with generalised additive models and their extensions. TEST, 29(2), 307–339. https://doi.org/10.1007/s11749-020-00711-5

Xu, Z., Liu, Y., & Wang, J. (2025). Using a multi-output neural network model to standardize heterogeneous fisheries data. *Journal of Ocean University of China, 24*, 1373-1385. https://doi.org/10.1007/s11802-025-6136-0

Yuffra, M. (2017). *Caracterización y comparación de patrones espaciotemporales y nichos ecológicos*. https://doi.org/10.13140/RG.2.2.34419.68641

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Singer, Y. (2017). Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*.

Zhou, S., Campbell, R. A., & Hoyle, S. D. (2019). Catch per unit effort standardization using spatio-temporal models for Australia's Eastern Tuna and Billfish Fishery. *ICES Journal of Marine Science, 76*(6), 1489-1504. https://doi.org/10.1093/icesjms/fsz034

Zhu, Y., Minami, K., Iwahara, Y., Oda, K., Hidaka, K., & Hoson, O. (2021). Seasonal variation in fish school spatial distribution and abundance under the Kuroshio regular pattern and the large meander in Suzu coastal waters. *PLoS ONE, 16*(11), e0260629. https://doi.org/10.1371/journal.pone.0260629

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2.ª ed.). https://christophm.github.io/interpretable-ml-book/

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews, 82*(4), 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x

Nee, S., Colegrave, N., West, S. A., & Grafen, A. (2005). The illusion of invariant quantities in life histories. *Science, 309*(5738), 1236–1239. https://doi.org/10.1126/science.1114488

Smith, R. J. (1980). Rethinking allometry. *Journal of Theoretical Biology, 87*(1), 97–111. https://doi.org/10.1016/0022-5193(80)90222-2

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., & De Bastiani, F. (2023). P-splines and GAMLSS: A powerful combination, with an application to zero-adjusted distributions. *Statistical Modelling, 23*(5-6), 510–524. https://doi.org/10.1177/1471082X231176635

Stasinopoulos, M., & Rigby, R. (2023). *gamlss.dist: Distributions for generalized additive models for location scale and shape* (Versión 6.1-1) [Paquete de R]. CRAN. https://doi.org/10.32614/CRAN.package.gamlss.dist

Swartzman, G., Bertrand, A., Gutiérrez, M., Bertrand, S., & Vasquez, L. (2008). The relationship of anchovy and sardine to water masses in the Peruvian Humboldt Current System from 1983 to 2005. *Progress in Oceanography, 79*(2-4), 228–237. https://doi.org/10.1016/j.pocean.2008.10.021

ANEXOS

Tabla 12

Matriz de consistencia

Problemas	Objetivos	Hipótesis	Variable	Dimensiones	Indicadores
Problema General ¿La regresión clásica centrada en la media y transformaciones logarítmicas presenta limitaciones frente a la regresión distribucional en escala original, en cuanto a representar adecuadamente la naturaleza heterogénea, altamente asimétrica y no gaussiana de los datos de la densidad acústica s_A de los cardúmenes de anchoveta, lo que limita su aplicabilidad como base de un servicio pesquero predictivo orientado al uso de ecosondas comerciales no calibradas? **Problemas Específicos** ¿Qué enfoques de modelamiento estadístico, regresión clásica centrada en la media con transformaciones logarítmicas y regresión distribucional en la escala original, resultan adecuados para predecir la densidad acústica s_A de los cardúmenes de anchoveta en el mar peruano? ¿Qué variables explicativas presentan efectos sobre la densidad típica ($Q_{0.5}$) y las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta en el mar peruano? ¿Cómo diagnosticar la especificación residual y el desempeño predictivo de los modelos de regresión clásicos centrados en la media con transformaciones logarítmicas y la regresión distribucional en la escala original? ¿En qué medida los modelos desarrollados pueden servir como base metodológica para un servicio pesquero de predicción de la densidad acústica (s_A) de los cardúmenes de anchoveta para ecosondas comerciales no calibradas?	**Objetivo General** Modelar la densidad acústica (sA) de los cardúmenes de anchoveta mediante regresión clásica centrada en la media con transformaciones logarítmicas y regresión distribucional en escala original, evaluando su adecuación estructural, validez residual y capacidad predictiva para sustentar un servicio pesquero con ecosondas no calibradas. **Objetivos específicos** Aplicar la regresión clásica centrada en la media con transformaciones logarítmicas y regresión distribucional en la escala original, con el fin de evaluar si representan adecuadamente la estructura de los datos, en términos de variabilidad, asimetría y presencia de colas pesadas. Determinar la magnitud del efecto de las variables explicativas sobre la densidad típica ($Q_{0.5}$) y las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta en el mar peruano. Diagnosticar los patrones residuales y el desempeño predictivo de los modelos de regresión clásicos centrados en la media y la regresión distribucional en la escala original. Evaluar la aplicabilidad de los modelos ajustados como base metodológica de un servicio pesquero de predicción de la densidad de los cardúmenes de anchoveta para ecosondas comerciales no calibradas.	**Hipótesis General** H_0: La regresión clásica centrada en la media con transformaciones logarítmicas es equivalente a la regresión distribucional en la escala original, en el modelado de la densidad acústica de los cardúmenes de anchoveta, en términos de correcta especificación residual, representación de la estructura de los datos y desempeño predictivo. H_1: La regresión clásica centrada en la media con transformaciones logarítmicas no es equivalente a la regresión distribucional en la escala original, en el modelado de la densidad acústica de los cardúmenes de anchoveta, debido a diferencias en la especificación residual, representación de la estructura de los datos y desempeño predictivo. **Hipótesis Especificas** H_0: La regresión clásica centrada en la media con transformaciones logarítmicas y la regresión distribucional en la escala original no difieren en su capacidad para representar la estructura de los datos, en términos de variabilidad, asimetría y colas pesadas. H_1: La regresión distribucional en la escala original representa de manera más adecuada la estructura de los datos, en términos de variabilidad, asimetría y colas pesadas, que la regresión clásica centrada en la media con transformaciones logarítmicas. H_0: Ninguna variable explicativa presenta efecto sobre la densidad típica ($Q_{0.5}$) ni sobre las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta. H_1: Al menos una variable explicativa presenta efecto sobre la densidad típica ($Q_{0.5}$) o sobre las densidades más altas ($Q_{0.95}$) en los modelos para los cardúmenes de anchoveta. H_0: No existe evidencia de diferencias en la especificación entre la regresión clásica centrada en la media y la regresión distribucional en la escala original; los residuos cuantílicos son compatibles con una distribución normal estándar N(0,1), sin patrones sistemáticos persistentes en los diagramas de gusano o de cubo, y no existen diferencias sustanciales en las métricas de desempeño predictivo. H_1: La regresión clásica centrada en la media presenta mala especificación, evidenciada por residuos cuantílicos con desviaciones sistemáticas respecto a N(0,1); en contraste, la regresión distribucional en la escala original reduce dichas desviaciones y presenta mejor desempeño predictivo que los modelos clásicos. H_0: Los modelos ajustados no son aplicables como base metodológica de un servicio pesquero predictivo de la densidad acústica (s_A) a escala de cardumen. H_1: Los modelos ajustados son aplicables como base metodológica de un servicio pesquero predictivo de la densidad acústica (s_A) a escala de cardumen.	Servicio pesquero predictivo	Descriptores morfológicos del cardumen	Altura media (m) Largocorregido (m) Área corregida (m^2) Volumen 3D (m^3)
				Factores espacio-temporales	Profundidad media (m) Distancia a la costa (mn) Posición geográfica (°) Periodicidad horaria (h)
				Validez estadística del servicio	Residuos cuantílicos ~ N(0,1) Diagramas de gusano y cubo Funciones de autocorrelación y autocorrelación parcial
				Capacidad predictiva del servicio	Error Absoluto Medio (m^2/mn^2) Raíz del Error Cuadrático Medio (m^2/mn^2) Puntuación probabilística continua ordenada (m^2/nn^2)
			Densidad de los cardúmenes de anchoveta	Densidad relativa	Coeficiente del área de dispersión (m^2/mn^2)

Nota. Elaboración propia.