Trabajo Fin de Máster

Análisis estadístico para la predación y mortalidad por canibalismo del bacalao (*Gadus morhua*) en Flemish Cap.

Adrián Mencía Martínez

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Análise estadística para o estudo da predación e mortalidade por canibalismo do bacallau (*Gadus morhua*) en Flemish Cap.

Título en español: Análisis estadístico para el estudio de la predación y mortalidad por canibalismo del bacalao (*Gadus morhua*) en Flemish Cap

English title: Statistical analysis of the study of predation and mortality dut to canibalism of cod (*Gadus morhua*) in Flemish Cap.

Modalidad: Modalidad A/B

Autor/a: Adrián Mencía Martínez, Universidade de Santiago de Compostela

Director/a: Javier Roca Pardiñas, Universidad de Vigo; ,

Tutor/a: Ester Román Marcote, Centro Oceanográfico de Vigo - Instituto Español de Oceanografía; Mª Concepción González Iglesias,

Breve resumen del trabajo:

Análisis de contenidos estomacales de 5830 bacalaos (*Gadus morhua*) del período 1993-2018 en el Banco de Flemish Cap (Atlántico noroeste) con el fin de estudiar la pauta alimenticia basada en el canibalismo que esta especie desarrolla. Esta predación varía anualmente porque es tamaño-dependiente; depende de la distribución de tallas de la población, debiendo estar presente la fracción de la población que puede ser presa y también la fracción que desarrolla esta práctica alimenticia (más habitual solo durante una parte de cicle vital de los individuos). La abundancia de las diferentes clases anuales determina la intensidad de esta predación, y el consumo ocasionado puede ser una de las principales causas de motalidad natural, lo cual debe ser considerado en los modelos de evaluación pesquera.

Recomendaciones:

Otras observaciones:

Don Javier Roca Pardiñas de la Universidad de Vigo; doña Esther Román-Marcote y Mª Concepción González Iglesias del Centro Oceanográfico de Vigo - Instituto Español de Oceanografía, informan que el Trabajo de Fin De Máster titulado

Análisis estadístico para la predación y mortalidad por canibalismo del bacalao (Gadus morhua) en Flemish Cap

fue realizado bajo su dirección por don Adrián Mencía Martínez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su confomidad para su presentación y defensa ante un tribunal.

En Bilbao, a 31 de Agosto de 2022.

Don Javier Roca Pardiñas

Doña Esther Román-Marcote

Mª Concepción González Iglesias

Adrián Mencía Martínez

Agradecimientos

Las datos y muestras fueron obtenidos en las campañas cientçificas de investigación mediante arrastre de fondo, llevadas a cabo por España en el Atlántico noroeste en el Área de Regulación NAFO, en la Div. 3M correspondiente al Banco de Flemish Cap. Estas campañas han sido realizadas Instituto Español de Oceanografía (IEO-CSIC) con la participación del Instituto de Investigaciones Marinas (IIM-CSIC) y el Instituto Português do Mar e da Atmosfera (IPMA).

Las campañas han sido cofinanciadas por la Union Europea a través del *European Maritime and Fisheries FUND (EMFF)*/ Fondo Europeo Marítimo y de Pesca (FEMP) dentro del programa nacional de recogida, gestión y uso de datos en el sector de pesquerías y apoyo del consejo científico para el asesoramiento de la Política Pesquera Común. La financiación por la Unión Europea a través del Fondo Europeo Marítimo y de Pesca está enmarcada dentro del Programa Plurianual de la Unión para la recopilación, gestión y uso de los datos de los sectores de la pesca y la acuicultura para el período 2017-2019 cuyo marco legal es el REGLAMENTO (CE) Nº1004/2017 DEL PARLAMENTO Y DEL CONSEJO.

Una mención especial también a Mª Concepción González Iglesias por su dedicación, disposición y paciencia, sin la que varios aspectos de este trabajo no serían tan acertados.

Índice general

Resumen

Resumen en español

Analizados los contenidos estomacales de 5830 individuos de bacalao (*Gadus morhua*), correspondientes a un período de muestreo 1993-2018 en el Banco de Flemish Cap (Atlántico noroeste) con el fin de estudiar la pauta alimenticia basada en el canibalismo que esta especie desarrolla.

Esta predación varía anualmente porque es tamaño-dependiente; depende de la distribución de tallas de la población, debiendo estar presente la fracción de la población que puede ser presa y a la vez la fracción que desarrolla esta práctica alimenticia (más habitual solo durante una parte de ciclo vital de los individuos). La abundancia de las diferentes clases anuales determina la intensidad de esta predación, y el consumo ocasionada puede ser una de las principales causas de motalidad natural, lo cual debe ser considerado en los modelos de evaluación pesquera.

English abstract

Having analised stomachal contents of 5830 individuals of atlantic cod (*Gadus morhua*), belonging to a 1993-2018 sampling period in the Flemish Cap Bank (northwest Atlantic), the cannibalist dietary pattern this species develops will be studied.

This predation varies annually because of its size-dependancy; it depends on the size distribution of the population, where the presence of the population fraction that can be a prey is a must and so is the presence of that performs this dietary practice (which is more usual only during a certain part of the vital cicle of the individuals). The abundancy of the different classes determines the insentisy of this predation, and the causes consumption can be a major cause of natural mortality, which should be included in the fisheries assessment models.

Capítulo 1

Introducción

El estudio se sitúa en el contexto del seguimiento y gestión de la actividad pesquera en el Banco de Flemish Cap [1] (Figura 1.1), con énfasis en la pesquería de bacalao *G. morhua* por parte de la flota española. El seguimiento es llevado a cabo por el Programa de Prospección y Evaluación de los Recursos Pesqueros en Aguas Lejanas, del Área de Pesquerías del Instituto Español de Oceanografía (IEO), actualmente integrado en el Consejo Superior de Investigaciones Científicas-CSIC. Dicho seguimiento se lleva a cabo tanto mediante muestreo a bordo de los buques comerciales a lo largo de todo el año, como realizando la campaña científica "Flemish Cap"en verano todos los años desde 1988.

Este banco está situado en el Área de Regulación de *Northwest Atlantic Fisheries Organization* (NAFO), en la División 3M, correspondiente con el punto rojo de la Figura 1.1.

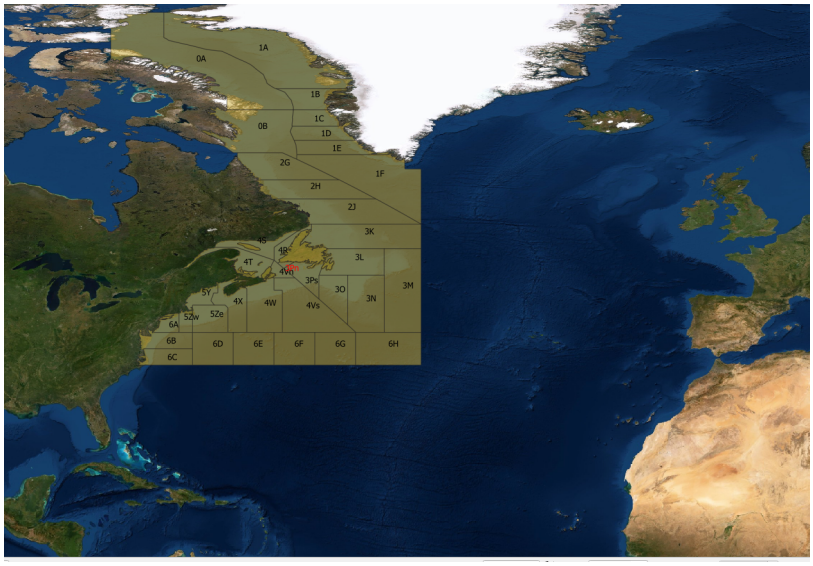


Figura 1.1: Área de la Convención NAFO

Se trata de un área de aguas poco profundas en el Atlántico Norte, formada por una plataforma submarina de 42.000 km^2 y un rango de profundidades que varía entre los 122 m y hasta más de 1000 m. Flemish Cap, centrado aproximadamente en 47 °N y 45 °W (Figura 1.2), es un banco aislado separado del Gran Banco de Terranova por el Flemish Pass, una zona de mayor profundidad. A pesar de ser un ecosistema aislado resulta altamente productivo con menos fluctuaciones abióticas estacionales e

[1] El Flemish Cap recibe este nombre debido a los numerosos pesqueros de Flandes que solían operar en la zona [1].

inter-anuales que en otras zonas de Terranova.

El Flemish Pass es una zona de mayor profundidad que lo separa del Gran Banco de Terranova, haciendo de Flemish Cap una zona aislada en la que la migración de especies está limitada, con características hidrográficas pecularíares.

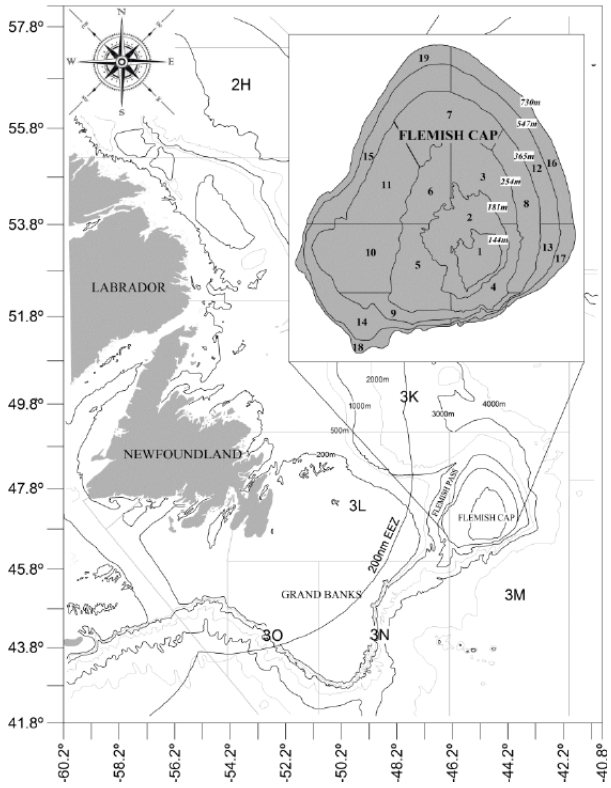


Figura 1.2: Área y detalle de Flemish Cap.

Las características de circulación oceánica sobre el banco de Flemish Cap están condicionadas por la Corriente del Labrador que fluye en dirección sur hacia el ecuador, transportando agua fría y de baja salinidad; y la Corriente del Atlántico Norte, que lo hace hacia el este y el noreste, cruzando el atlántico y transportando agua caliente y de alta salinidad. La Corriente de Labrador, se separa en dos ramas, pasando una de ellas a través del Flemish Pass, mientras que la otra fluye hacia el este por su flanco norte, y deriva después hacia el sureste (Figura 1.3).

Historicamente, la pesquería del bacalao ha sido de gran interés a nivel mundial debido a la importante actividad económica que ha generado dado al valor que esta especie ha tenido y tiene en el mercado.[2] En el Atlántico noroeste, dentro del área de regulación NAFO, la gestión de esta especie se divide en tres poblaciones que corresponden a su localización en diferentes áreas geográficas: la de la Div. 3M (en Flemish Cap), en las Divs. 3NO (sur del Gran Banco de Terranova) y en la Div. 3L (noroeste del Gran Banco). Las tres poblaciones estuvieron cerca del colapso a mitad de los años 90 debido a la sobrepesca, y desde 1992, en el caso del stock de Flemish Cap, la pesquería estuvo en moratoria (prohibida la pesca dirigida), hasta 2009; los otros dos stocks siguen en moratoria (De Cárdenas, 1996)[3]. La moratoria fue establecida en el intento de recuperación tras el tremendo declive de la población debido a la sobreexplotación unido a condiciones oceanográficas poco favorables para el reclutamiento. [2]

Fueron varios los factores que contribuyeron a la sobrepesca [4]. La tecnología pesquera, cada vez más eficiente, permitió a la industria ejercer un esfuerzo pesquero sin precedentes debido a la aparición

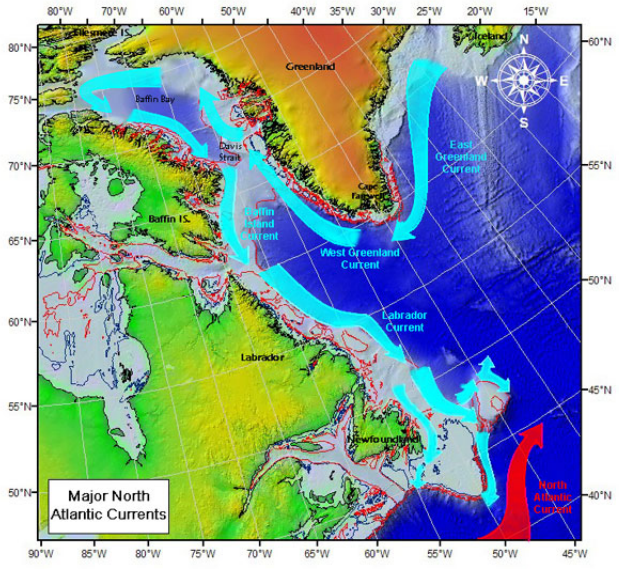


Figura 1.3: Corrientes en la zona de Flemish Cap.

de barcos de mayor eficacia (mayor potencia, mayor capacidad de almacenaje, mayor duración de las mareas [2], etc.) que eran capaces de faenar durante meses en las pesquerías. Sin embargo, las regulaciones para la preservación de las pesquerías no evolucionaron al mismo ritmo que la capacidad para su explotación, y a menudo los gobiernos y las organizaciones internacionales asignaron cuotas basándose en motivos económicos, en vez de en motivos ecológicos. A pesar de que la conservación se popularizó a partir de 1960, se cometieron sobreestimaciones en las poblaciones y como resultado también en las cuotas asignadas, que resultaron no ser sostenibles.

Los estudios de los hábitos alimenticios en el ecosistema del Banco de Flemish Cap muestran variaciones dependientes de la disponibilidad de especies como la gallineta, los hipéridos y los camarones. Dentro de su dieta, el bacalao también recurre al canibalismo, donde los individuos de mayor tamaño se alimentan de los individuos de menor tamaño, especialmente cuando ocurre un solapamiento en la distribución espacial y abundan los individuos más pequeños.

Hasta ahora la predación del bacalao se ha asumido constante a la hora de realizar las evaluaciones de la NAFO. Sin embargo se sabe que no es así, por lo tanto, el objetivo será el de identificar la pauta de depredación de esta especie y las variables que la afectan: abundancia, biomasa, y composición demográfica por tallas; presencia/ausencia de otras especies; temperatura, estación, profundidad, etc.

La contribución principal de este TFM es la de generar conocimiento sobre la pauta de predación del bacalao y las variables que le afectan, haciendo incapié en el canibalismo y en cómo afectan la composición demográfica por tallas y las dinámicas de predación a la disponibilidad de stock y preferiblemente a la distribución de tallas

Para ello se van a utilizar los datos obtenidos mediante las campañas del Centro Oceanográfico Español (IEO), que a su vez se basa en la metodología presentada en Doubleday (1981) [5] para definir la pauta de investigación. Se utilizarán los datos obtenidos sobre los contenidos estomacales de los individuos capturados.

[2] tiempo total que tarda un barco desde su salida al regreso.

Capítulo 2

Metodología de Muestreo

2.1. Procedimientos de generación y selección de localizaciones de pesca

La obtención de los datos por parte del IOE, se realiza siguiendo los estándares marcados desde la NAFO, desde donde se plantean procedimientos de muestreo con la intención de crear un sistema universal de muestreo de modo que la comparación de datos obtenidos por diferentes países sigan un procedimiento tal, que puedan ser manipulados de manera conjunta y/o comparados. En Junio de 1975 se ve la necesidad de la creación de un manual para estandarizar los métodos de muestreo de pesca de arrastre de fondo para la evaluación de poblaciones. En un entorno de cooperación como es el del área de la NAFO, donde diferentes países llevan a cabo las evaluaciones de las divisiones (en algunos casos de manera conjunta) la creación de métodos estandarizados es necesaria para la utilización eficiente de la información obtenida por cada una de las partes.

En este contexto se presenta el manual de W. G. Doubleday (1981) [5], donde se especifican las pautas para la realización de un muestreo correcto.

Tal y como especifica Doubleday, la distribución de los peces no es para nada uniforme, y se tiene poco control sobre las condiciones de pesca, lo que resulta en grandes variaciones en muestras de mismas especies en las mismas estaciones, por eso la aleatoriedad juega un papel importante. Se propone la utilización de muestreo estratificado por profundidades.

Son varias las ventajas de este tipo de muestreo:

- El muestreo se esparce por todo el área a muestrear, asegurando el número necesario de estaciones de pesca en cada estrato.
- Se puede aumentar el ratio de estaciones por unidad de área, en caso de querer obtener mayor precisión.
- Se pueden agregar los datos por estratos, dependiendo del estudio de interés.

Estratificación

Doudbleday (1981) presenta un esquema de estratificación para las áreas de la división 3M donde se consideran 19 estratos hasta 730 m de profundidad, que fue extendido por el *Department of Fisheries and Oceans*(DFO) ampliándo el esquema hasta 39 estratos que abarcan hasta 1460 m de profundidad, basándose en cuestiones biológicas e hidrográficas. Se puede observar un resumen de este esquema en la Tabla 2.2. A la hora de realizar la estratificación en profundidad se tiene en cuenta, por ejemplo, que las líneas de profundidad delimiten los hábitats naturales de las especies: 50 fath (91 m) para la limanda de cola amarilla (*Limanda ferruginea*), 150 fath (274 m) para la platija americana (*Hippoglossoides platessoides*).

En la Figura 2.1, se ve el sistema de estratificación de la división 3M. Donde se crean diferentes áreas en función de la profundidad, nombrándolas consecutivamente en sentido horario alrededor de la zona menos profunda. Por experiencia previa se sabe que los fondos de los estratos 26 y 27 presentan abundancia de esponjas, lo que hace que sus fondos no sean aptos para la pesca de arrastre que se emplea durante el muestreo, lo mismo pasa con los estratos del 35 al 39 por, presumiblemente, presencia masiva de corales. Todos ellos se eliminan de la prospección, resultando en 32 estratos muestreables. En el caso de la Figura 2.1, los puntos en verde corresponden a la localización de las pescas realizadas en la campaña de 2019.

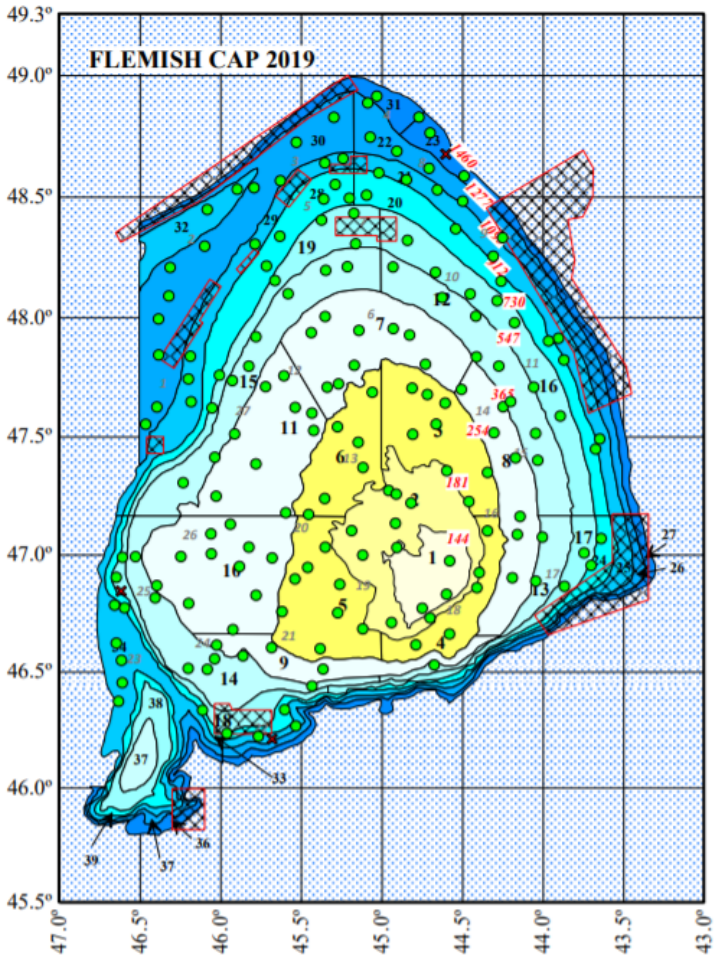


Figura 2.1: Esquema de estratificación de la división 3M y plan de pescas de 2019. Origen: IEO

Una vez definidos los limites de estratificación, se divide cada zona en unidades de misma área, de aproximadamente 35 millas náuticas cuadradas siendo las medidas del rectángulo de 5' de latitud y 10' de longitud, se denomina a esto rectángulo". Después se vuelve a dividir dicho rectángulo en 10 rectángulos de misma área y dimensiones 2,5' de latitud y 2' de longitud que se denominan *fishing units*.

La selección de estaciones se realiza tras numerar de manera consecutiva las unidades pesqueras y

	Área (millas cuadradas)	Estratos	Rectángulos	Unidades Pesqueras	n.pescas
profundidad < 730 m	10.555	19	309	3090	121
profundidad 730-1460 m	5.515	13	169	1.690	61
Total	16.070	32	478	4.780	181

Cuadro 2.1: Especificación y características del área prospectada y número de pescas.

seleccionar números al azar hasta haber obtenido el número necesario para llevar a cabo el plan de pescas. Para ello se establecen las siguientes condiciones:

- El número de pescas en cada estrato (Tabla 2.2) está fijado manteniendo la proporcionalidad con el número de unidades pesqueras de cada estrato, y garantizando un mínimo de dos pescas por estrato.
- Dentro de cada estrato se eligen aleatoriamente los rectángulos, sin reemplazamiento. Además no pueden seleccionarse dos rectángulos adyacentes.
- Dentro de cada rectángulo, la elección de entre las diez unidades pesqueras en la que hacer la pesca se realiza de manera aleatoria.

Se utiliza información previa, de otras campañas o comercial para eliminar unidades pesqueras que puedan ser problemáticas, siguiendo los siguientes criterios:

- enganche en el fondo.
- roturas en el copo o roturas importantes del arte.
- arrastres por debajo de 20 minutos.
- mal funcionamiento del arte

Rechazar una pesca por algunos de los problemas anteriores, significa que ésta no puede usarse para la cuantificación de biomasa ni para determinar la estructura de la población. Esta captura lograda, cuando existe, no se tría, no se clasifican las especies ni se realizan mediciones de talla. Sin embargo sí que sirve para otro tipo de muestreo biológico.

Si una vez seleccionada una estación, durante el muestreo se observa que el fondo no es adecuado para la realización del estudio, existen dos opciones:

1. antes de la realización del muestreo se seleccionan, también al azar, estaciones sustitutivas.
2. se selecciona la unidad pesquera adyacente en la dirección en que mejor convenga a la ruta planeada.

En este procedimiento se consideran dos características que pueden sesgar los resultados:

1. La distribución y abundancia de peces en zonas de fondo rocoso pueden diferir de aquellas en otros tipos de fondos.
2. Al sustituir una estación de fondo rocoso por una adyacente, la muestra puede no ser representativa de estaciones en las que no se puede realizar el arrastre de fondo.

2.1.1. Plan de Pescas

Tal y como se especifica en el Protocolo de la Campaña de Investigación Pesquera en Flemish Cap [6], se establecen las siguientes condiciones para el plan de pescas:

- El número de pescas en cada estrato está fijado manteniendo la proporcionalidad con el numero de estaciones en cada estrato y garantizando un mínimo de dos pescas por estrato.
- Dos pescas no pueden realizarse ni en la misma estación, ni en estaciones adyacentes.
- Dentro de cada estación la selección de la unidad pesquera se realiza de manera aleatoria.
- Se utiliza información, tanto de campañas anteriores como de la pesca comercial, para eliminar de las posibles unidades pesqueras las pescas problemáticas.

La velocidad objetivo en arrastre es de 3.5 nudos. La velocidad no puede mantenerse debido al peso insuficiente de las puertas de arrastre, por lo tanto, las pescas de mayor profundidad se hacen a la máxima velocidad posible, en torno a 3 nudos. Las pescas tienen una duración de 30 minutos de arrastre nominal. El orden de realización de las pescas seleccionadas se determina durante la campaña, fijando cada día las que se realizan el día siguiente, tratando de minimizar las rutas entre las pescas. Un plan detallado con el orden de todas las pescas resulta impracticable, pues las alteraciones debido a enganches y roturas del arte son imprevisibles.

2.2. Muestreo de Tallas

Se siguen las recomendaciones de la NAFO sobre los intervalos de talla y discriminación de sexos en los datos de muestreo (NAFO 1999). Según ello, la medición de talla en peces se harán sobre su longitud total y al centímetro inferior. Se anota además el sexo de cada individuo. En el caso del bacalao, la medición se hace al centímetro.

Como norma se miden todos los individuos presentes en la captura, y solo si ésta supera los 200 individuos de una especie se realizará sobre una muestra aleatoria, anotando siempre el peso de la muestra.

2.3. Procedimientos de muestreo de contenidos estomacales de peces en campañas científicas

2.3.1. Metodología general

Actualmente el muestreo en las campañas científicas de NAFO se realiza cada dos años, periodicidad iniciada el 2008. Previamente el muestreo era realizado todos los años, de tal forma que en la campaña de Flemish Cap hay una serie histórica iniciada en 1993. El muestreo y análisis de los contenidos estomacales ser realiza a bordo; es llevado a cabo por un grupo de dos personas, las mismas a lo largo de toda la campaña, con experiencia en esta tarea y designadas por el jefe de campaña. La toma de muestras es mediante un muestreo aleatorio, pero estratificado por sexo y rango de talla (igual que la recogida de muestras en campañas comerciales). Para las especies medidas por la longitud total (LT) al cm inferior (tiburones, rayas y la mayoría de los peces óseos) estos rangos son de 10 cm (0-9, 10-19, 20-29 cm, etc); se recogen 50 individuos de cada sexo por cada rango de talla.

Este protocolo de recogida que se viene realizando así desde 2004, en los años anteriores se procedía de forma diferente, de tal forma que el número de muestras se hacía atendiendo a cada lance muestreado donde se intentaba muestrear en función de las tallas capturadas. Este procedimiento incrementaba mucho el muestreo además de ocasionar un supra-muestreadas las tallas más abundantes.

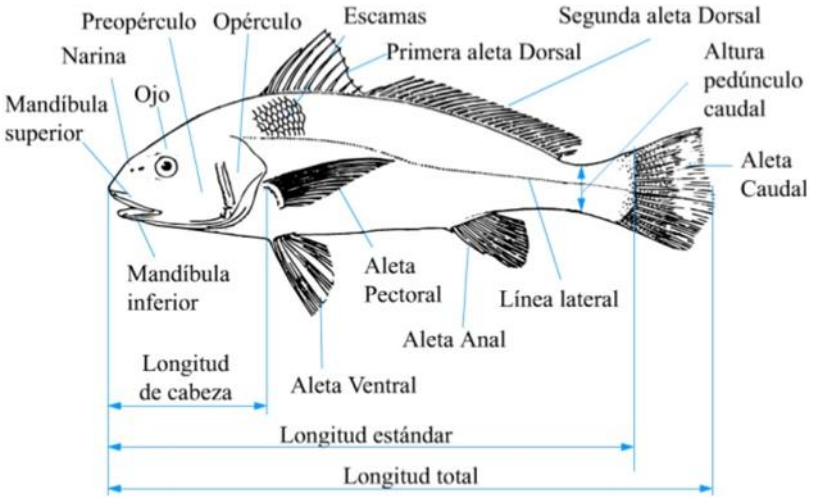


Figura 2.2: Funcionamiento de Isolation Forest

Si en la muestra tomada aparecen individuos cuyo estómago esté invaginado (cuando una parte del estómago se introduce en una porción adyacente) o contenga presas ingeridas en el copo, estos serán descartados. Aquellos individuos que hayan perdido todo o parte del contenido estomacal (regurgitación total o parcial del alimento) son válidos, pero no se tomarán datos sobre las presas, solo se indicará presencia/ausencia de contenido estomacal. Se observa y anota el tamaño y color de la vesícula biliar(amarilla,verde,transparente;vacía;mediada;llena) de acuerdo al criterio indicado por ICES (1991), para distinguir estómagos vacíos o con poco alimento de otro regurgitado total o parcialmente. Los datos anotados de cada ejemplar predador son: especie, talla, sexo, madurez sexual macroscópica, peso vivo, y repleción estomacal total en volumen (en cc) o en peso (en gr). Dicha cuantificación del contenido/presas se hará con la báscula de precisión en gramos, o con el trofómetro en cc.

El análisis específico del contenido estomacal consiste en la separación de los componentes por tipos o especies de presas, para su identificación al nivel taxonómico más bajo posible. Se anota cada una de las especies o presas, junto con el número de ejemplares, estado de digestión en el que se encuentran (1,fresco; 2, semidigerido o 3, totalmente digerido) y el porcentaje del volumen o del peso que representa respecto del contenido estomacal total.

La presa será medida siempre que el estado de digestión lo permita. La medición (en mm) es: longitud total en peces, cefalotórax en crustáceos decápoda tipo Caridea (ej. camarón o gamba) o Anomura (tipo ermitaño), caparazón en crustáceos decápoda tipo Brachyura (ej cangrejo), manto en cefalópodos decápodos (tipo calamar o sepia), cabeza en cefalópodos octópodos (tipo pulpo), ancho de la concha en bivalvos, longitud de la concha en gasterópodos y ancho del disco en ofiuras.

N.º de estrato	Intervalo de profundidad (*fathom*)	Área (mi^2 *náuticas*)	N.º de Unidades Pesqueras
1	70-80	342	100
2	81-100	838	250
3	101-140	628	180
4	101-140	348	100
5	101-140	703	200
6	101-140	496	150
7	141-200	822	240
8	141-200	646	190
9	141-200	314	90
10	141-200	951	280
11	141-200	806	240
12	201-300	670	200
13	201-300	249	70
14	201-300	602	170
15	201-300	666	200
16	301-400	634	190
17	301-400	216	60
18	301-400	210	70
19	301-400	414	120
20	401-500	525	160
21	501-600	486	150
22	601-700	533	160
23	701-800	284	90
24	401-500	253	80
25	501-600	486	150
28	401-600	530	160
29	501-600	226	70
30	601-700	1134	350
31	701-800	284	90
32	501-600	488	150
33	401-500	98	30
34	501-600	238	70
Total estratos (1-25 y 28-34)		16070	4780

Cuadro 2.2: Esquema de estratificación

Capítulo 3

Técnicas Estadísticas

En esta sección se realiza una introducción a las técnicas estadísticas empleadas en el análisis de datos del Capítulo 3.

3.1. Regresión

La regresión o análisis de regresión es un conjunto de técnicas estadísticas cuyo objetivo es estimar la relación entre una variable dependiente, y una o varias variables independientes.

El modelo general de regresión puede escribirse tal que

$$Y = f(X, \beta) + \mathcal{E}, \tag{3.1}$$

donde

Y = variable dependiente

X = variable independiente

β = parámetro escalar desconocido

$\mathcal{E}$ = término de error

y donde el objetivo determinar la función $f(X, \beta)$ que mejor se ajuste a los datos.

Aunque esa relación pueda ser de muchas formas, la forma más común es la regresión lineal (Sección 3.1.1), que trata de determinar la recta o combinación lineal que mejor se ajuste a los datos de acuerdo a algún criterio matemático.

3.1.1. Regresión Lineal

La regresión lineal se basa en asumir que la función $f(\cdot)$ de la Ecuación 3.1, es lineal. La regresión lineal simple se suele formalizar como la media condicionada de la variable respuesta en función de los valores que tome la variable explicativa

$$\beta(x) = \mathbb{E}(Y/X = x)$$

La variable respuesta se puede descomponer a través de la media condicionada de X, a lo que se le suma un error (diferencia entre el valor predicho y el observado) $\mathcal{E}$ no observable que verifica

que todos los errores son Normales e independientemente distribuidos de media 0 y misma varianza $\mathcal{E}_i \sim NID(0, \sigma^2)$ para todo x. También se le suma un intercepto β_0 quedando su expresión general en el caso de que hubiera más de una variable explicativa:

$$Y = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \mathcal{E} \quad \text{donde} \quad i = 1, ..., n \tag{3.2}$$

donde p es el número de variables explicativas y n el número de datos.
A la hipótesis de linealidad se le suman también otras hipótesis:

- **Homocedasticidad:** La varianza del error es la misma cualquiera que sea el valor de la variable explicativa:
$$Var(\mathcal{E}/X = x) = \sigma^2$$
- **Normalidad:** El error tiene distribución Normal
$$\mathcal{E} \in N(0, \sigma^2)$$
- **Independencia:** Las variables aleatorias son mutuamente independientes, lo que significa que sus errores de estimación asociados también son independientes entre ellos.

A partir de aquí se explica el procedimiento como si solo hubiera una variable explicativa. Denotando como $\hat{\beta_0}$ y $\hat{\beta_1}$ los estimadores de los parámetros, la predicción en y para el valor x sería tal que $\hat{\beta_0} + \hat{\beta_1} x$. De modo que los errores de predicción pueden escribirse

$$\hat{\mathcal{E}}_i = Y_i - \hat{\beta_0} - \hat{\beta_1} x_i \quad \text{para } i \in \{1, ..., n\}$$

Uno de los criterios matemáticos más comunes es, escoger $\hat{\beta_0}$ y $\hat{\beta_1}$ de modo que generen los residuos más pequeños. Con este objetivo, y para evitar que se compensen los residuos positivos con los negativos, se utiliza la minimización de la suma de los cuadrados de los residuos, y se elijen los coeficientes que minimizan esa suma

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Calculando las derivadas parciales respecto de β_0 y β_1, igualando a cero y despejando se obtienen los valores candidatos a mínimo, y se comprueba que constituyen un mínimo mediante la segunda derivada. Los estimadores obtenidos son

$$\hat{\beta_0} = \bar{Y} - \frac{S_{xY}}{S_x^2} \bar{x} \quad \hat{\beta_1} = \frac{S_{xY}}{S_x^2}$$

Además la varianza del error σ^2 se estima

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\mathcal{E}}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} x_i)^2$$

3.1.2. Regresión Cuantil

Así como la regresión por mínimos cuadrados ordinarios se encuentra vinculada con la media, la regresión cuantílica está vinculada a los cuantiles. El cuantil θ de una muestra, con $0 < \theta < 1$, será aquel valor b que deje una proporción θ de las observaciones por de bajo de b y una proporción $(1 - \theta)$ por encima. En el caso de la mediana $\theta = 0{,}5$, quedarán el 50 % de los datos por debajo de $b = M_e$.

Siguiendo la forma de la Ecuación 3.2, la regresión cuantil se expresa:

$$Q_r(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + ... + \beta_p(\tau)x_{ip} \quad \text{donde} \quad i = 1, ..., n \quad \text{y} \quad p \in \mathcal{R}$$

Lo que significa que en vez de ser constantes, son ahora una función con dependencia del cuantil, por lo que la función a minimizar también cambia. En el caso de una sola variable explicativa,

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \rho_\tau(y_i - (\beta_0(\tau) + \beta_1 x_{i1}(\tau))$$

Donde ρ es una función que asigna diferentes pesos a los datos dependiendo del cuantil.

$$\rho_\tau(u) = \tau \,\text{máx}(u, 0) + (1 - \tau)\,\text{máx}(-u, 0) \tag{3.3}$$

en este caso u es el error de cada punto. Si el error es positivo, la función 3.3 multiplica el error por τ, mientras que si es negativo lo multiplica por $(1 - \tau)$.

3.1.3. Regresión Linealizable

3.1.3.1. Regresión Exponencial

Cuando la función f de la Ecuación 3.1 no es lineal y se puede asumir que la relación entre las variables es del tipo:

$$Y = \beta_0 e^{\beta_1 X} + \mathcal{E} \tag{3.4}$$

se denomina regresión exponencial. Este tipo de regresión,sin embargo, es linealizable, ya que si la relación (X, Y) es exponencial, al aplicar logaritmos en la Ecuación 3.4,

$$ln(Y) = \beta_0 + \beta_1 X + \mathcal{E} \tag{3.5}$$

entonces $(X, ln(Y))$ es lineal y se puede trata como tal. Una vez hallada la solución valdría con deshacer el logaritmo para obtener la función de la regresión exponencial.

3.1.4. Regresión No-Paramétrica

En el caso de la regresión no paramétrica, no se asume linealidad ni ninguna otra distribución en la función desconocida $f(\cdot)$ de la Ecuación 3.1, así que el *modelo de regresión no-paramétrico general* se expresa

$$Y = f(X) + \mathcal{E}$$

Las siguientes condiciones caracterizan un modelo de regresión no paramétrico:

- No se asume una forma específica para $f(\cdot)$, solo que es una función suave.
- La varianza condicional σ^2 no tiene por qué ser constante, por lo que el modelo no tiene por qué ser homocedástico.
- Los errores $\mathcal{E}$ verifican $\mathbb{E}(\mathcal{E}/X = x) = 0$, pero no tienen por qué ser Normales.
- Las observaciones son independientes.

Los modelos de regresión no paramétrica se basan en las observaciones para especificar la forma del modelo, es decir, que la curva en cualquier punto se basa en las observaciones en ese punto y en algunos de las observaciones cercanas.

3.1.4.1. Estimador Nadaraya-Watson

Con kernel uniforme

El estimador de Nadaraya-Watson, propuesto por Nadaraya (1964) [7] y Watson (1964) [8] propone para la regresión la idea de unos puntos de evaluación centrados en las observaciones de la muestra que pondera uniformemente los puntos cercanos a tal punto de evaluación que se encuentran a menos de cierta distancia.

La predicción para $f(x)$ como una media de los valores respuesta cuyas covariables son cercanas a x:

$$\hat{f}_{NWU}(x) = \frac{\sum_{i=1}^{n} Y_i I(|X_i - x| < h)}{\sum_{i=1}^{n} I(|X_i - x| < h)}$$

Que también puede ser expresado como $\hat{m}_{NWU}(x) = \sum_{i=1}^{n} W_{i,h}(x) Y_i$ donde $W_{i,h}(x)$ es el peso de cada observación Y_i, que depende del ancho de ventana h y en un kernel uniforme $[-1, 1]$ dado por $I(|\cdot - x| < h)/2$.

Otros kernel

El kernel uniforme puede ser reemplazado por un kernel general. Quedando el estimador de Nadaraya-Watson como:

$$\tilde{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right)} \tag{3.6}$$

En general, K es una densidad totalmente especificada y simétrica centrada en 0. Aunque la importancia de la selección de la ventana suele cobrar una mayor importancia que la selección del tipo de kernel. Algunas funciones núcleo usuales son:

- **Uniforme:** $K(u) = \frac{1}{2} I\left(|u| < 1\right)$.
- **Gaussiana:** $K(u) = \frac{1}{\sqrt{2\pi}} exp\left(-u^2/2\right)$.
- **Epanechnikov:** $K(u) = \frac{3}{4}\left(1 - u^2\right) I(|u| < 1)$

3.1.4.2. Estimador Polinómico Local

Partiendo de la idea de ajustar un modelo lineal solo a los puntos que se encuentren a cierta distancia h de x. Se construye el modelo lineal en $(x - h, x + h)$:

$$I_i = \beta_0(x) + \beta_1(x) X_i + \mathcal{E}_i, \quad X_i \in (x - h, x + h)$$

Los estimadores para $\beta_0(x)$ y $\beta_1(x)$ se pueden obtener mediante mínimos cuadrados minimizando

$$\sum_{i=1}^{n} \left(Y_i - \beta_0(x) - \beta_1(x) X_i\right)^2 I\left(|X_i - x| \leq h\right)$$

Y si se considerasen mínimos cuadrados ponderados,

$$\sum_{i=1}^{n} \left(Y_i - \beta_0(x) - \beta_1(x) X_i\right)^2 K_h(x - X_i) \tag{3.7}$$

El estimador lineal local en x se define como:

$$\hat{f}_{LL}(x) = \hat{\beta}_0(x) + \hat{\beta}_1 x$$

donde $\hat{\beta}_0(x)$ y $\hat{\beta}_1$ son los valores que minimizan la anterior suma de cuadrados ponderada.

Partiendo de esta idea de ajustar un modelo lineal, se puede generalizar para realizar el ajuste local de un polinomio de grado p. Esta idea esta basada en la expansión de Taylor (asumiendo que f es suficientemente suave):

$$f(x_0) \approx f(x) + f'(x)(x_0 - x) + \frac{f''(x)}{2!}(x_0 - x)^2 + \cdots + \frac{m^p}{p!}(x_0 - x)^p$$

Que se puede ajustar minimizando

$$\sum_{i=1}^{n} \left(Y_i - \sum_{l=0}^{p} \beta_l(x)(x - X_i)^l \right)^2 K_h(x - X_i)$$

3.2. Diagnóstico del modelo

A continuación se presentan diferentes pruebas para comprobar la normalidad e independencia de los datos.

3.2.1. Normalidad

3.2.1.1. Shapiro-Wilk

El test de Shapiro Wilk (Shapiro y Wilk, 1965 [9]) sirve para determinar si una muestra proviene o no de una distribución normal.

$$H_0 : X \sim N(\mu, \sigma) \tag{3.8}$$

Se basa en el *qq-plot*, donde se acepta la normalidad si los puntos aparecenen en una línea recta. Así que el coeficiente de correlación de esos puntos puede utilizarse como indicar de la bondad de ajuste. El estadístico es:

$$D = \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{3.9}$$

donde $X_{(1)} < X_2 < \ldots < X_n$ son las muestras ordenadas y a_i son los componentes del vector $m^t V^{-1}((m^t V^{-1})(V^{-1}1m))^{1/2}$, donde $m = (m_1, \ldots, m_n)$ son los valores esperados de los estadísticos de orden de variables aleatorias independientemente e idénticamente distribuídas muestreadas de una distribución N(0,1), y V es la matriz de covarianzas de esos mismos estadísticos.

Si la hipótesis nula es cierta, el estadístico toma valores cercanos a 1.

3.2.1.2. Jarque-Bera

El Test de Jarque-Bera (Jarque y Bera, 1987 [10]), comprueba si una muestra $X_1, X_2, ..., X_n$ proviene o no de una distribución Normal,

$$H_0 : X \sim N(\mu, \sigma) \tag{3.10}$$

al comparar los estadísticos de la muestra con los estadísticos procedentes de una distribución Normal. Estos estadísticos son el coeficiente de asimetría (Ecuación 3.11) y el coeficiente de curtosis (Ecuación3.12),

$$\hat{A} = \frac{\hat{c}_3}{\overline{3}} \tag{3.11}$$
$$\hat{c}_2^{\overline{2}}$$

$$\hat{K} = \frac{\hat{c}_4}{\hat{c}_2^2} \tag{3.12}$$

donde $\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^j$.

En una Normal estos estadísticos toman los valores $A = 0$ y $K = 3$.
El estadístico de Jarque-Bera se define como

$$T_{JB} = n\left(\frac{\hat{A}^2}{6} + \frac{(\hat{K}-3)^2}{24}\right) \tag{3.13}$$

bajo la hipótesis de Normalidad, la distribución asintótica de T_{JB} es χ^2_2.

3.2.2. Independencia

3.2.2.1. Ljung-Box

Propuesto por Ljung y Box (1970) [11], este estadístico contrasta la independencia de un conjunto de observaciones $X_1, X_2, ..., X_n$. Se define el estadístico como:

$$Q = n(n+2)\sum_{j=1}^{m}\frac{r_j^2}{n-j} \tag{3.14}$$

r_j = autocorrelaciones muestrales acumuladas

m = el retardo temporal/ profundidad de la comparación.

Se rechaza la hipótesis nula cuando $Q > \chi^2_{1-\alpha,h}$, es decir, cuando el valor del estadístico es superior al valor dado por una distribución chi-cuadrado para un nivel de significación α y h grados de libertad.

3.3. Clustering

3.3.1. k-means

k-means es una de las técnicas más utilizadas para la identificación de clústeres, y se basa en la distancia (generalmente euclidea) para identificar un número determinado de clústeres k, se refiere a técnica y no algoritmo porque describe una manera de abordar el agrupamiento que es utilizado por diferentes algoritmos.

Una de las debilidades atribuidas a esta técnica es el hecho de tener que definir el número de grupos deseado (se habla de otras limitaciones en la sección **??**), teniendo en cuenta que no existe información previa, sin embargo, existen métodos para la selección de k.

Aunque existe una gran variedad de algoritmos que siguen el procedimiento de *k-means*, estos son los que se utilizan más comúnmente:

- Lloyd (o Lloyd-Forgy).
- MacQueen
- Hartigan-Wong

Aunque conceptualmente son similares, cuentan con diferencias que afectan a su coste computacional y a su desempeño en determinadas situaciones.

3.3.1.1. Algoritmos

Algoritmo de Lloyd/Forgy

El Algoritmo de Lloyd (Lloyd 1957 [12], publicado 1982) y el Algoritmo de Forgy (Forgy, 1965 [13]) son lo que se llaman modelos *batch (*también llamados modelos *offline)*, lo que significa que los centroides [1] se actualizan todos a la vez al final de cada iteración.

La diferencia entre los algoritmos es que el Algoritmo de Lloyd considera que la distribución de los datos es discreta, mientras que el algoritmo de Forgy la considera continua.

Para un conjunto de observaciones $[x_1, x_2, .., x_n] \in R^d$, donde R^d es un espacio d-dimensional, el algoritmo intenta buscar un conjunto de clústeres k donde sus centroides $C = [c_1, c_2, .., c_k] \in R^d$ sean una solución para el problema de minimización:

Para la distribución discreta:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n} d(c_i, x_i j) \tag{3.15}$$

Para la distribución continua:

$$E = \sum_{i=1}^{k} \int \rho(x) d(c_i, x_i j) dx \tag{3.16}$$

Donde $\rho(x)$ es la función de densidad y *d()* es la función de distancia. En caso de que la función de densidad no sea conocida, tiene que ser estimada a partir de los datos.

El primer paso del algoritmo es seleccionar los k centroides iniciales. Esto se puede realizar, por ejemplo:

- Basándose en conocimientos empíricos anteriores.
- Usando k observaciones aleatorias del conjunto de observaciones.
- Seleccionando las k observaciones que más alejadas estén entre sí.
- Eligiendo valores al azar dentro de R^d.

Después se realizan iteraciones de los siguientes pasos. En el primero, cada observación es asignada a su centroide más cercano (basándose en una métrica anteriormente definida). El segundo es actualizar cada centroide calculando la media de las observaciones pertenecientes a su mismo subespacio $C(R^d)$. Estos pasos se repiten hasta que ninguna observación cambia de centroide o hasta un número limitado de veces previamente establecido.

A modo esquemático:

1. Definir k.

2. Seleccionar métrica a utilizar.

3. Definir método de selección de los centroides iniciales.

4. Crear k centroides.

5. Para cada observación:

a) Calcular la distancia de cada observación a cada centroide.

[1] Un centroide es el centro geométrico (o la posición media de todos los puntos en todas las coordenadas en el caso de un espacio d-dimensional) de un objeto convexo y puede tomarse como una generalización de la media. Un objeto es convexo si dados dos puntos cualquiera, el segmento que une tales puntos también está contenido dentro del objeto.

***b*)** Asignar cada observación al clúster del centroide más cercano.

6. Recolocar el centroide a la media de las observaciones pertenecientes a su clúster.

7. Repetir los pasos 5-6 hasta que ninguna observación cambie de clúster o se alcance un número máximo de iteraciones.

Algoritmo de MacQueen

El algoritmo de MacQueen (MacQueen, 1967 [14]) es muy similar al de Lloyd, con la diferencia de que los centroides se recalculan cada vez que una observación cambia de clúster, y también una vez se han realizado una iteración completa.

Los centroides iniciales se determinan de la misma manera que en Lloyd/Forgy, y cuando una observación cambia de clúster solo se actualizan los dos centroides involucrados. Este algoritmo es más eficiente ya que los centroides se actualizan más a menudo.

A modo esquemático:

1. Definir k.

2. Seleccionar métrica a utilizar.

3. Definir método de selección de los centroides iniciales.

4. Crear k centroides.

5. Para cada observación:

***a*)** Calcular la distancia de cada observación a cada centroide.

***b*)** Asignar cada observación al clúster del centroide más cercano.

6. Recolocar el centroide a la media de las observaciones pertenecientes a su clúster.

7. Para cada observación:

***a*)** Calcular la distancia de cada observación a cada centroide.

***b*)** Asignar cada observación al clúster del centroide más cercano.

***c*)** Si una observación cambia de clúster, actualizar la localización de los dos centroides involucrados.

8. Repetir paso anterior hasta que todas las observaciones han sido consideradas.

9. Actualizar los centroides.

10. Repetir el paso 7 hasta que ninguna observación cambie de clúster o se alcance un número máximo de iteraciones.

Algoritmo de Hartigan y Wong

El algoritmo de Hartigan y Wong (Hartigan y Wong, 1979 [15]) busca la creación de los clústeres a partir de la minimización de la suma de cuadrados del error (SSE). Lo que significa que puede asignar observaciones a clústeres cuyo centroide no es el más cercano a la observación.

La suma de cuadrados del error, o simplemente la suma de cuadrados, es la diferencia entre cada observación y su valor predicho (en este caso el centroide), al cuadrado y sumada para todas las observaciones.

$$SSE = \sum_{i \in k} (x_i - c_k)^2 \tag{3.17}$$

Los clústeres son inicializados de la misma manera que en los casos anteriores y se asignan las observaciones a los clústeres con el centroide más cercano. Después, para cada observación se calcula la suma de cuadrados del error del clúster en el que se encuentra dicha observación *como si esa observación no estuviera en el clúster*, y el SSE del resto de clústeres *como si esa observación fuera parte de esos clústeres*.

Entonces si el SSE menor es el del clúster al que pertenece la observación, se pasa a evaluar la siguiente observación. Si por lo contrario, un clúster al que no pertenece la observación resulta tener un SSE menor, la observación se asigna al nuevo clúster y se actualizan los centroides de los dos clústeres involucrados. Se seguirá tal dinámica hasta que ninguna observación cambie de clúster o se llegue a un número máximo de iteraciones prefijado.

A modo esquemático:

1. Definir k.

2. Seleccionar métrica a utilizar.

3. Definir método de selección de los centroides iniciales.

4. Crear k centroides.

5. Para cada observación:

- ***a***) Calcular la distancia de cada observación a cada centroide.
- ***b***) Asignar cada observación al clúster del centroide más cercano.

6. Recolocar el centroide a la media de las observaciones pertenecientes a su clúster.

7. Para cada observación

- ***a***) Calcular el SSE de su clúster correspondiente, omitiendo la observación que se está evaluando.
- ***b***) Calcular el SSE del resto de clústeres, como si la observación que se está evaluando perteneciera a cada uno de ellos.
- ***c***) Asignar la observación al clúster con menor SSE.
- ***d***) Si la observación cambia de clúster, actualizar los centroides que corresponda.

8. Si ningún caso ha cambiado de clúster, parar. Si no, repetir 7.

3.3.1.2. Métricas

Índice Davies-Bouldin

Este índice presentado en Davies y Bouldin, 1979 [16] calcula el ratio de la dispersión de cada clúster con la distancia entre los centroides de los clústeres. El hecho de que los casos pertenecientes a cada clúster estén agrupados y a la vez los clústeres estén separados entre sí se traduce en un mejor resultado, y vendría indicado por un menor valor en el índice.

La dispersión se calcula tal que

$$dispersion_k = \left(\frac{1}{n_k} \sum_{i \in k}^{n_k} (x_i - c_k)^2 \right)^{1/2} \tag{3.18}$$

n_k = número de observaciones del clúster k

x_i = i-ésima observación del clúster k

c_k = es el centroide del clúster k

La separación entre clústeres *j*-*k* se calcula tal que

$$separacion_{j,k} = (\sum_{1 \leq j \leq k}^{N} (c_j - c_k)^2)^{1/2} \tag{3.19}$$

c_j, c_k = centroides de los clústeres j y k respectivamente.

N = número total de clústeres

Por lo tanto el ratio se calcula como

$$r_{j,k} = \frac{dispersion_j + dispersion_k}{separation_{j,k}} \tag{3.20}$$

Este ratio se calcula para cada par de clústeres. Para cada clúster, el mayor ratio de entre él y el resto de clústeres se define como R_k.

$$R_k \equiv \max_{j \neq k} r_{j,k} \tag{3.21}$$

El índice Davies-Bouldin es la media de esos ratios máximos:

$$DB = \frac{1}{N} \sum_{k=1}^{N} R_k \tag{3.22}$$

Fijando la distancia entre dos clústeres, si se reduce la dispersión de cada clúster, el índice se hace menor. Fijando la dispersión de los clústeres, si se aumenta la distancia entre ellos, el índice también se hace menor. En teoría, cuanto menor sea el índice, mejor es la separación entre los clústeres.

Estadístico pseudo-F
Es el ratio entre la *suma de cuadrados inter-clústeres* entre la *suma de cuadrados intra-clústeres.*

La suma de cuadrados inter-clústeres es la diferencia al cuadrado entre el centroide de un clúster y el centroide de todos los datos del conjunto, ponderado con el número de observaciones del clúster y sumado para todos los clústeres.

$$SS_{inter} = \sum_{k}^{N} n_k (c_k - c_g)^2 \tag{3.23}$$

La suma de cuadrados intra-clústeres es la diferencia al cuadrado entre cada observación y su centroide, sumado para todos los clústeres.

$$SS_{intra} = \sum_{k}^{N} \sum_{i \in k}^{n_k} (x_i - c_k)^2 \tag{3.24}$$

Quedando el estadístico como

$$PseudoF = \frac{SS_{inter}/(k-1)}{SS_{intra}/(n-k)} \tag{3.25}$$

N = número de clústers

n_k = número de observaciones en cada clúster k

c_k = centroide de un clúster k

c_g = centroide de todos las observaciones

Manteniendo la misma dispersión de cada clúster, cuánto más separados estén los centroides de los clústeres, mayor será el valor del índice. Y manteniendo la distancia, si disminuye la dispersión de cada clúster, mayor será el índice también. Por lo que cuanto mayor la diferenciación de cada clúster, mayor es el índice.

3.4. Detección de datos atípicos

Un dato atípico o *outlier*, en inglés, es una observación (o medición) anómalo en comparación con el resto de datos contenidos en un determinado conjunto de datos. Puede deberse a diversos motivos: una medición tomada, grabada o introducida incorrectamente en el conjunto de datos; el dato procede de una población diferente a la que se pretende estudiar... Existe una extensa literatura sobre estos datos anómalos:

- "Un dato atípico es una observación que se desvía tanto del resto de observaciones como para crear la sospecha de que fue creada por un mecanismo generador diferente (Hawkings, 1980) [17].
- "Un dato atípico es una observación (o conjunto de observaciones) que son inconsistentes con el resto de datos (BarnetLewis 1994) [18]
- "Un outliers es una observación que se encuentra fuera del patrón general de una distribución- (Moore y McCabe 1999)[19].

Por su parte, Rousseeuw (2017)[20] dice que los atípicos pueden ser, dependiendo de la circunstancia:

(a) Errores no deseados que pueden afectar negativamente al análisis.

(b) Valiosas pepitas de información inesperada.

Eliminar un dato de una muestra por haberlo considerado atípico puede llevar a perder información relevante debido a una singularidad del mecanismo generador, y a su vez incluir un dato atípico en una muestra puede confundir los resultados. Ambos casos alteran los análisis posteriores y pueden dirigir a conclusiones incorrectas si se llegara a tomar la decisión equivocada. Por lo tanto la importancia reside en identificar de manera adecuada qué datos son atípicos y cuáles no.

3.4.1. Método Clásico: Box-plot

El *Boxplot*, también conocido como Diagrama de Cajas y Bigotes es una forma estandarizada de representar la distribución de una muestra basándose en un sumario de 5 cinco números:

1. **Q1:** valor intermedio entre la mediana y el menor valor del conjunto de datos, percentil 25.
2. **Q3:** valor intermedio entre la mediana y el valor más alto del conjunto de datos, percentil 75.
3. **Rango Intercuantílico (RI):** rango entre los percentiles 25 y 75.
4. **Máximo:** Valor correspondiente a $Q3 + 1{,}5 * RI$

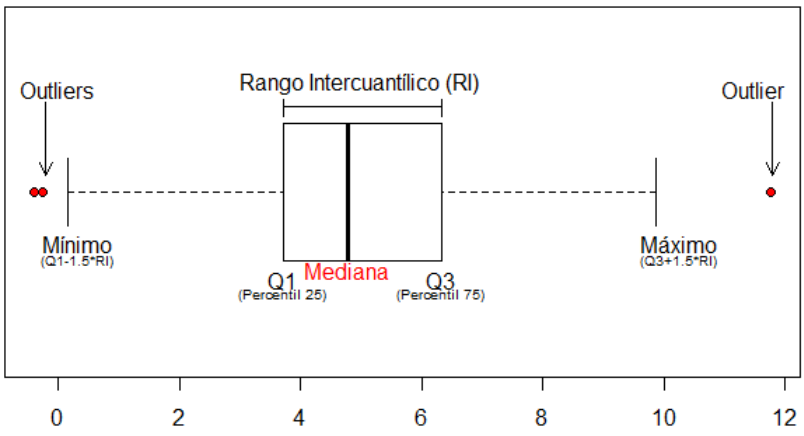


Figura 3.1: Componentes del Boxplot

5. **Mínimo:** Valor correspondiente a $Q1 - 1{,}5 * RI$

Esta herramienta considera que los datos provienen o se parecen a una distribución normal, a excepción de los posibles datos atípicos. Para ilustrar mejor estos resultados, en la Figura 3.2 se compara el boxplot de una muestra proveniente de una distribución normal y la función de densidad de una distribución normal, ambos casos de media 0 y varianza 1.

Se puede ver que el boxplot almacena en la caja, es decir, entre el Q3 y el Q1, el 50 % de los datos y entre los bigotes el 99.3 % de los datos. Los datos más extremos, que se encuentran a menos de $Q1 - 1{,}5 \cdot RI$ o más de $Q3 + 1{,}5 \cdot RI$ representan solo el 0.7 % de la probabilidad. Esos datos, son exactamente, los que la herramienta considera atípicos, ya que la probabilidad de que aparezcan es muy pequeña respecto al resto.

Esta herramienta, por lo tanto, trabaja de manera autosuficiente, es decir, el investigador no ajusta ningún tipo de parámetro para la búsqueda de valores atípicos.

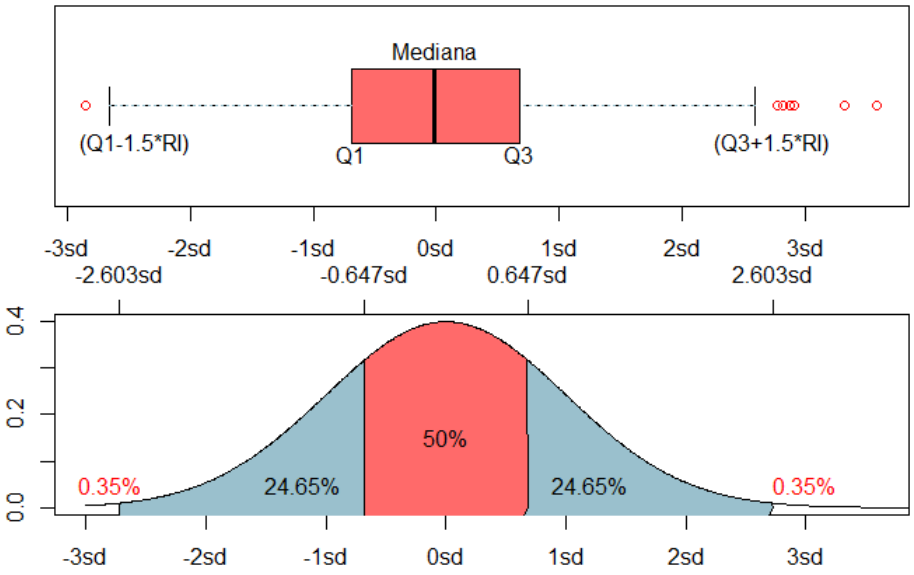


Figura 3.2: Comparación de Boxplot y densidad de una Normal(0,1)

3.4.2. Método Basado en un Modelo de Regresión: distancia de Cook

Los puntos de datos con grandes residuos (valores atípicos) y / o alto apalancamiento pueden distorsionar el resultado y la precisión de una regresión. La distancia de Cook mide el efecto de eliminar una observación determinada. Se considera que los puntos con una gran distancia de Cook merecen un examen más detenido en el análisis.

Habiendo ajustado un modelo de regresión 3.2, la distancia de Cook (1977)[21] se define como la suma de todos los cambios en el modelo de regresión cuando la observación i se quita de ella:

$$D_i = \frac{\sum_n^{j=1}(\hat{y}_j - \hat{y}_{j(i)})^2}{p\dot{s}^2} \tag{3.26}$$

Existen diferentes criterios en cuanto a la decisión de qué distancias de Cook deben tomarse como relevantes a la hora de reflexionar sobre una observación.

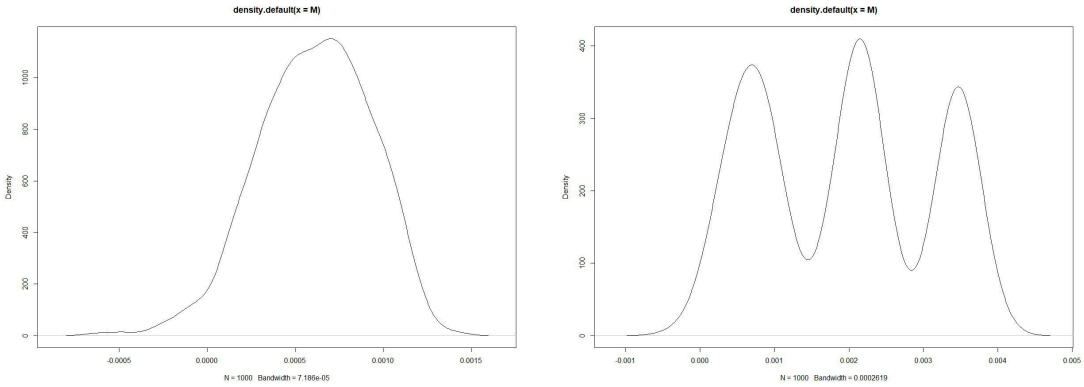
- Dado que la distancia de Cook está en la métrica de una distribución F con p y $n-p$ grados de libertad, el punto medio $F_{0,5}(p, n-p)$ se puede utilizar como límite. [22] Este valor es cercano a 1 para grandes tamaños de muestra, por lo que una opción simple es la de comparar los valores que cumplen $D_i > 1$ [23].
- Cualquier valor superior a $4/n$.

Sin embargo, la distancia de Cook ha demostrado no proporcionar siempre buenos resultados y suele convenir combinarla con otras técnicas para la comparación de resultados.[24]

3.4.3. Método Basado en Remuestreo: Bootlier

3.4.3.1. Bootlier-plot

Bootlier plot (bootstrap based oulier detection plot), propuesto por Singh y Xie (2003) [25] es una herramienta gráfica que busca la presencia de valores atípicos en una muestra. Su funcionamiento se



(a) Bootlier-plot sin atípicos (b) Bootlier-plot con un atípico

Figura 3.3: Bootlier-plots para: (a) 1000 datos de una normal; (b) Dato atípico 4,5 sumado al caso anterior

basa en un fenómeno muy sencillo: cuando existe un valor atípico en una muestra, algunas remuestras de esa muestra contendrán ese valor atípico mientras que otras no lo harán.

Se espera que la presencia de un potencial atípico, cause un incremento significativo en la media bootstrap, lo que conlleva a que la distribución bootstrap de la media de la muestra sea una distribución de mixturas. Esto quiere decir que si hubiera un dato atípico, el histograma bootstrap de la media de la muestra sea multimodal. Sin embargo, realizando un remuestro estándar, esa multimodalidad no es realmente visible a no ser que el dato atípico sea realmente severo. Para hacer que la herramienta sea más sensible se le da mayor peso a los valores extremos, utilizando el estadístico "media-media recortada".

El gráfico de densidad basado en bootstrap de "media - media recortada"(MTM) es una herramienta gráfica no paramétrica para la detección de outliers. Este gráfico es multimodal ante la presencia de datos atípicos.

Siendo $Y^b = \left[Y_1^b, \ldots, Y_n^b\right]$ $(b = 1, 2, \ldots, B)$ el equivalente bootstrap de Y, y k el número de observaciones recortadas de cada lado y donde k/n es una fracción pequeña e $Y_{(i)}^b$ los estadísticos ordenados de forma ascendente , la media recortada queda tal que:

$$\bar{Y}^b(k) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} Y_{(i)}^b \tag{3.27}$$

El MTM de la b-ésima remuestra bootstrap, M^b, es la diferencia entre la media aritmética y la media recortada:

$$M^b = \frac{1}{n} \sum_{i=1}^{n} Y_i^b - \bar{Y}^b(k) \tag{3.28}$$

El histograma $\frac{1}{2k} \sum_n^{n-k+1} Y_{(i)}$ tendrá al menos dos modas si hubiera datos atípicos: la correspondiente a las remuestras donde Y^* está libre de los valores extremos, y la que corresponde a los casos en las que esa Y^* sí que contiene los valores extremos.

3.4.3.1.1 Propiedades Bootlier-plot

Generalmente se utilizan versiones suavizadas de la estimación de la densidad para representar los resultados. La elección del número de datos a recortar k repercute en lo suave/dentada que resulte la

función. Singh y Xie (2003) [25] recomiendan mantener $k = 2$.

Para el caso en el que se crea que puedan existir datos atípicos tanto en el extremo superior como en el inferior de la muestra, se ha de actuar de diferente manera. Ya que habrá muestras sin ninguno de los atípicos, muestras con atípicos de solo un extremo y finalmente muestras con atípicos de ambos extremos. Estas últimas, suavizan la función de la densidad y seguramente dificultan la identificación de las modas. En ese caso se han de emplear *two-sided bootlier plots*, es decir dos histogramas:

$$T_L(M^b) = \bar{Y}^b - \frac{1}{n-k}\sum_{k+1}^{n} Y_{(i)}^b$$

$$T_U(M^b) = \bar{Y}^b - \frac{1}{n-k}\sum_{1}^{n-k} Y_{(i)}^b$$

En tal caso la recomendación es tomar $k = 3$ o $k = 4$.

Esta herramienta busca separaciones significativas en los extremos con sus vecinos más próximos, para hacer esto se tiene en cuenta tanto la dispersión de los datos vecinos a los extremos, como la dispersión de todo el conjunto. En consecuencia el bootlier tiene un enfoque no paramétrico. Sin embargo, si la brecha entre los datos atípicos y la muestra está cubierta por unos pocos datos, la multimodalidad desaparece.

3.4.3.2. Test de distribución libre Bootlier

En la sección 3.4.3.1 se define una herramienta gráfica para la detección de la presencia de atípicos, pero no para su identificación. Si bien es verdad que más adelante Singh y Xie (2003) [25] propone una herramienta para identificarlos, llamada *bootlier index*, que mide el grado de multimodalidad de una función de densidad y remarca la posibilidad de crear test basados en la distribución de ese estadístico. Sin embargo, no presentan un marco estadístico para usarlo. Mejorando esta idea, Candelon y Metiu (2013) [26] propone un test de distribución libre para contrastar la hipótesis nula de inexistencia de datos atípicos en una muestra dada, siendo $f_T(.)$ la función de densidad de $T(Y^*)$:

$$\begin{cases} H_0 : f_T \text{ tiene exactamente una moda (y ningún mínimo local)} \\ \qquad \text{en el interior de un intervalo cerrado } \Im \\ H_1 : f_T \text{tiene más de una moda en } \Im \end{cases}$$

H_0 es equivalente a la hipótesis nula de que no hay outliers y H_1, a la hipótesis alternativa de que hay uno o más outliers. Para poder comprobar estas hipótesis, se utilizan conjuntamente el Bootlier-plot y un test de distribución libre para multimodalidad propuesto por Silverman (1981) [27], que se basa en la propiedad de que el estimador kernel de la densidad es un estimador no paramétrico de la función de densidad.

Para el estadístico $T(Y*)$, de los estadísticos extraídos $Y_1^*, Y_2^*, \ldots, Y_n^*$ de la densidad $f_{T(.)}$, el estimador kernel de la densidad en cualquier punto x se expresa tal que:

$$\hat{f}(x,h) = \frac{1}{bh}\sum_{b=1}^{B} K\left(\frac{x - M^b}{h}\right), \tag{3.29}$$

donde h es una ventana y $K(.)$ una función kernel. Para una gran cantidad de clases de funciones kernel, incluida la normal estándar, el número de modas de la función kernel es una función continua por la derecha y decreciente de la ventana h. En consecuencia, para una ventana suficientemente grande $\hat{f}(.,h)$ tiene una sola moda en el interior del intervalo cerrado $\Im$. Además, existe una ventana estrecha h_{crit} para la que la densidad kernel estimada con tal ventana es unimodal. La llamada *ventana crítica*, definida como $h_{crit} = inf(h; \hat{f}(.,h))$ tiene precisamente una moda en $\Im$.

Esta ventana crítica es más grande para funciones multimodales que para unimodales, ya que en el primer caso se requiere de una ventana mayor para suavizar las diferentes modas. Utilizando esta propiedad Silverman (1981) [27] propone un procedimiento bootstrap para comprobar la multimodalidad de cualquier función de densidad.

Combinando el Bootlier-plot con el test de Silverman, se puede obtener un test de distribución libre para la presencia de datos atípicos en una muestra obtenida de cualquier función de densidad. El método, denominado "*Bootlier test*" se describe como:

1. Tomar una muestra grande $b = 1, 2, \ldots, B$ de muestras aleatorias de Y con reemplazamiento, y para cada remuestra Y^b computar el estadístico de media recortada M^b.
2. Obtener la estimación kernel estimada de la ecuación 3.29 para los estadísticos "media-media recortada" $M^1, \ldots, M^B$, denotada $\hat{f}_M(., h)$.
3. Estimar la ventana crítica $\hat{h}_{crit}$ de la densidad $\hat{f}_M(., h)$ y reestimar la función de densidad kernel utilizando la ventana crítica obtenida, obteniendo $\hat{f}_M(., \hat{h}_{crit})$.
4. Algoritmo de Silverman (1981):
   - Crear $M^{1*}, \ldots, M^{B*}$, una remuestra bootstrap tomada de la distribución con densidad $\hat{f}_M(., \hat{h}_{crit})$.[2]
   - Obtener la función de densidad kernel estimada en la ecuación 3.29 para el estadístico bootstrap "*mean-trimmed mean*" $M^{1*}, \ldots, M^{B*}$, denotándola $\hat{f}_{M*}(., h)$.
   - Estimar la ventana crítica bootstrap $\hat{h}_{crit}$ de la densidad bootstrap $\hat{f}_{M*}(., h)$.
   - Repetir los pasos (a)-(c) un gran número de veces.
5. La hipótesis nula de unimodalidad (no hay outliers en $\mathbf{Y}$) se rechaza si $Prob(\hat{h}^*_{crit} \leq \lambda_\alpha \hat{h}_{crit}) \geq 1 - \alpha$, siendo α el tamaño nominal y λ_α un parámetro de escalado que asegura que el tamaño empirico coincide con el nominal.

3.4.3.2.1 Identificación de outliers

Para localizar los outliers se utiliza el siguiente algoritmo: considerando el estadístico de forma ascendente $Y_{(i)} = [Y_{(1)}, Y_{(2)}, \ldots, Y_{(n-1),Y_{(n)}}]$ primero se emplea el Bootlier test en $Y_{(i)}$. Si la hipótesis nula de unimodalidad se rechaza, entonces Y contiene uno o más outliers y estos deben localizarse en las colas de $Y_{(i)}$. Secuencialmente se eliminan las observaciones de las colas siguiente las submuestras: $[Y_{(1)}, \ldots, Y_{(n-1)}], [Y_{(2)}, \ldots, Y_{(n)}], [Y_{(1)}, \ldots, Y_{(n-2)}], [Y_{(2)}, \ldots, Y_{(n-1)}], [Y_{(3)}, \ldots,$
$Y_{(n)}], [Y_{(1)}, \ldots, Y_{(n-3)}], etc$, y se emplea el Bootlier para cada submuestra hasta que no se pueda rechazar la hipótesis nula, siendo los datos que se quedan fuera de la primera muestra en la que no se rechace la hipótesis nula, los datos atípicos del conjunto Y.

3.4.4. Método alternativo 1: HDoutliers

Este algoritmo presentado por Wilkinson, 2016 [29] está diseñado para cumplir varios criterios a la vez:

- Permite identificar datos atípicos en conjuntos de datos que mezclan variables categóricas y variables continuas.
- Reduce problemas de alta dimensionalidad al realizar proyecciones aleatorias.
- Reduce problemas de tamaño de muestra utilizando un algoritmo de agregación de datos.

[2] En la práctica se computan remuestras insesgadas siguiendo a Efron (1979:) [28].

- Reduce problemas de detección de atípicos cuando estos crean un grupo uniforme que puede escaparse a la detección por métodos tradicionales.
- Funciona tanto para datos unidimensionales como multidimensionales.

El algoritmo se compone de estos pasos:

1. Si hay alguna variable categórica en el conjunto de datos, se convierte a una variable continua utilizando Análisis de Correspondencia [30].

2. Si hay más de 10000 columnas se utilizan proyecciones aleatorias para reducir el número de columnas a $p = 4 \cdot \log n/(\epsilon^2/2 - \epsilon^3/3)$ donde ϵ es el error de la raíz de la distancia.

3. Normalizar las columnas de la matriz resultante X_{nxp}.

4. Sea $row(i)$ la i-ésima fila de X.

5. Sea $\delta = \dot{1}/(\log n)^{1/p}$.

6. Se iniciliza *exemplars*, una lista de ejemplares con entrada inicial $[row(1)]$.

7. Se iniciliza *members* una lista de listas con entrada inicial [1]; cada ejemplar tendrá una lista propia de índices de miembros.

8. Se ejecuta lo siguiente:

9. Para cada $row(i)$ donde $i = 1, \ldots, n$:

a) Calcular la distancia para el ejemplar más cercano de *exemplars*.

b) Si $d < \delta$, añadir i a la lista *members* asociado al ejemplar más cercano.

c) Si $d > \delta$:

- Añadir $row(i)$ a *exemplars*.
- Añadir una nueva lista a *members* inicilizada con $[i]$.

10. Calcular las distancias a los vecinos más próximos en cada par de ejemplares presentes en la lista *exemplars*.

11. Ajustar una distribución exponencial para la cola superior de las distancias de los vecinos más próximos y calcular el punto $1 - \alpha$ de la función acumulada de la distribución ajustada.

12. Para cualquier ejemplar que sea significativamente distante del resto de ejemplares basándose en ese punto de corte, marcar todos los *members* correspondientes a ese *exemplar* como dato atípico.

3.4.5. Método alternativo 2: Isolation Forest

Isolation Forest, o iForest, es uno de los algoritmos más utilizados en el campo de la detección de anomalías debido a que sus propiedades le permiten adaptarse correctamente a distintos tipos de conjuntos de datos. La idea se basa es aislar puntos a través de construir iTrees, aquellos puntos que se aislen de un modo más sencillo, es decir con menos árboles, serán candidatos a atípicos. Este método consta solo de dos variables: el número de árboles a construir y el tamaño de sub-muestra. La siguiente figura ilustra la idea del algoritmo, los valores atípicos serán más fáciles de aislar bajo una partición aleatoria de los datos.

iForest no utiliza ninguna medida de distancia ni densidad para detectar atípicos. Esto elimina una gran parte de costes computacionales que tienen otros métodos. Además, tiene la capacidad para trabajar con grandes conjuntos de datos de alta dimensión. Una vez que el algoritmo término, cada punto

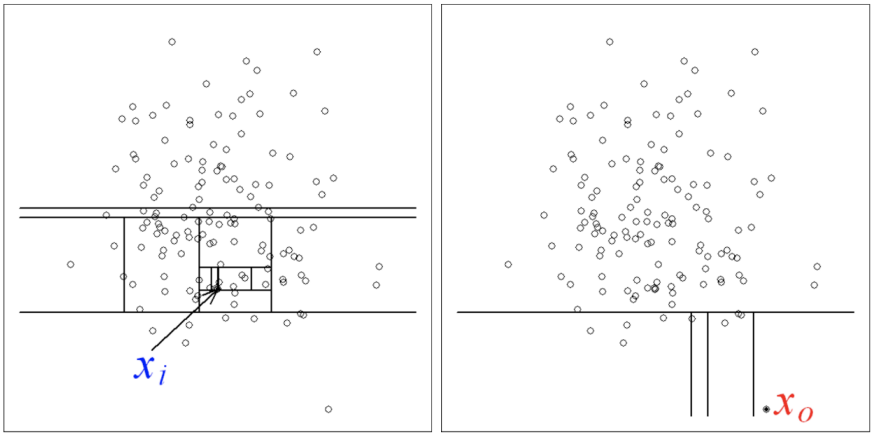


Figura 3.4: Funcionamiento de Isolation Forest

recibe una longitud de camino y una puntuación de atípico (anomaly score) por lo que ordenándolos en base a estas variables se puede hacer un ranking de los puntos más candidatos a atípico. Dado que este método no es el motivo principal del trabajo no vamos a profundizar más en el, salvo detallar que:

- Puntuaciones o scores cercanos a 1 indicarán que el punto es atípico.
- Puntuaciones o scores más pequeños que 0.5 calificarán al punto como no atípico.
- Si todas las puntuaciones o scores son proximas al valor 0.5 la muestra entera no tiene atípicos.

Este método se puede encontrar en la librería solitude en R [31].

3.5. Tests

3.5.1. U de Mann-Whitney

La prueba de la U de Mann-Whitney (Mann y Whitney, 1947 [32]) (también llamada de Mann-Whitney-Wilcoxon, prueba de suma de rangos Wilcoxon, o prueba de Wilcoxon-Mann-Whitney) es una prueba no paramétrica aplicada a dos muestras independientes que tiene en cuenta la localización y la forma de los datos de ambos grupos para determinar si un grupo tiende a tener valores diferentes al otro. [33]

Siendo X e Y las dos poblaciones a comparar y F y G sus funciones de distribución, respectivamente, se quiere comparar la independencia de dos muestras independientes $X_1, X_2, ..., X_n$ e $Y_1, Y_2, ..., Y_m$, comprobando las hipótesis

$$H_0 : \ F(x) = G(x)$$
$$H_1 : \ P(Y \geq X) > 0{,}5.$$

La U de Mann-Whitney, realiza la comparación entre las distribuciones mediante este estadístico:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I(Y_j > X_i). \tag{3.30}$$

El estadístico cuenta los pares (X_i, Y_j) en los que Yes mayor que X, por lo que cuanto mayor sea el estadístico, más diferentes serán las muestras.

La distribución del estadístico se aproxima adecuadamente a la distribución Normal para tamaños muestrales grandes de n y m, dada por la expresión:

$$\frac{U - \mu_U}{\sigma_U} \approx N(0,1). \tag{3.31}$$

Donde μ_U y α_U son la media y la desviación estándar de U si la hipótesis nula es cierta, y vienen dadas por las siguientes fórmulas:

$$\begin{aligned} \mu_U &= nm/2 \\ \sigma_U &= \sqrt{\frac{nm(n+m+1)}{12}} \end{aligned} \tag{3.32}$$

3.6. Otras Técnicas

3.6.1. Transformación Box-Cox

La transformación Box-Cox, es una familia de transformaciones basada en un parámetro cuyo valor óptimo se computa basándose en un objetivo específico. Suele utilizarse para corregir sesgos en la distribución de errores, para corregir la heterocedasticidad y normalmente para corregir la no linealidad entre dos variables. Viene dada por la siguiente expresión:

$$\forall y \in (0,\infty) \quad t_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{si} \quad \lambda \neq 0 \\ \ln(y), & \text{si} \quad \lambda = 0 \end{cases} \tag{3.33}$$

Cada valor de λ produce una transformación diferente, y la selección del parámetro se puede realizar de acuerdo a diferentes criterios, Box y Cox (1964) [34] propusieron hacerlo basándose en la maximización de la función de verosimilitud, aunque se puede realizar de muchas otras formas como maximizando el coeficiente de correlación de la gráfica de probabilidad o el estadístico de bondad de ajuste de Shapiro-Wilk.

Tomando el objetivo de maximizar la función de log-verosimilitud y asumiendo que las observaciones transformadas según la Ecuación (3.33) provienen de una distribución Normal con media μ y desviación estándar σ, se puede definir la función de log-verosimilitud como:

$$\log[L(\lambda,\mu,\sigma)] = \frac{-n}{2}\log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 + (\lambda - 1)\sum_{i=1}^{n} log(x_i) \tag{3.34}$$

Para un valor fijado de λ, la función de máxima verosimilitud se maximiza reemplazando μ y σ por sus estimadores de máxima verosimilitud:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.35}$$

$$\hat{\sigma} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2} \tag{3.36}$$

Aunque en la teoría se maximice la ecuación 3.34 derivando, en su implementación práctica es más usual que se haga iterando con diferentes valores de λ y utilizando los valors de μ y σ dados por las Ecuaciones (3.35) y (3.36). Es este el caso de la función *boxcox* del paquete MASS (v7.3-53; Venables, W. N. & Ripley, B. D. (2002))[35].

Capítulo 4

Análisis de los Datos

4.1. Análisis Exploratorio

Se dispone de datos de contenidos estomacales correspondientes a la alimentación de 5830 ejemplares de bacalao en el Banco de Flemish Cap recogidos mediante los procedimientos de muestreo especificados en la sección 2.3. Las campañas han sido realizadas en verano, habiéndose obtenido datos desde 1993 hasta 2018.

La base de datos disponible cuenta con la siguiente información: año, código indicador de cada campaña pesquera, lance (cada pesca realizada en una posición), número de registro del predador, número de registro de la presa, especie de predador muestreada (*G. morhua* en este caso), talla del predador, sexo del predador, estado de madurez sexual macroscópica, clasificación juvenil/adulto, presencia/ausencia de contenido estomacal, peso del contenido estomacal, grupo de la presa, especie presa, grupo taxonómico de la presa, número de presas de una misma especie presentes en el contenido estomacal, estado de digestión, talla mínima en caso de haber más de una presa de la misma especie, talla máxima en el mismo caso, talla de la presa en caso de haber una sola presa o en caso de haber más de dos (talla intermedia), talla de la presa calculada a partir de restos duros cuando no se puede determinar de manera visual, peso del depredador (vivo o calculado), peso vivo del pedrador, peso de la presa, profundidad del muestreo, tipo de campaña (científica o comercial), mes, estación, muestreo diurno/nocturno, temperatura de fondo de la zona de muestreo y temperatura superficial de la zona de muestreo.

Antes de iniciar el análisis propiamente, interesa obtener una visión global de los datos con los que se va a trabajar.

La Figura 4.1 muestra el número de predadores muestreados en cada año del estudio, que aunque contenga datos desde 1993 hasta 2018, no se ha realizado todos los años. Se ha realizado anualmente desde 1993 hasta 2006, y después en 2008, 2010, 2011, 2012, 2014, 2016 y 2018. El número de predadores muestreados varía considerablemente a lo largo de los años, siendo el año 2003 en el que menos ejemplares se obtuvieron (302 en total) y 2008 el que más (1749). Que el número varíe tanto, puede ser reflejo de la situaciones de las poblaciones en ese mismo momento, dado que la pesquería se encontraba en moratoria desde 1999, tendría sentido que el bajo número de individuos localizados sea un simple reflejo de las poblaciones de *G. morhua* en esos años.

Aunque en algunos casos existan diferencias entre el número de ejemplares macho y hembra muestreados (en 1997 o 2014 por ejemplo) y en otros casos el número es similar, no parece preocupante ya que la proporción de cada uno parece estar bastante equilibrada: hay 5721 ejemplares hembra (56 %) en los datos disponibles, y 9317 machos (44 %), tal y como se muestra en la Figura 4.2.

Al mirar la representación de tallas de los predadores en la estimación de la densidad de la Figura

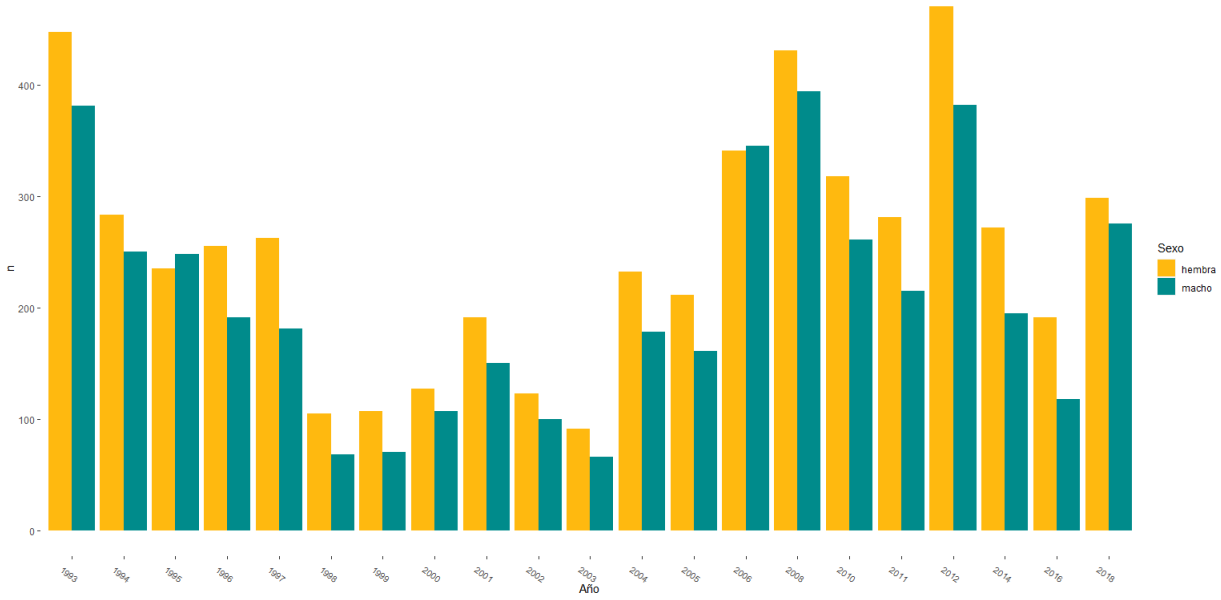


Figura 4.1: Número de predadores muestreados cada año, separados por sexos.

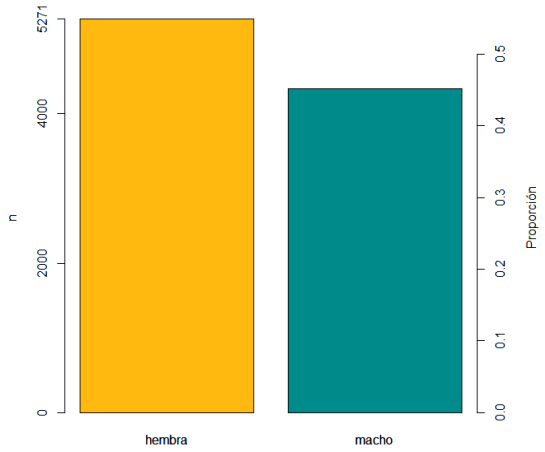


Figura 4.2: Número de predadores muestreados por sexo y proporción del total de datos que representan.

4.3, se puede observar que no se han tomado el mismo número de muestras para todas las tallas. Existe una asimetría positiva, con un mayor número de ejemplares de pequeño y mediano tamaño mientras que las tallas más grandes, a partir de 100 cm, el número de predadores es bastante reducido, debido a la menor abundancia. Las tallas menores de 13 cm no están representadas ya que no son muestradas. La multimodalidad de la curva, puede deberse a que al muestrear cada año/dos años, se vayan reflejando los cambios en el crecimiento de los individuos.

Lo mismo ocurre en la Figura 4.4, donde se puede observar la estimación de la densidad de las tallas de los ejemplares obtenidos cada año del muestreo. En la secuencia de densidades puede observarse

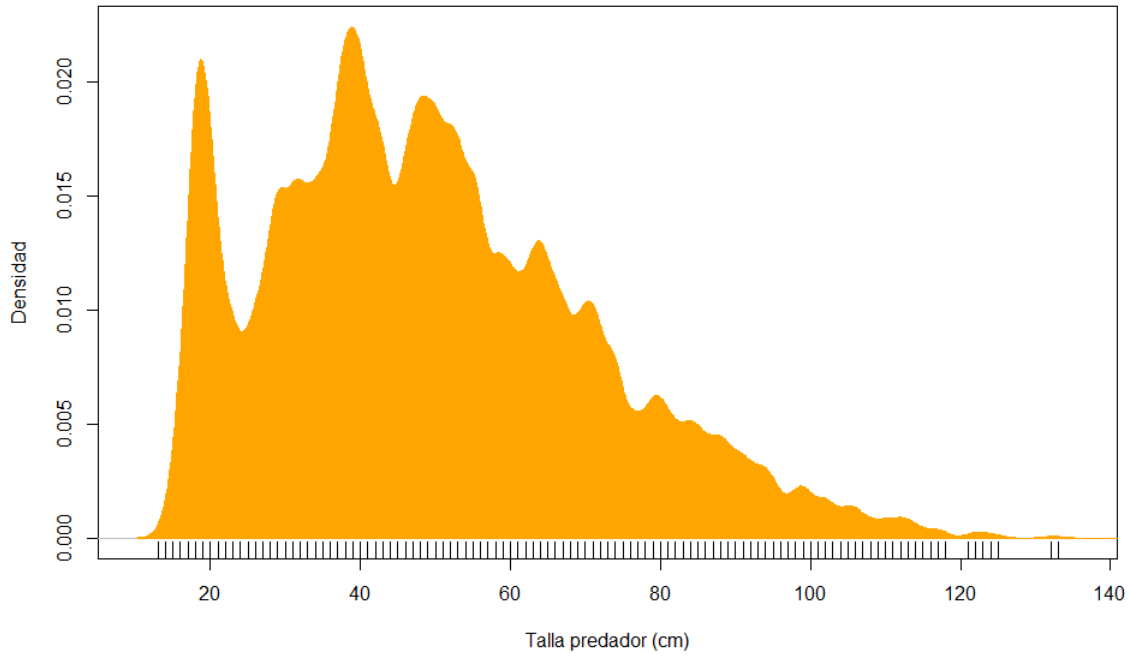


Figura 4.3: Estimación Kernel de la densidad para la talla de los predadores.

que la curva se va desplazando hacia la derecha. Esto ocurre de manera más acentuada en los años en los que la pesquería se encuentra en recuperación. A partir del año 2004 la curva empieza a aplanarse, lo que significa que se han muestrado individuos de más o menos todo el rango de tallas, efectivamente como corresponde a unos cambios en el diseño del muestreo con el objetivo de obtener similar representatividad de todas las tallas. No se puede afirmar que los datos reflejen los cambios fisiológicos de las poblaciones, o si por otro lado, es el muestreo el que lleva a estas asimetrías. Sin embargo, el hecho de que la recuperación de las poblaciones que conlleva a la reapertura de la pesquería y el hecho de que la curva se vaya aplanando parece que refuerce la idea de la recuperación de la pesquería y el crecimiento de la población.

En la Figura 4.5 se muestra la estimación de la densidad para la profundidad a la que se han realizado los muestreos. En este caso también se observa asimetría positiva y multimodalidad. La multimodalidad puede deberse a que los estratos del muestreo son intervalos y que por lo tanto ciertas profundidades dentro del intervalo estén más representadas que otras. Sin embargo la asimetría positiva puede estar más relacionada con que la presencia de ejemplares es menor al aumentar la profundidad donde se ha muestreado, haciendo de las pescas en profundidades mayores a 400 m algo anecdótico.

4.1.1. Conclusiones Análisis Exploratorio

Los datos disponibles no muestran la proporcionalidad y similar representación en función de las variables (tiempo, talla, sexo, profundidad, ...) deseada. Por otro lado, el número de ejemplares es reflejo de la abundancia y frecuencia de tallas, con estas características:

- Hay mayor número de individuos de las tallas más pequeñas, que son más abundantes.
- Hay un mayor número de hembras.

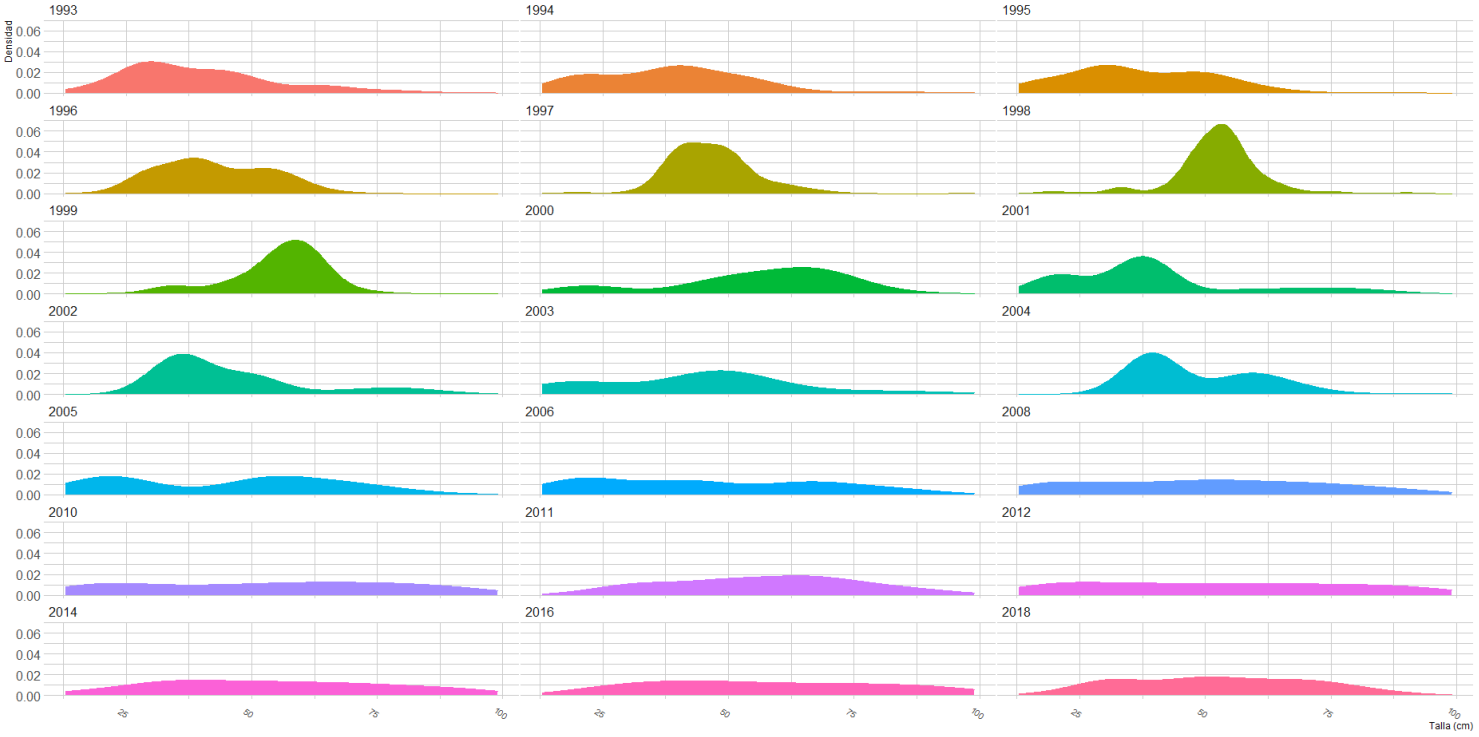


Figura 4.4: Número de predadores muestreados por talla cada año de estudio.

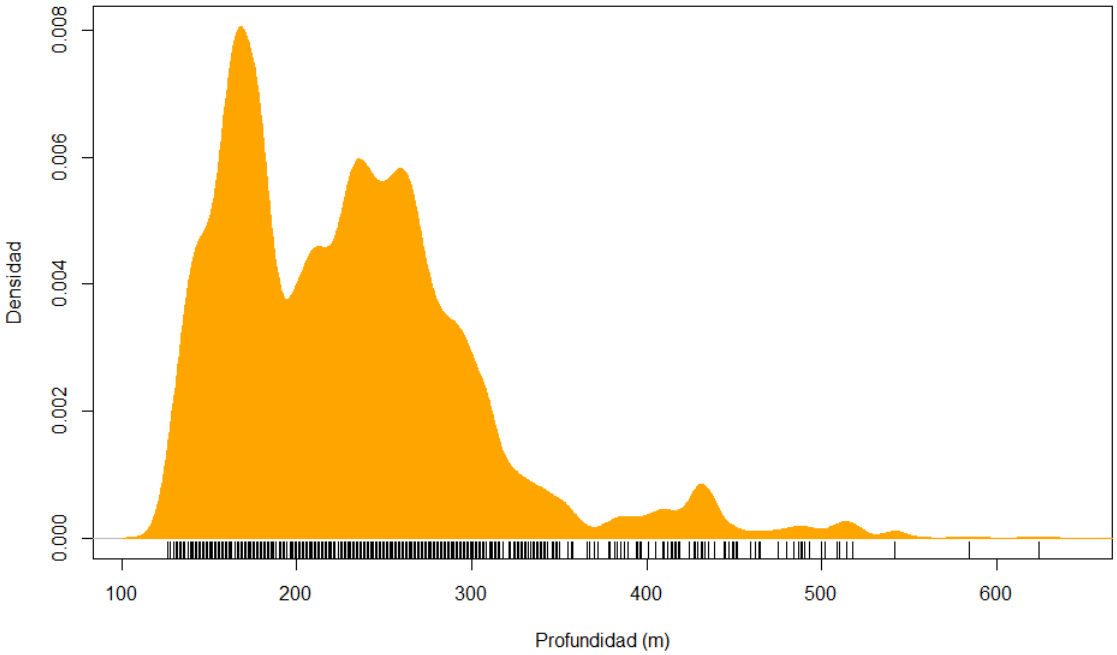


Figura 4.5: Estimación Kernel de la densidad para la profundidad de muestreo.

- Hay un menor número de individuos de la fracción poblacional de reclutamiento.
- Disminuye el número de individuos con la profundidad dado que el hábitat preferente de esta especie es a menos de 400 m.

4.1.2. Detección de datos atípicos

En este apartado se van a emplear diferentes técnicas para la detección de datos atípicos. La primera limitación que se encuentra en las diferentes técnicas presentadas en la Sección 3.4 es aquella que tiene que ver con la multidimensionalidad de los datos, ya que algunas técnicas solo sirven para comprobar cada variable de manera independiente.

Aplicar el diagrama de cajas y bigotes o bootlier permite únicamente comprobar la presencia de datos atípicos en una variable en concreto. En la Figura 4.6 se presenta este diagrama para las tallas de las presas, donde se detectan tres datos atípicos, empleando el algoritmo de bootlier sin embargo devuelve 0 datos atípicos. Sin embargo esta estrategia puede ser mejorada usando técnicas que permitan la comprobación de datos atípicos en más de una dimensión.

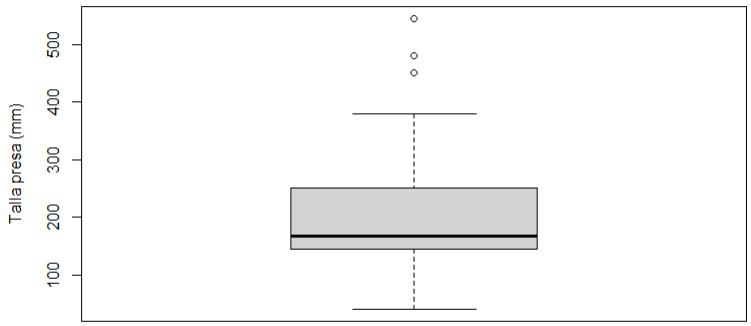


Figura 4.6: Diagrama de cajas y bigotes para la talla de la presa.

Tomando las variables de peso del predador y la talla del predador, graficadas en la Figura 4.7 se va a comprobar la presencia de este tipo de datos. Se han elegido estas variables porque las variables que relacionan la talla del predador con la presa presentan una alta variabilidad que confunde a las técnicas de detección de atípicos.

A continuación se va a ajustar un modelo que permita calcular la distancia de Cook para evaluar cada dato y tomarlo o no como una observación atípica.

Se parte de una distribución exponencial para después ajustar un modelo no paramétrico más exacto partiendo de unos valores iniciales. El modelo resultante y el ajuste es el presente en la Figura 4.8.

De esta forma se pueden calcular las distancias de Cook para cada dato 4.9, de acuerdo con la Ecuación 3.26. Se representan en rojo aquellos valores que cumplen $D_i > 1$ de acuerdo con los criterios marcados en el apartado 3.4.2.

Los resultados son demasiado numerosos así que se opta por seleccionar solo los datos que corresponden al cuantil 0.95. Al marcar en rojo esos puntos en el diagrama de dispersión de la Figura 4.10 se puede observar que los resultados no son muy buenos. Por un lado, existen algunos casos bastante obvios que sí que son identificados, pero que tampoco resulta en un resultado mucho mejor que el que se podría hacer tras una identificación visual. Además, parece que por un lado identifica los casos más externos al modelo (probablemente debido a que los residuos aumentan en función de la talla del predador). Por otro lado marca también un conjunto de datos correspondientes a las tallas y pesos más grandes, lo que podría deberse a que esos datos se encuentran más aislados simplemente por ser menos probable encontrar bacalaos de gran tamaño y peso por ser estos menos abundantes.

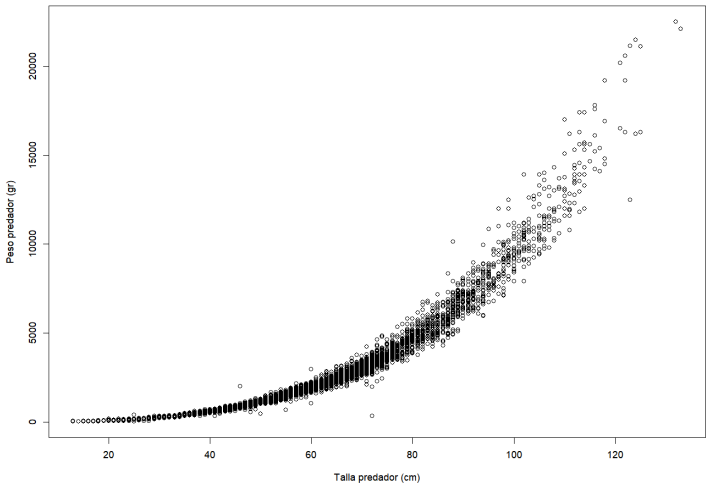


Figura 4.7: Diagrama de dispersión del peso de los predadores frente a su longitud.

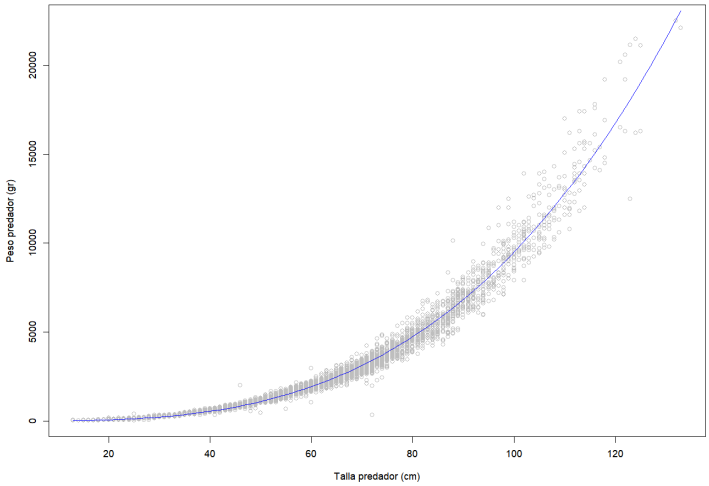


Figura 4.8: Diagrama de dispersión con ajuste de modelo exponencial.

Los resultados obtenidos con *Isolation Forest* se representan en la Figura 4.11, donde parece que se realiza una discriminación de datos atípica más acertada.

Finalmente los resultados obtenidos con *HDOutliers* se representan en la Figura 4.12. En comparación con el resto de métodos parece que se obtienen los resultados más fiables, ya que se eligen aquellos datos que distan del conjunto incluso visualmente, así que serán estos los datos que se saquen del conjunto para proseguir con el análisis.

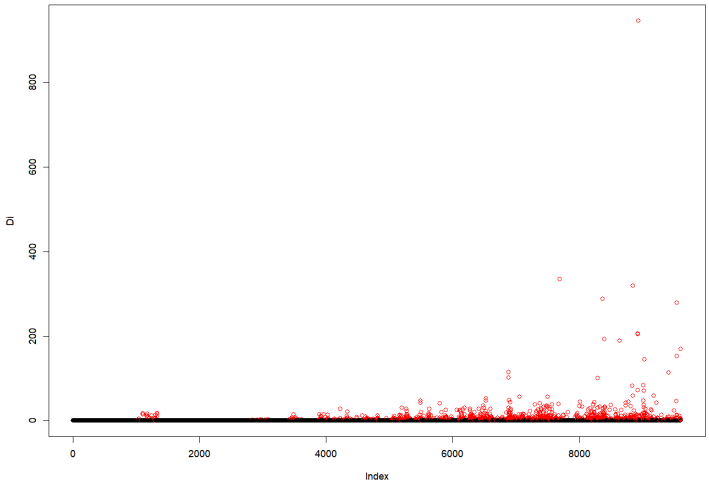


Figura 4.9: Distancias de Cook.

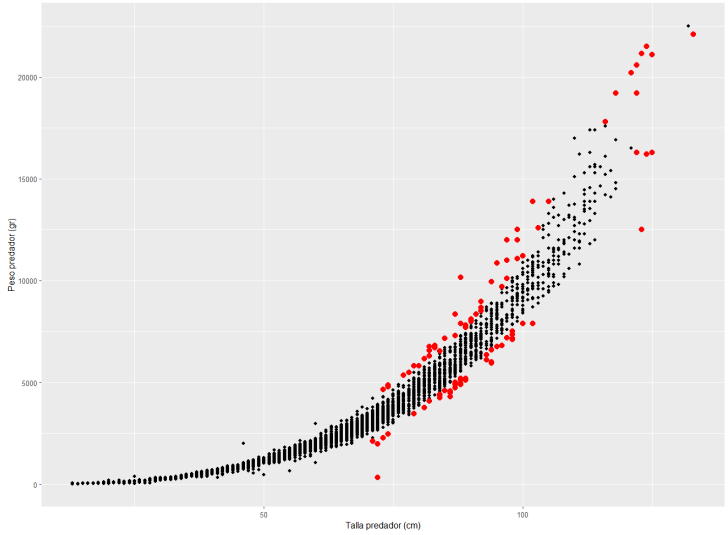


Figura 4.10: Atípicos representados.

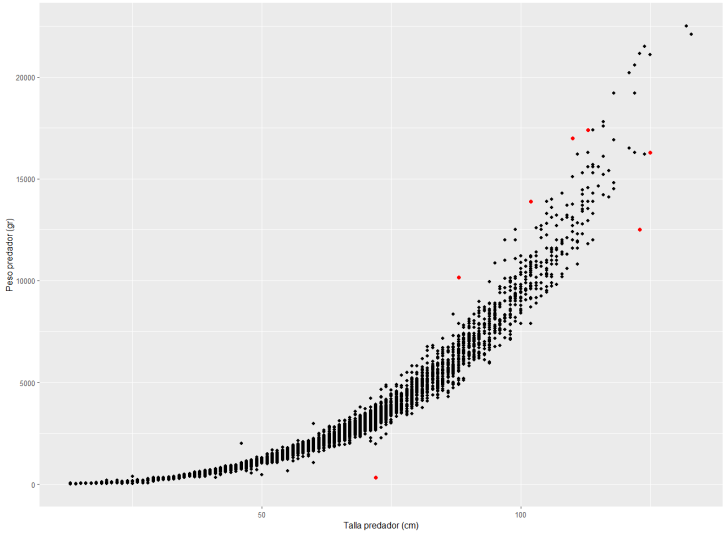


Figura 4.11: Atípicos representados utilizando *Isolation Forest* con parametro $s = 0{,}75$.

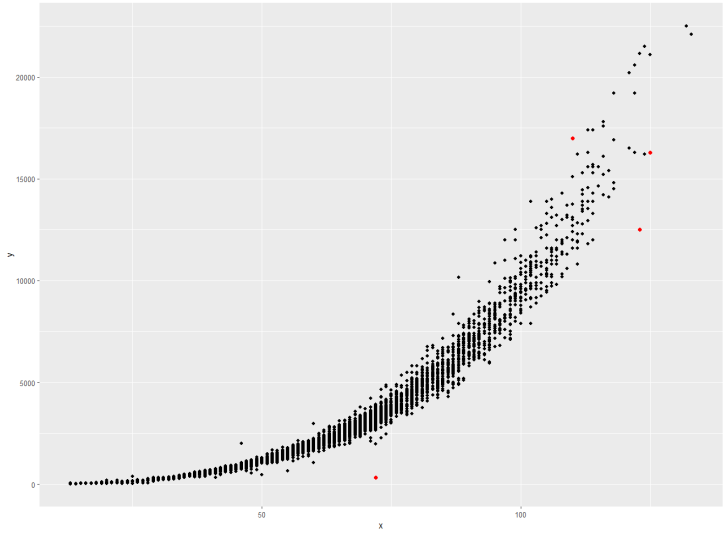


Figura 4.12: Atípicos representados utilizando *HDOutliers*.

4.2. Índices para el análisis de alimentación de peces

Antes de empezar con el análisis es necesario definir el índice con que trabajar. Se puede definir "individuo de interés" como aquella presa que cumple ciertas características en lo que a una "característica de interés" se refiere, por ejemplo: que pertenece a cierta especie, que tiene cierta medida, que pertenece a un predador muestreado a cierta profundidad, etc.

Hyslop (1980) [36] hace una revisión de los métodos de análisis descriptivo usualmente utilizados en el análisis de contenidos estomacales, de donde podemos definir los siguientes índices:

1. **Frecuencia de ocurrencia(%FO):** conteo de veces que un estómago contiene uno o más individuos de interés, normalmente expresado como porcentaje del total de estómagos. Tiene la ventaja de ser rápido y de no requerir ningún aparato específico, sin embargo falla en representar la importancia de cada individuo repecto del predador que lo ha ingerido.

2. **Numérico (%N):** conteo del número de individuos de interés que aparecen en los contenidos estomacales, que usualmente se expresa como porcentaje respecto del número total de individuos. Contar los individuos es relativamente rápido y simple si la identificación de las presas es factible, ya que esta tarea puede complicarse debido al estado de digestión. Puede ser más tedioso en el caso de microorganismos, por la necesidad de realizar submuestras. Es especialmente apropiado si las presas se encuentran en el mismo rango de tallas, y este rango es lo suficientemente pequeño.

3. **Volumétrico (%V):** cálculo del volumen total de un individuo de interés, normalmente expresado como porcentaje del volumen total encontrado en los contenidos estomacales analizados.

4. **Gravimétrico (%W):** peso que representa un individuo de interés respecto del peso total encontrado en los contenidos estomacales de esa característica. Un tipo de índice gravimétrico más informativo es el "Mean Weight Fullness Index´´:

   – **Mean Weight Fullness Index (MWFI):** fracción de la suma de pesos de los individuos de interés presentes en el contenido estomacal de un depredador entre el peso de ese mismo depredador.

$$FI = \frac{\sum_{i=1}^{N} \frac{Wsc_i}{Wp_i}}{N} \cdot 100 \tag{4.1}$$

   donde,

   Wsc_i = peso total del contenido estomacal en el predador i

   Wp_i = peso del predador i

   N = número total de predadores

El primer y segundo índice tienen la desventaja de no considerar el tamaño del predador, por lo que dan la misma importancia a una presa pequeña que a una de gran tamaño: obtiene el mismo valor, por ejemplo, un hipérido (un crustáceo, considérese "pequeño") que una presa de pez como sebastes o bacalao. Las presas pequeñas son consumidas más habitualmente y en mayor número, por lo que al final éstas reciben una importancia mayor.

El tercer y cuarto índice, sobrevaloran las presas más grandes, pero sin embargo son más acertados por presentar una relación más directa con el valor calórico de las presas. Además, en el caso del WFMI, se tiene en cuenta también de alguna manera el tamaño del predador a través de su peso.

En este estudio se mostrarán resultados usando la mayoría de la índices arriba indicados, aunque preferentemente se hará uso de MWFI por ser un índice que minimiza el efecto de la talla (variable que afecta a la cantidad de consumo que el individuo necesita para cubrir sus necesidades energéticas.) Sin embargo, la representación de los distintos índices nos informa de diferentes aspectos de la dieta reseñables.

Los contenidos estomacales de los individuos de bacalao muestreados a lo largo del periodo de estudio reflejan hasta 198 items diferentes predados, es decir, un amplio espectro específico (ancho de nicho ecológico), aunque el alimento mayoritario es a base de un menor número de especies y que generalmente muestran gran abundancia y disponibilidad. A continuación se ilustran los resultados obtenidos con los diferentes índices descritos:

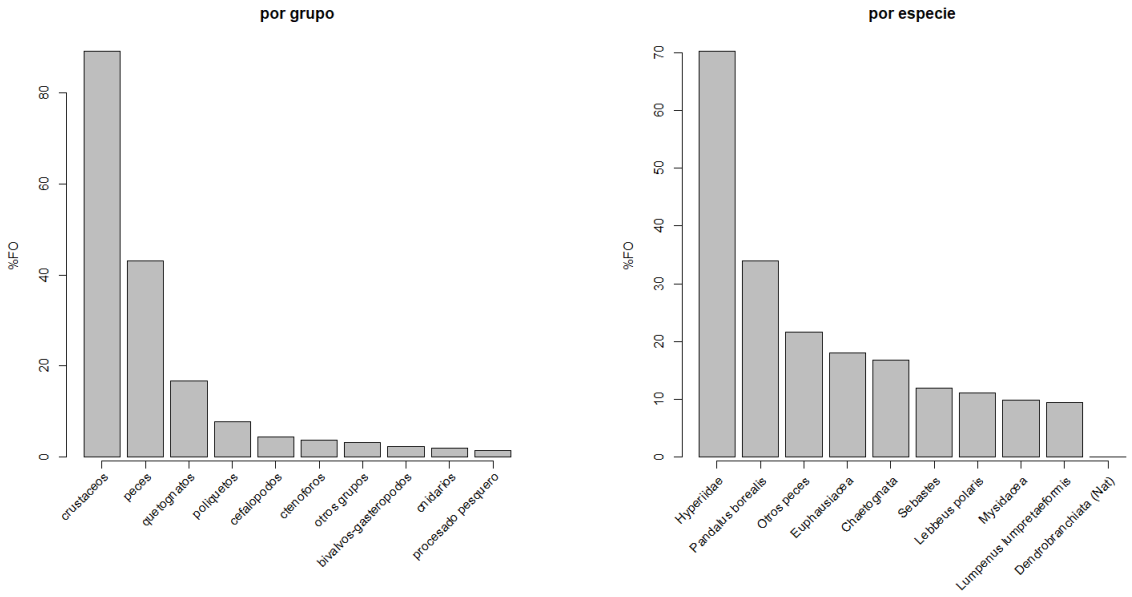


Figura 4.13: %FO de los registros según grupo y especie de presa.

Los diagramas de barras de la **Figura 4.13** reflejan la **frecuencia de ocurrencia (%FO)** de los 10 grupos y 10 especies de presas más frecuentes. En un mismo estómago puede encontrarse más de un grupo/especie de presas. Como era de esperar, los individuos de menor tamaño y mayor abundancia son los que mayor importancia obtienen según este índice. Del subfilo de los crustáceos (presentes en el 80.31 % de los estómagos) son los individuos de la familia de los hipéridos los más frecuentes, además de presas de especies como *Pandalus borealis* y *Lebeus polaris*, e individuos de otras categorías taxonómicas más grandes como Euphausiacea, Mysidacea y Dendrobranchiata. Los peces son el segundo grupo con mayor frecuencia de ocurrencia (43.16 %). La categoría "otros peces", que aúna peces digeridos, peces no identificados y peces identificados pero que son ingeridos con una frecuencia muy baja, presenta como es lógico una frecuencia de ocurrencia alta. Los individuos del género *Sebastes* y de la especie *Lumpenus lumpretaeformis* son los peces con mayor frecuencia de ocurrencia. La especie *G. morhua*, que no está entre las 10 especies con mayor FO, está presente en tan solo el 3.31 % de los estómagos analizados, por lo que este índice no le otorga demasiada relevancia.

El problema de dar mayor importancia a las presas consumidas más frecuentemente y que son generalmente más pequeñas, es más evidente al utilizar el método **numérico (%N)**, ya que no solo cuenta la presencia/ausencia, sino que también el número de individuos consumidos.

Los diagramas de la **Figura 4.14** muestran el **índice gravimétrico (%W)** de las 10 presas con mayores valores según su grupo de presa y según la especie. En este caso los peces son el grupo con mayor relevancia, representando el 53.58 % del peso total encontrado en los estómagos, seguido por los crustáceos que representan el 30.2 %. Los resultados cambian completamente aquellos de la Figura 4.13. El procesado pesquero toma también una relevancia mencionable, con hasta el 12.81 % del peso total aunque el número de ingestas es muy reducido; en este sentido, la relevancia que tiene el procesado pesquero está sobrevalorada. En cuanto a las especies se pueden apreciar en los primeros puestos individuos del grupo *Sebastes* y en el cuarto, el propio *G. morhua* (consecuencia de comportamiento de canibalismo en la alimentación).

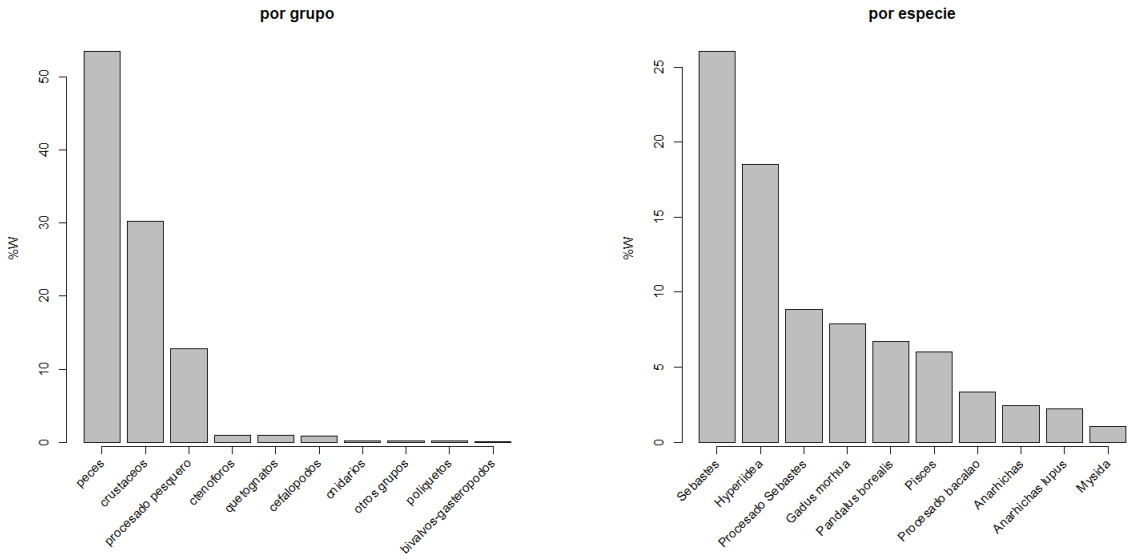


Figura 4.14: %W de los registros según grupo y especie de presa.

A continuación se muestra el índice **MWFI**, en función de ciertas variables de interés.

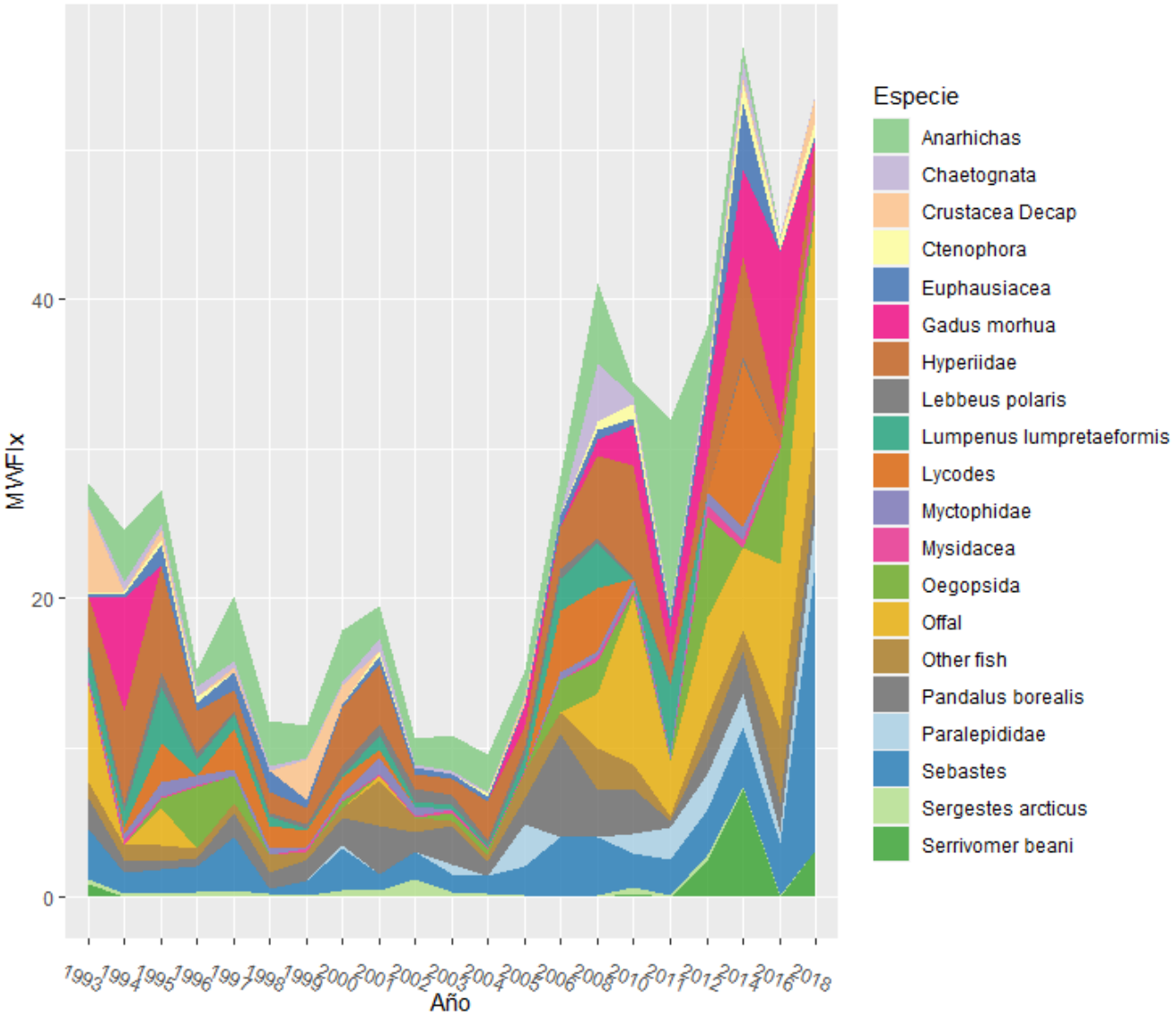


Figura 4.15: MWFI por especie de presa y año.

En la Figura **4.15** se muestra el índice mencionado para el predador bacalao en **función del año y separado por especies de presa**. Es decir, en este caso el **individuo de interés** es aquel que pertenece a cada especie de las que aparecen en la leyenda, y la **característica de interés** es el año. Por un lado se puede observar que el índice total, cambia a lo largo del tiempo. Donde la intensidad de alimentación entre 1993 y 1999 es menor. Solo se puede observar predación sobre el propio bacalao en los primeros años de este periodo, después en los años en los que la población estaba en declive no se observa en la dieta, lo cual muestra la validez de estos estudios como bioindicador. También se aprecia en esos primeros años 'offal'(restos de procesado pesquero) casquería, utilizados por la industria pesquera, que desaparece y vuelve a aparecer coincidiendo con la mayor o menor actividad pesquera. El índice empieza a crecer otra vez en 2004-2005, coincidiendo con el inicio de la recuperación de la recuperación de la población en Flemish Cap, y vuelve a subir a partir de 2008-2009 coincidiendo con el cese de la moratoria y reapertura de la pesquería.

En la Figura **4.16** se muestra el mismo índice pero siendo esta vez la característica de interés la profundidad (en intervalos de 100 metros) a la que se han capturado los ejemplares de bacalao, siendo la profundidad mínima 126 m y la máxima 624 m. Este gráfico muestra que el rango de profundidad donde más se alimentan los bacalaos es entre los 100 m (126 m en realidad por los datos que se tienen) y hasta los 399 m. Parece lógico poder relacionarlo con la distribución en profundidad de las presas. Se puede pensar que en el intervalo en el que mayor es el índice, mayor presencia de presas hay. Por lo general, la disponibilidad de presas disminuye con la profundidad, por ejemplo *Sebastes* parece ser de las únicas presas presentes en profundidades mayores a 500 m.

Pero también va unido a dónde se distribuye la fracción de la población del bacalao que ejerce mayor actividad predadora, es decir por su talla o edad, precisa mayor alimento para cubrir sus necesidades energéticas. Tal y como ocurre en la mayoría de las especies los cambios de intensidad de alimentación varía a lo largo de la vida de los individuos, lo que se puede observar en la Figura 4.17.

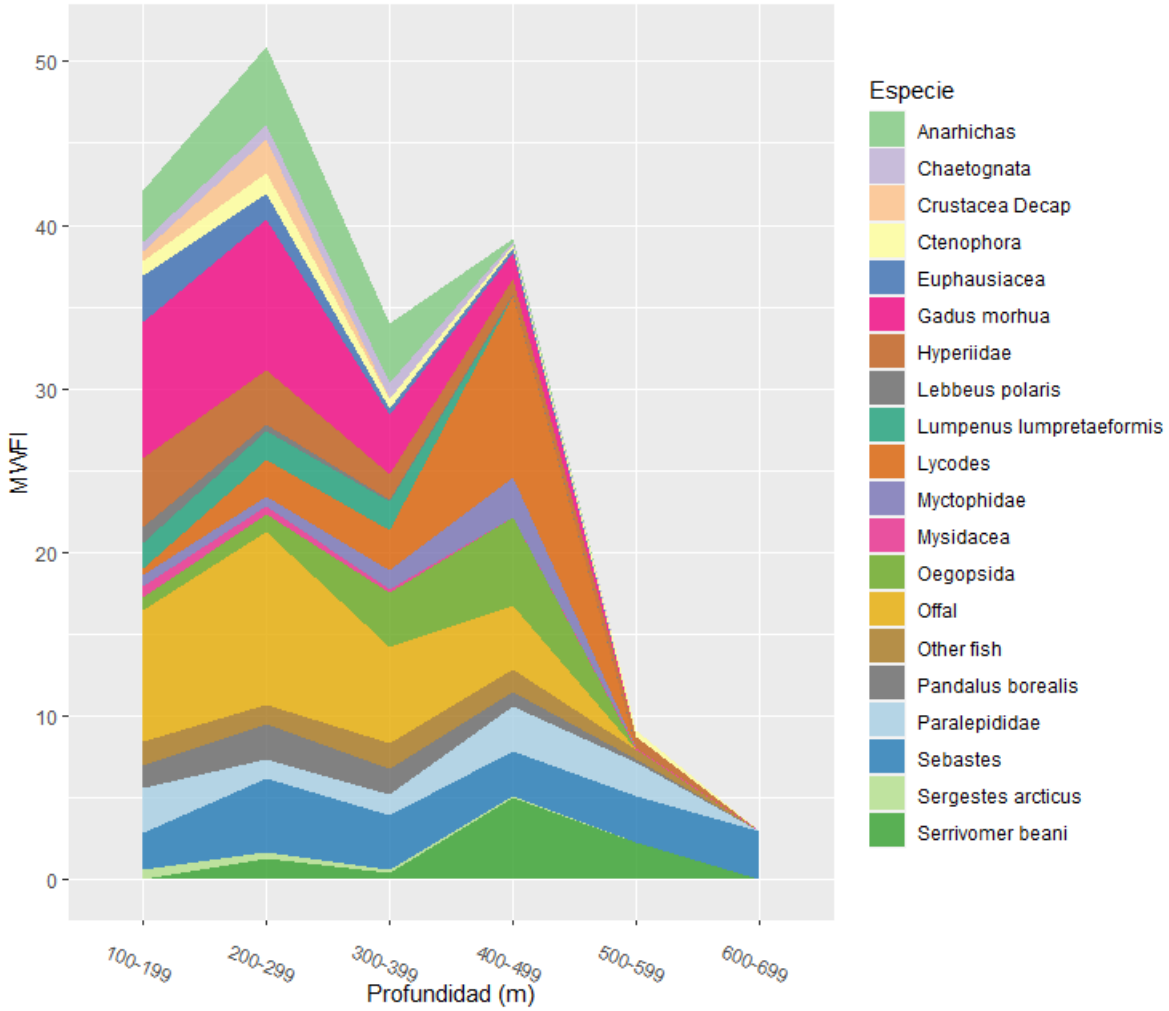


Figura 4.16: MWFI por especie de presa y profundidad.

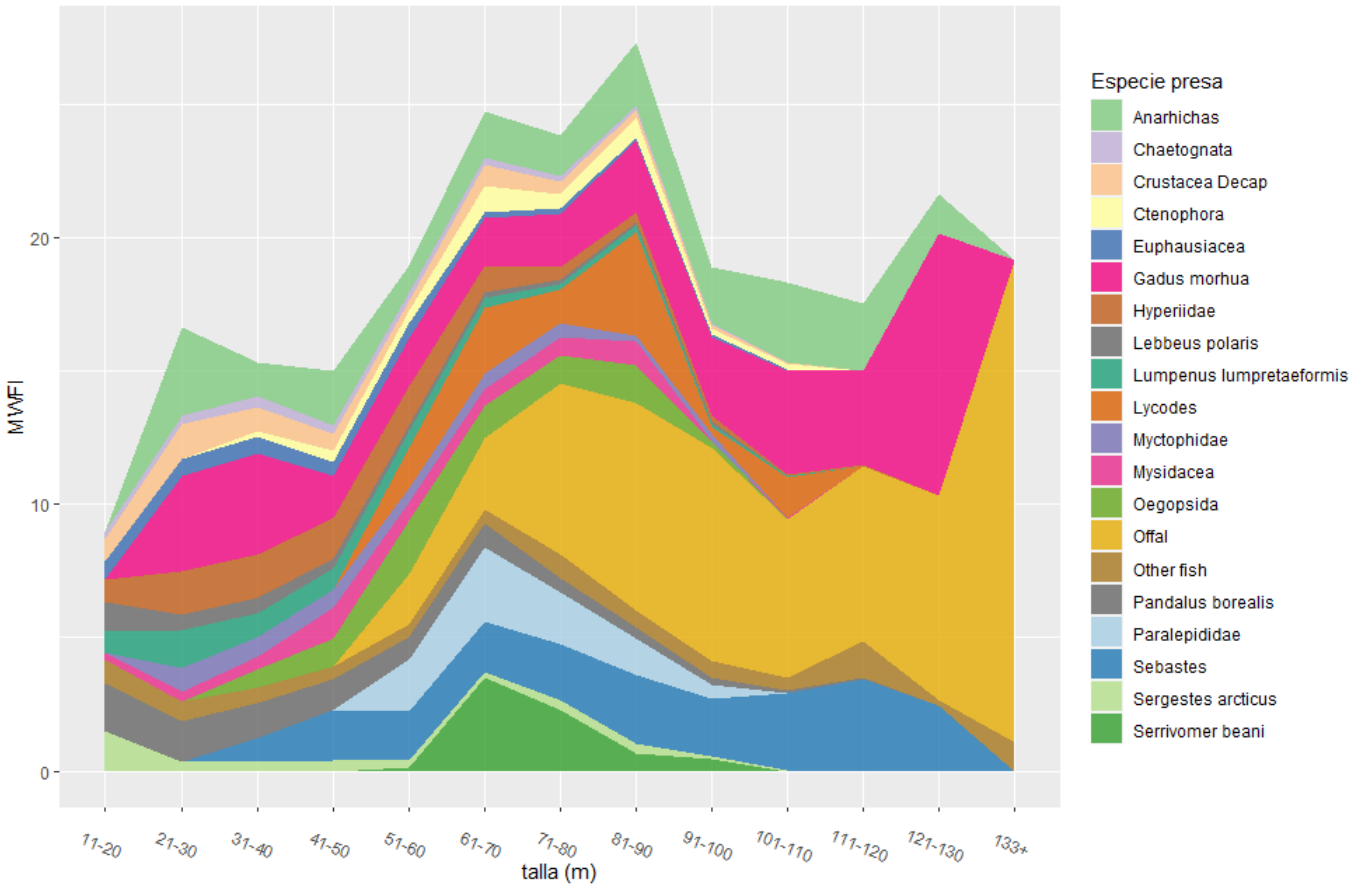


Figura 4.17: MWFI por especie de presa y talla.

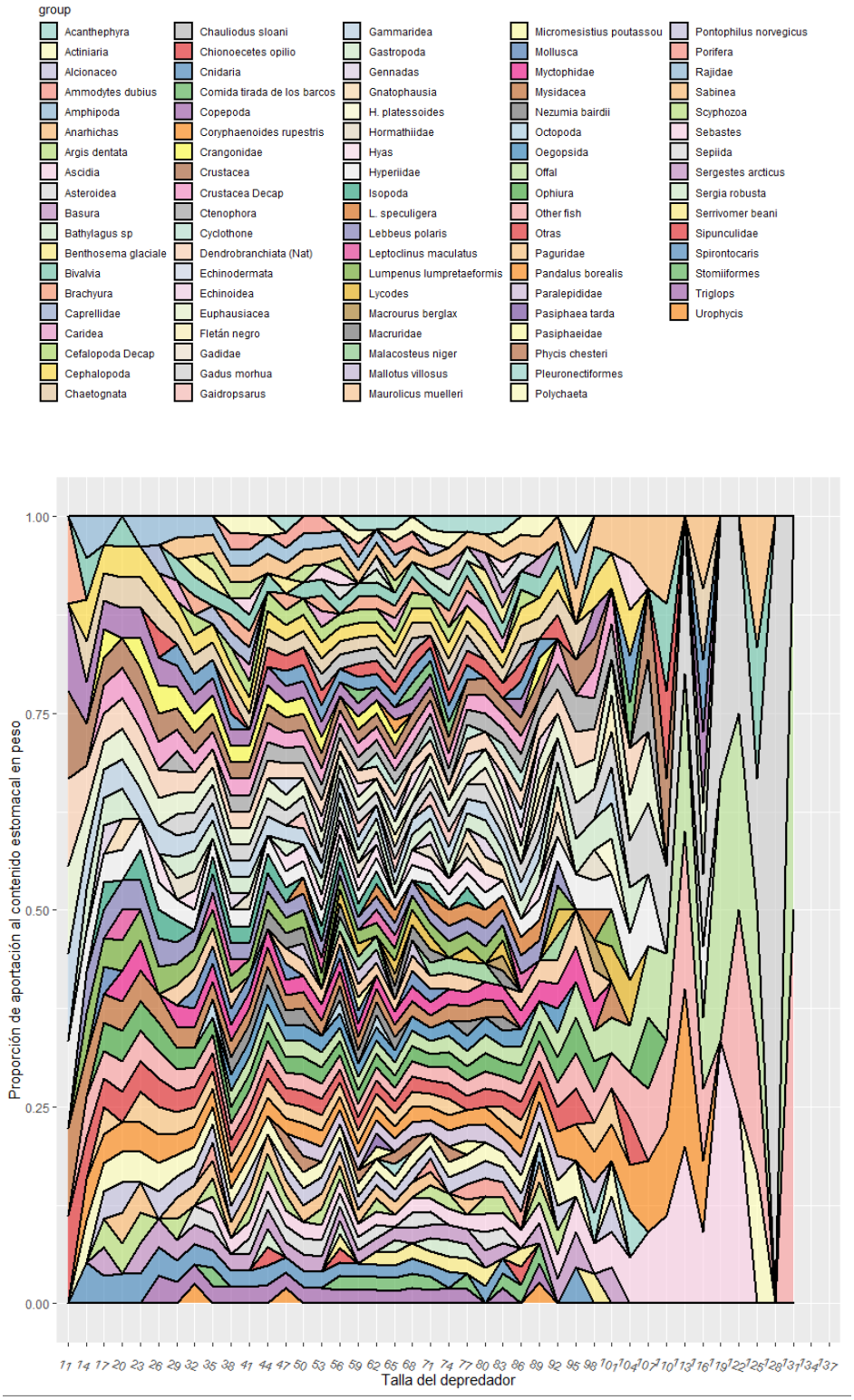


Figura 4.18: Proporción de aportación a la dieta por especie en función de la talla del predador.

La dieta del bacalao se compone de más de 200 especies (Zatsepin y Petrova 1939 [37]); Mehl 1991 [38]), pero normalmente se alimentan en función de la abundancia y disponibilidad de la presa, siendo los peces la presa preferida (en peso). Sin embargo, al analizar los datos por tallas se pueden ver diferencia relacionadas con el tamaño del depredador 4.18. Se observa como *G. morhua* o *Sebastes* no forman parte importante de la dieta de los bacalaos hasta que éstos alcanzan un tamaño considerable. Es interesante observar la propoción de aportación al contenido estomacal en peso porque se puede la gran amplitud del nicho alimenticio del bacalao, que se compone prácticamente de todo aquello que pueda ingerir.

Las conclusiones de este apartado son las siguientes:

- Las principales presas en la alimentación del bacalao en Flemish Cap son: hypéridos, camarón boreal y gallineta.
- La importancia de estas presas puede interpretarse de manera diferente en función del índice utilizado.
- La alimentación muestra variaciones a lo largo de la serie analizada.
- La dieta varía con la talla del predador y la profunidad de distribución. Ambos factores cambian anualmente, lo cual influye también en los cambios anuales.
- El MWFI proporciona mayor información minimizando el efecto de la talla.
- La variación batimétrica, ligada a la estacional y por lo tanto de temperatura se minimiza con el protocolo de muestreo en la misma época; auqnue se pierde información del cambio trófico a lo largo del año, cuando es sabido que la alimentación e insidad es muy diferente de verano a invierno.

4.3. Pauta de predación del bacalao sobre presas de su misma especie

Conocer el comportamiento de los individuos derivados de los cambios ontogénicos en respuesta a las fases vitales, es fundamental para conocer la dinámica de la población y los efectos esperables de la dinámica trófica en el sistema. La relación de tallas entre predador y presa, es decir, qué fracción de la población (comprendido entre un rango de tallas) ejerce mayor incidencia predadora y sobre qué fracción de la población (mayoritariamente también comprendido en un rango de tallas).

El diagrama de dispersión de Figura 4.19 que muestra que esta relación de tallas entre las presas y los predadores (ambos *G. morhua*), y al que se le ha ajustado una recta de regresión. A grandes rasgos apreciamos que este hábito alimenticio se incrementa en los individuos mayores de 50 cm; la mayoría de las presas estarían comprendidas entre 10 y 40 cm; y el rango de tallas de los individuos predados incrementa con la talla del predador.

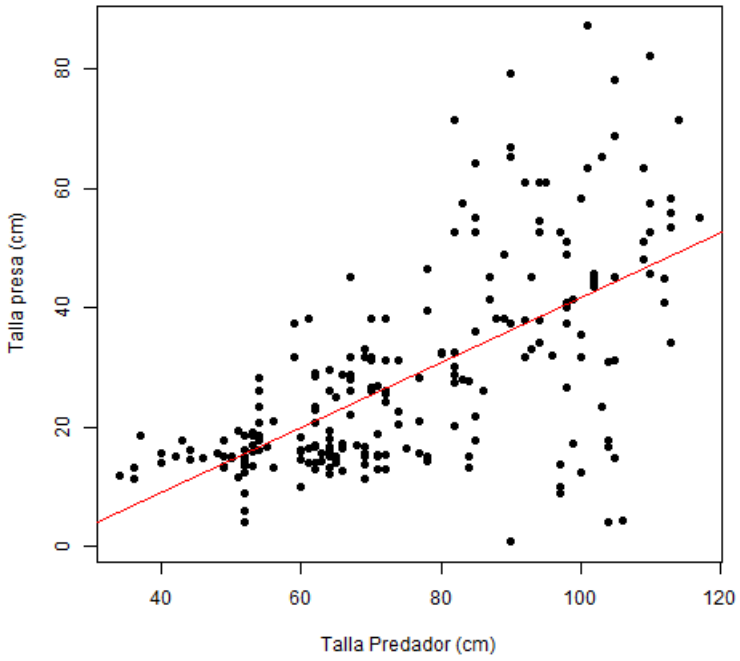


Figura 4.19: Ajuste de un modelo lineal a la relación entre la talla de la presa y la del depredador.

Al diagrama de dispersión se le ha ajustado un modelo lineal simple utilizando mínimos cuadrados ordinarios:

- $Y = -12{,}45X + 0{,}54090$

donde ambos coeficientes resultan significativos, con respectivos p-valores del t-test $3{,}27e-4$ y $< 2e-16$. El modelo, con un valor del estadístico F de 153,5 con 1 y 230 grados de libertad, obtiene un p-valor $< 2{,}2e-16$. Por lo tanto el modelo es significativo. Aunque el modelo explique a grandes rasgos la preferencia por tamaño del bacalao, se obtiene un coeficiente de determinación $R^2 = 0{,}4$, por lo que el modelo no es capaz de describir una gran parte de la variabilidad presente en los datos.

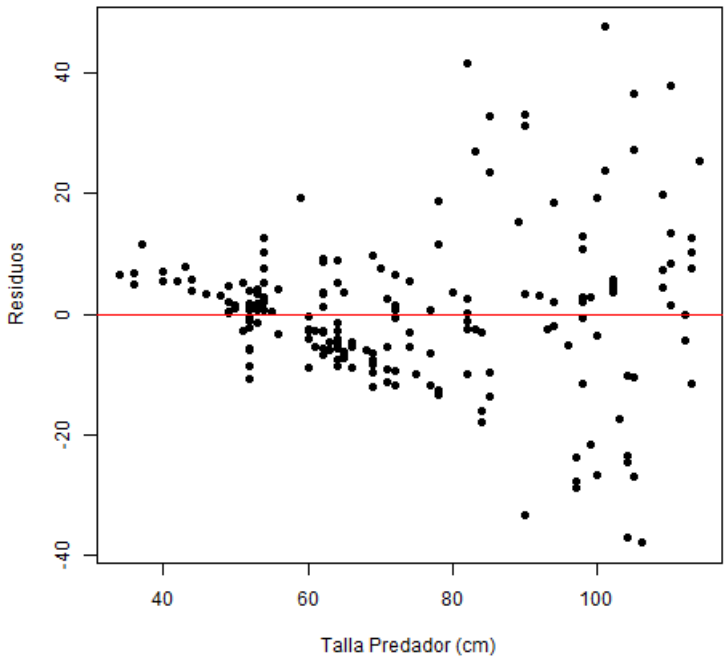


Figura 4.20: Análisis de los residuos del modelo.

El análisis de los residuos no es satisfactorio, aunque éstos cumplan la hipótesis de normalidad, confirmado por el test de Normalidad de Shapiro-Wilk p-valor= $5{,}161e-07$ y Jarque-Bera p-value = $1{,}364e-09$, no cumplen las hipótesis de homocedasticidad e independencia.
En la Figura 4.20 se puede observar como la magnitud de los errores es mayor a medida que aumentan la talla del depredador, lo que indica que los residuos presentan homocedasticidad.
Por otro lado, el text de Ljung-Box rechaza la independencia, con un p-valor $6{,}291e-12$.

Transformar los datos es una estrategia usual cuando las hipótesis básicas del modelo no se cumplen, en este caso se opta por la transformación Box-Cox que supone un método automático para seleccionar el tipo de transformación a realizar. Este método, en principio, selecciona un tipo de transformación para que la distribución de los datos se acerque a la Normal, y aunque los datos ya cumplan la condición de Normalidad, puede ayudar también a que los datos transformados presenten la misma o similar varianza. En la Figura 4.21 se observa que el máximo de la función de verosimilitud se alcanza en el valor $\lambda = 0{,}34$, lo que indica que transformar la variable y como $y = y^{0{,}34}$ es la transformación más aconsejable.

Con los datos transformados se vuelve a ajustar un modelo lineal utilizando mínimos cuadrados ordinarios, visible en la Figura 4.22.

Como se puede apreciar en la Figura 4.23, no se ha solucionado el problema de la heterocedasticidad, por lo que la transformación no ha conseguido el objetivo.

Se podría pensar en realizar otro tipo de ajustes que no fueran un modelo lineal, en la Figura 4.24 donde se han incluido, además del primer modelo tratado, un estimador de Nadaraya-Watson y un estimador polinómico local. Sin embargo estos modelos no aportan resultados mucho mejores que la regresión lineal.

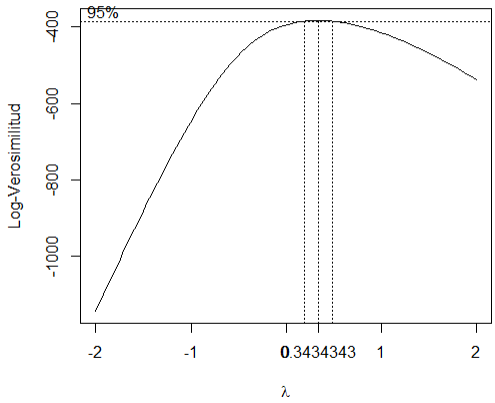


Figura 4.21: Valor de lambda en el que se alcanza el máximo de la función de verosimilitud para la transformación BoxCox.

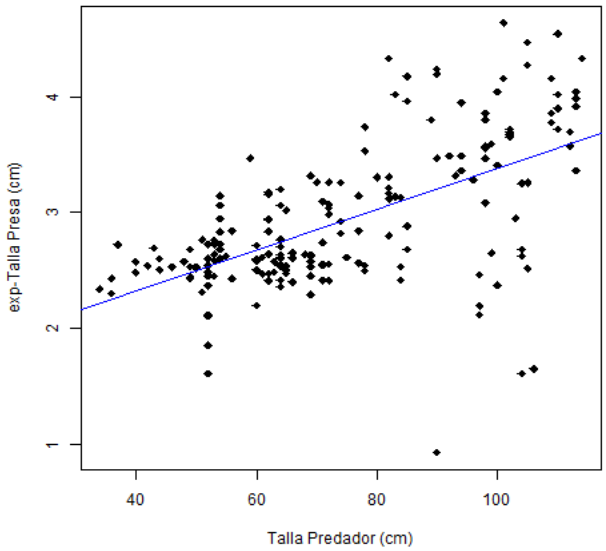


Figura 4.22: Ajuste de un modelo lineal a la relación entre los datos transformados de la talla de la presa y la del depredador.

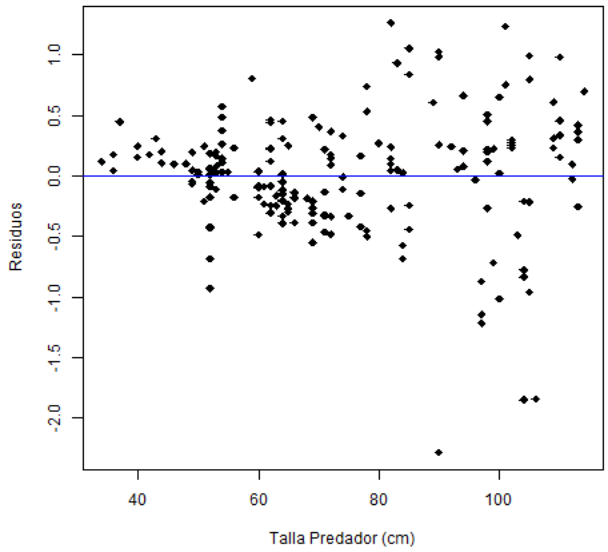


Figura 4.23: Análisis de los residuos del modelo ajustado a los datos transformados.

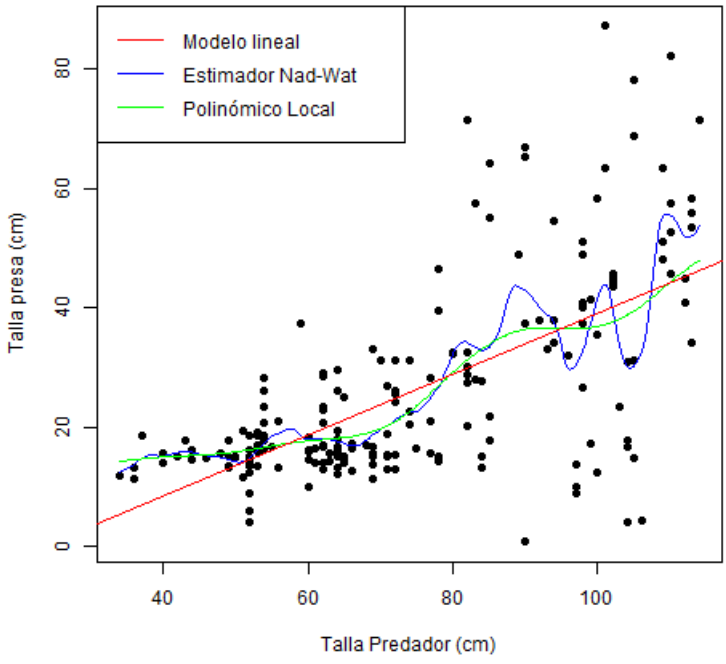


Figura 4.24: Diferentes ajustes a los datos

En comunidades acuáticas, el tamaño de de la presa y del predador es un atributo que se conecta directamente con el éxito en el forrajeo. Este éxito de los predadores aumenta a medida que aumenta su tamaño debido a varios factores: mayores velocidades de nado sostenido y explosivo y mayor agudeza visual y mayor tamaño de boca derivado principalmente de cuestiones morfológicas como mayor tamaño corporal (Keast & Webb 1966 [39], Webb (1976) [40], Beamish 1978 [41]).

A su vez, la respuesta de escape de la presa también aumenta con su tamaño, ya que la distancia de reacción aumenta y el nado mejora con el tamaño. (Folkvord & Hunter 1986, Blaxter & Fuiman 1990).

Para la mayoría de los peces y el bacalao no es una excepción, el tamaño de la presa aumenta con el tamaño del predador. (Keast & Webb 1966 [39], Popova 1967 [42], Nielsen 1980 [43], Persson 1990 [44], Juanes 1994 [45]).

Por lo general, el rango de tallas predado es mayor en predadores grandes, incrementándose el tamaño de presa máximo mientras que el tamaño de presa mínimo apenas cambia.

Siguiendo la estrategia de Scharf y Juanes y Rountree (2000) [46], se va a utilizar la regresión cuantil para examinar la anchura del espectro alimentario (en cuanto a rango de talla foco del canibalismo).

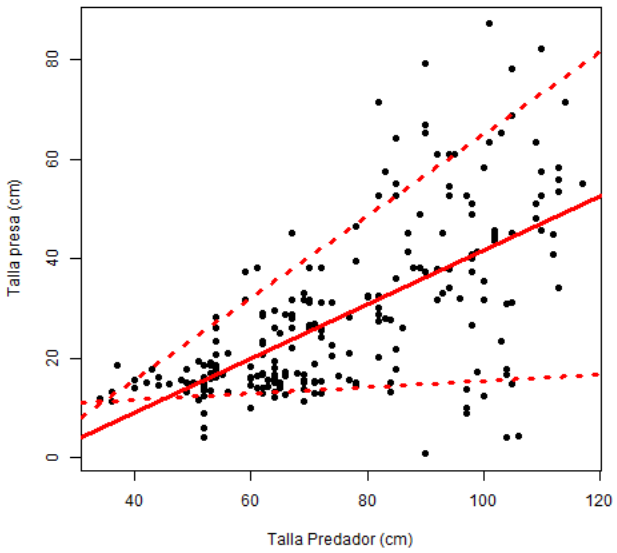


Figura 4.25: Ajuste de modelo lineal para la media. Ajuste de modelos a partir de regresión cuantil para el máximo y el mínimo.

Mínimo	Mediana	Máximo	n	Cuantiles
Y= 0.035X+10.44	Y= 0.5X-11.91	Y=0.75X-15.03	232	10/90

Cuadro 4.1: Resumen del ajuste.

La Figura 4.25 ilustra cómo se cumple lo indicado anteriormente. La talla mínima del bacalao consumido apenas cambia, sí lo hace, sin embargo la talla máxima, indicando que efectivamente, el rango de tallas consumido crece a la vez que crece el depredador, consecuencia de las ventajas que

supone el aumento de tamaño y también obligados por la mayor necesidad energética. El rango de tallas de las presas consumidas se expande con el aumento de tamaño del predador. El hecho de que las presas pequeñas sigan siendo parte de la dieta de los individuos cuando crecen, da lugar a una distribución asimétrica de talla predador - talla presa.

4.3.1. Pauta en función del sexo

El siguiente paso es aplicar técnicas para ver si los depredadores tienen conductas de consumo diferenciables según variables como el sexo o la talla.

El sexo es una variable factorial (macho o hembra) y se puede contrastar su efecto Mediante el Análisis de la Varianza (ANOVA), es decir, contrastar la hipótesis nula de que el tamaño medio de presa es igual para los predadores macho y los predadores hembra.

$$H_0 : \mu_{hembra} = \mu_{macho}$$
$$H_1 : \mu_{hembra} \neq \mu_{macho},$$

El ANOVA requiere de las hipótesis básicas de Normalidad, independencia y homocedasticidad para que sus resultados sean concluyentes. La suposición de normalidad no se cumple para los datos de las tallas de las presas (Test de Normalidad de Shapiro Wilk, W=0.8426, p-valor= $3{,}24e^{-}13$), por lo que será mejor utilizar un test no paramétrico para realizar la comprobación.

En lugar del ANOVA, ya que falla la Normalidad, se opta por utilizar métodos no paramétricos.

El test, por lo tanto se encarga de determinar si las diferencias observadas en la Figura 4.26, por un lado la diferencia en el valor de mediana y por otro las diferencias de distribución, son o no significativamente diferentes.

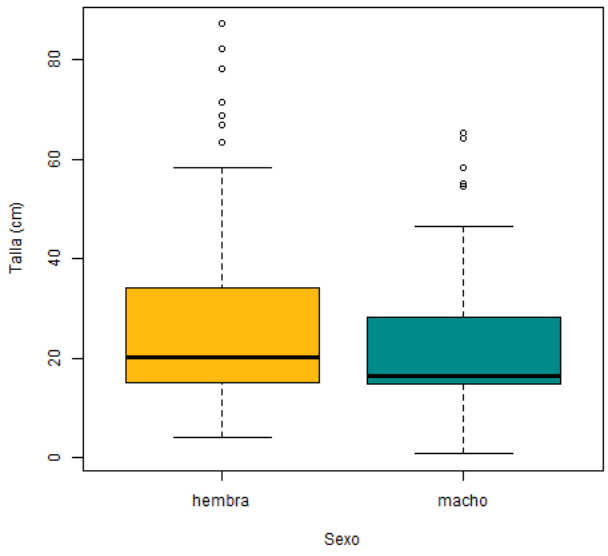


Figura 4.26: Boxplot para la Talla (cm) de presa por sexos

Utilizando el test de Kruskal-Wallis [47], se obtiene un $p-valor = 0{,}128$ por lo que no hay suficientes evidencias para rechazar la hipótesis nula a favor de la alternativa, es decir, no existen diferencias suficientemente significativas como para decir que exista un comportamiento diferente entre machos y

hembras en lo que se refiere a la predación por tallas de la presa. Esto hecho resulta interesante porque simplifica en gran medida la evaluación del impacto por canibalismo.

4.3.2. Clustering

Para identificar si existen rangos de tallas de predador en los que se aprecien preferencias de tallas de presa similares, es decir, un comportamiento parecido a la hora del forrajeo, se van a emplear diferentes algoritmos de agrupamiento no supervisado (En inglés *Clustering*). No supervisado significa que no se tiene más información que los datos a analizar, es decir, no se tienen etiquetas o clasificaciones previas con las que poder realizar un aprendizaje que se sepa que es correcto.

Un clúster es un conjunto de observaciones que son más similares entre ellas que lo que lo son a observaciones en otros clústeres. Y se denomina *Clustering* a las técnicas que intentan identificar estructuras de agrupamiento en un conjunto de datos [48] para construir los clústeres.

4.3.2.1. k-means

A tener en cuenta

La técnica *k-means* se basa en la búsqueda de mínimos locales en vez del mínimo global, por lo que puede resultar en particiones no-óptimas. Además de esto, estas técnicas tienden a crear clústeres del mismo tamaño y que suelen a ser esféricos en n-dimensiones (donde n es la dimensión del espacio de trabajo). Esto significa que los clústeres serán esféricos en dos dimensiones, una esfera en tres dimensiones y una hiperesfera en tres o más dimensiones. Por lo que funcionan mejor si los clústeres tienen forma globular, tamaño similar y densidades similares. [49]. El algoritmo es sensible a los *outliers* ya que utiliza la media (estadístico descriptivo no robusto sensible a casos extremos). Además, los algoritmos son sensibles a la selección de los centroides iniciales [49].

En cuanto a **Forgy/Lloyd**, es adecuado para analizar grandes conjuntos de datos, sin embargo es posible crear clústeres vacíos, si todas las observaciones son movidas a la vez de un clúster. El algoritmo de **Macqueen** converge más rápidamente que el de Forgy/Lloyd, sin embargo suele suponer un coste computacional mayor. Además tanto éste como el algoritmo de **Hartigan y Wong** son sensibles al orden en el que se evalúan las observaciones.

4.3.2.2. Selección del número de clústeres *k*

Una de las desventajas de *k-means* es tener que seleccionar el número de clústeres, sin embargo, existen diferentes criterios para realizar una selección objetiva del parámetro ‖. Para ello se optimizan las métricas que cuantifican la calidad de las soluciones.

A continuación se presentan tres métricas "internas", es decir, no se utiliza más información que la disponible a la hora de formar los clústeres. Una estrategia común para la selección del parámetro k es la de ir variar este parámetro y calcular las métricas en cada caso para poder seleccionar después la que objetivamente mejor haya resultado.

A continuación se crea un modelo *k-means* para ver si las técnicas son capaces de identificar grupos con compotarmiento similar en lo relativo al consumo de bacalao dependiendo de las tallas tanto de los predadores como de las presas. Para ello se emplea el paquete mlr3 (v0.8.0; Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B, 2019) disponible en R.

Para el análisis, se utilizan las variables que describen la talla y el peso del predador, y la talla y el peso de la presa.

En la Figura 4.27, los gráficos de densidad de la parte superior deberían ayudarnos a identificar las zonas con mayor densidad, lo que sería susceptible de ser un clúster. Eso mismo, en los gráficos de dispersión debería traducirse como zonas más negras.

Entre los superiores, a priori parece que se pueda hablar de al menos dos clústeres diferenciados (TallaPresa vs. TallaPredador) aunque no de manera demasiado clara. Las que más interesan son la

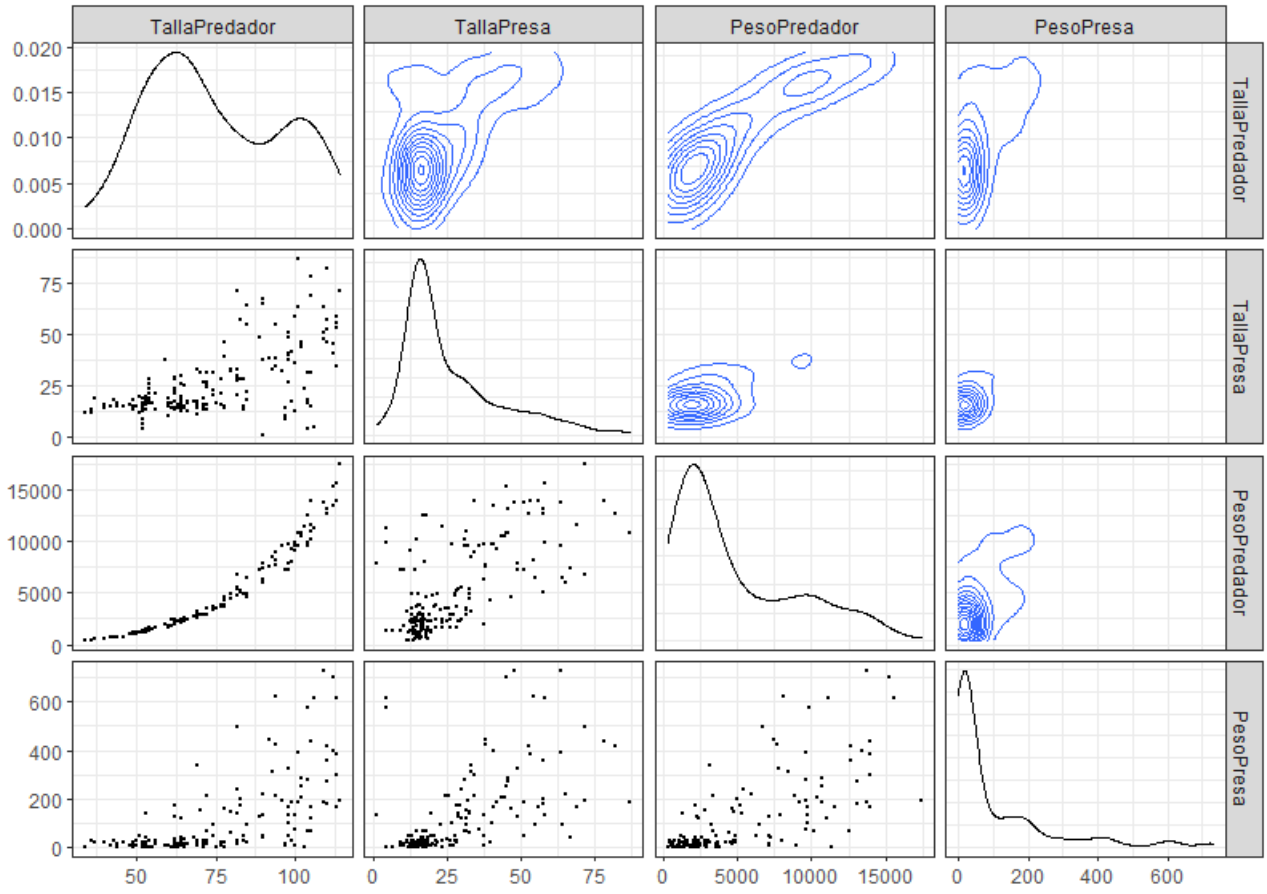


Figura 4.27: Encima de la diagonal: estimación de la densidad conjunta. Debajo de al diagonal: diagramas de dispersión. Diagonal: estimación de la densidad.

talla del predador y la de la presa, y que las otras variables han sido añadidas al modelo con intención de aportar más información que pudiera ser válida en el análisis, y ante la evidencia de que tal información puede resultar confusa en lugar de esclarecedora, se va a eliminar la variable PesoPresa del modelo, manteniendo sin embargo la del peso del predador. Esto se realiza así porque tal y como realiza el análisis de los contenidos estomacales, existe una mayor seguridad empírica sobre la adecuación de los datos del predador que de los de la presa.

Para identificar el número de grupos y las observaciones pertenecientes a cada uno de ellos, se aplican los tres algoritmos descritos basados en *k-means*: (Lloyd-Forgy, MacQueen y Hartigan-Wong) utilizando diferentes números de clústeres $k \in [1, 9]$. Y a eso se la aplican las métricas descritas (Davies-Bouldin y Pseudo-F) utilizando validación cruzada. Es decir, se hacen diferentes particiones de la muestra, se les aplican los algoritmos de k-means utilizando diferentes números de clústeres y después se aplica la métrica y se saca su media. Para identificar los mejores resultados, hay que tener en cuenta que la métrica Davies-Boulding se minimiza y Pseudo-F (o G1 en los gráficos) se maximiza.

En la Figura 4.28 aparecen los resultados obtenidos para cada algoritmo y utilizando diferente número de clústeres según los índices de Davies-Bouldin (db) y Pseudo-F (G1). La primera limitación apreciable es que según las métricas no es posible valorar el caso de un único clúster. En cuanto al criterio de Davies-Bouldin, teniendo en cuenta que hay que buscar el mínimo, los tres algoritmos parecen coincidir en que el caso en el que se crean 8 clústeres. Para el algoritmo de MacQueen, la calidad valorada por Davis-Boulding en el caso de la creación de 8 clústeres no dista en más de una décima

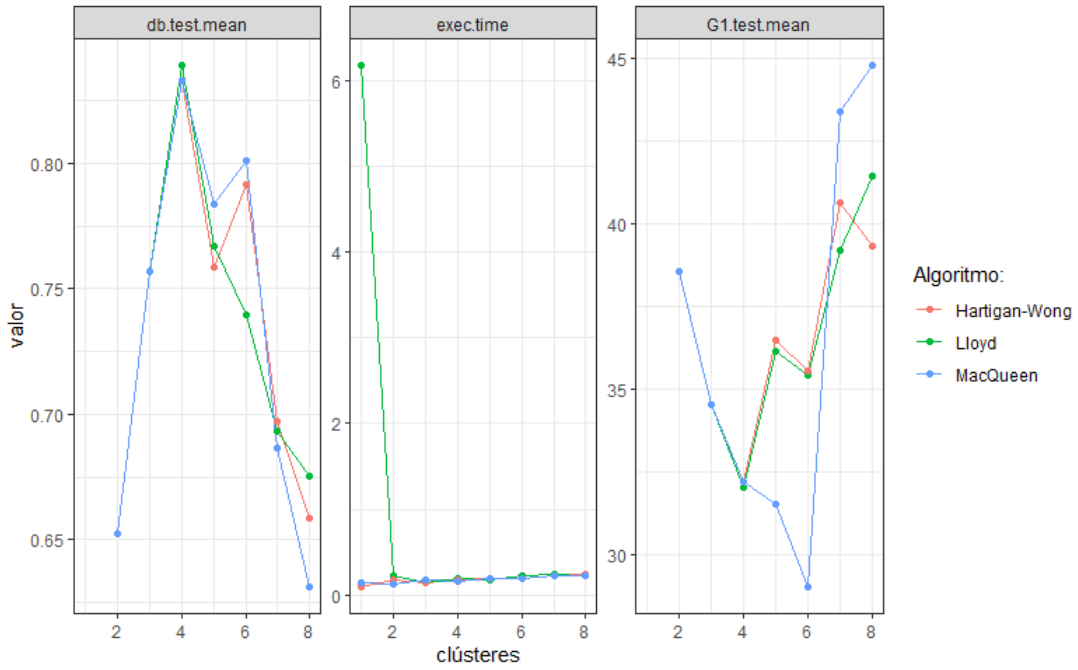


Figura 4.28: Métricas obtenidas para cada algoritmo formando diferentes números de clústeres.

del caso de creación de dos clústeres. Según la métrica Pseudo-F (G1), la cual hay que maximizar, Hatigan-Wong propone 7 clústeres, mientras que Lloyd/Forgy y MacQueen proponen 8. En el caso de MacQueen, el caso de dos clústeres proporciona un resultado lo suficientemente alto como para tenerlo en cuenta. En cuanto al tiempo de ejecución, no hay diferencias mencionables salvo en el caso del algoritmo Lloyd/Forgy para el caso de 1 clúster.

Así, fijando el número de clústeres en 8 y aplicando el algoritmo de MacQueen, se obtiene una agrupación en clústeres visible en la Figura 4.29, donde se pueden observar densidades, *boxplots*, regresiones, número de individuos, todo ello clasificado por clúster; además del conteo de casos por intervalos del eje x desagregado por clústeres.

Las agrupaciones generadas tienen indicios que podrían indicar un número no óptimo de clústeres. En el gráfico de densidad de TallaPredador vs. TallaPresa, los clústeres naranja, azul y verde aparecen divididos, lo que puede indicar que el número de clústeres debería ser mayor. Sin embargo, viendo la regresión de estas dos variables y el número de datos por clústeres en algunos casos, no parece tener demasiado sentido. Ya que puede ser que se trata de buscar patrones entre las observaciones disponibles, no crear patrones para describir las singularidades de cada dato, que es lo que se haría en el caso de que hubiera el mismo número de clústeres que observaciones hay en el conjunto de datos. Teniendo en cuenta que el objetivo es encontrar relaciones definibles por intervalos de tallas, este resultado de agrupamiento no ofrece demasiado información en este sentido ya que los diferentes clústeres generan grupos en los mismos intervalos. No tener un mayor volumen de datos pone en duda las propuestas de los Boxplot de estas dos variables a la hora de marcar algunos datos como atípicos.

En la Figura 4.30 se pueden observar los resultados pero seleccionando 4 clústeres sin hacer caso a las métricas. Se ha probado también con 3 clústeres, los resultados son similares pero unificando en uno solo los clústeres azul y naranja. En este caso se ve una diferenciación a priori bastante clara. Por un lado, se diferencian los ejemplares de pequeño tamaño que consumen ejemplares de pequeño tamaño (azul), ejemplares medianos que consumen ejemplares más pequeños y medianos (naranja) y por otro lado los ejemplares de gran tamaño (clústeres verde y morado). La diferenciación entre

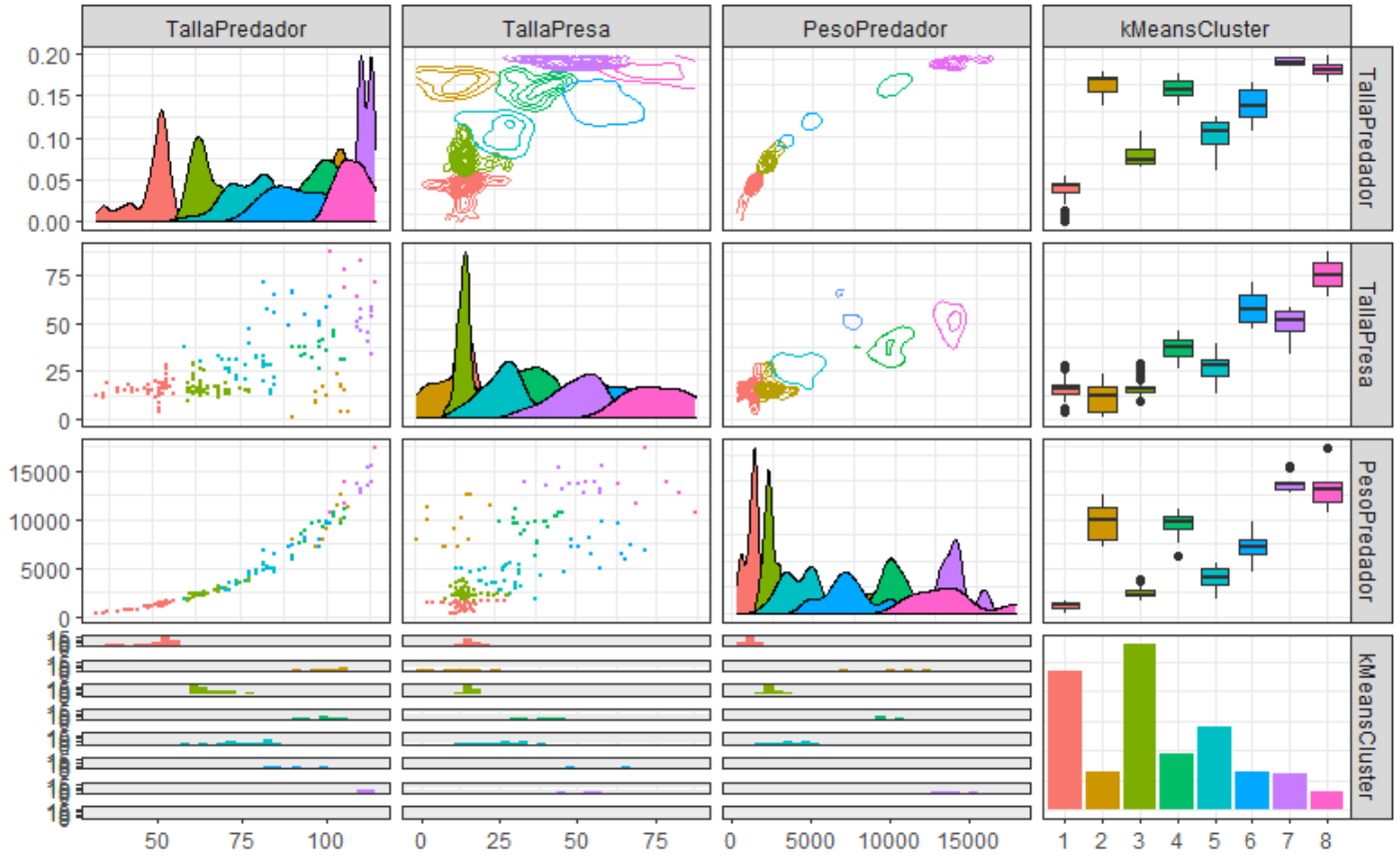


Figura 4.29: Resultados obtenidos aplicando MacQueen y creando 8 clústeres.

estos dos clústeres no parece tan inmediata, de modo que podría ser difícil, sabiendo el tamaño de un consumidor, qué tamaño de presa tendría como objetivo. El peso del predador, que se ha metido por si pudiera servir es estos casos, no supone un aporte de información útil. Se puede intuir que los individuos con mayor peso pertenecen al clúster morado, pero aun así no resulta en una diferenciación clara.

Se puede obtener la siguiente conclusión de esta separación de 4 grupos:

- Coincide con el cambio ontogénico por el cual el cambio por crecimiento supone un cambio en la dinámica trófica.
- Este cambio en relación a la alimentación a base de canibalismo denota cuatro fases en su pauta de predador,que de forma aproximada se puede agrupar en individuos menores de 60 cm, 60-80 cm, 80-90 cm y mayores de 90 cm.
- El análisis también esboza las fases bajo el papel trófico de presa por parte del bacalao en el ecosistema de Flemish Cap, de forma aproximada podemos agrupar los individuos en menores de 10 cm, 10-25 cm, 30-50 cm y mayores de 50 cm.

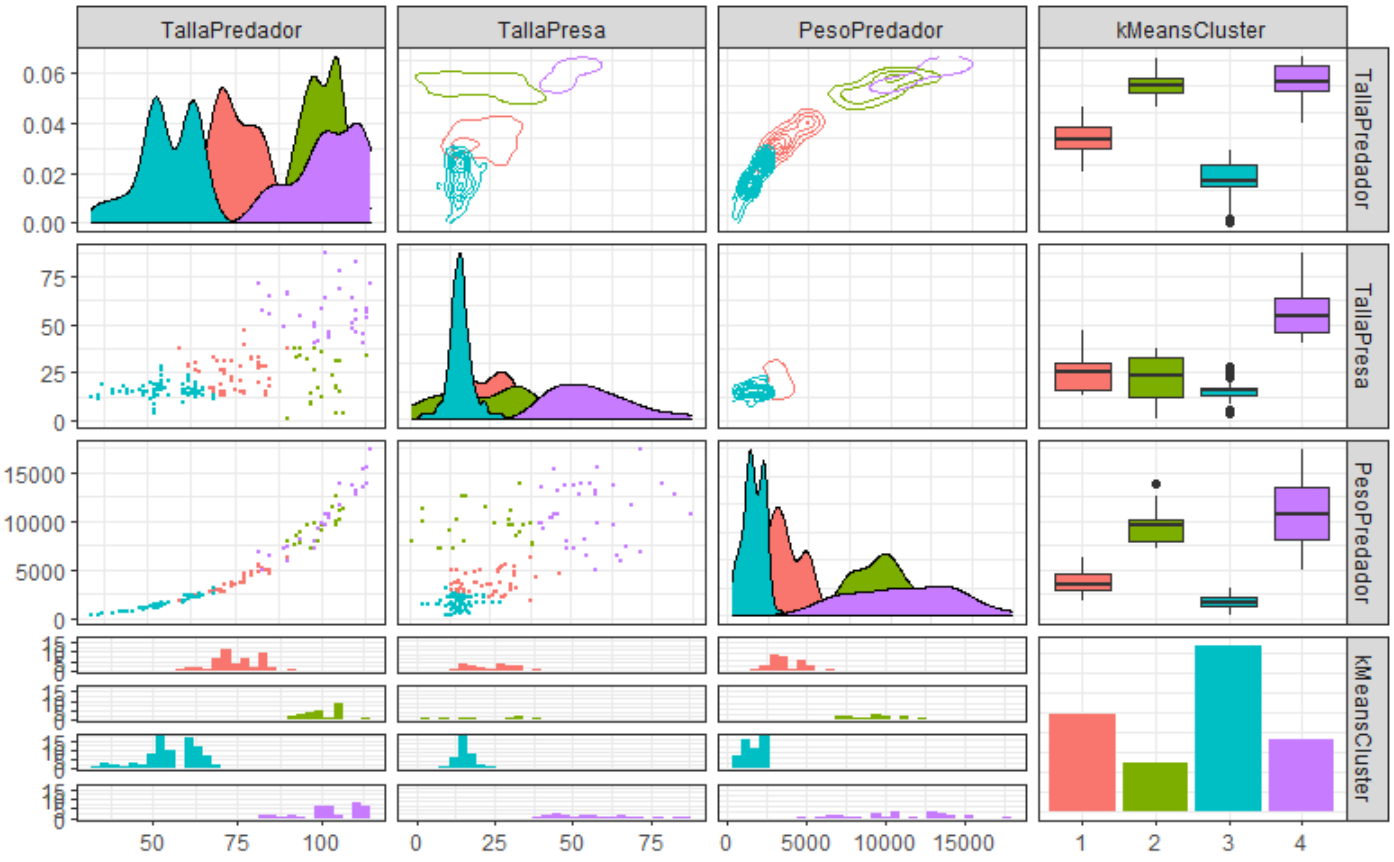


Figura 4.30: Resultados obtenidos aplicando MacQueen y creando 4 clústeres.

El *Feeding Intensity Index*(FI) es un índice utilizado para describir la intensisdad de alimentación (presencia/ausencia de alimento en el estomago, no la cantidad), en este caso del bacalao a lo largo del período 1993-2018, faltando los años 2007 y 2009.

Se define como:

$$FI = \frac{n}{N} \cdot 100 \tag{4.2}$$

n = individuos con contenido estomacal

N = número total de individuos

en el caso de $\text{FI}_{\text{bacalao}}$, se cuentan los individuos cuyo contenido estomacal también es bacalao.

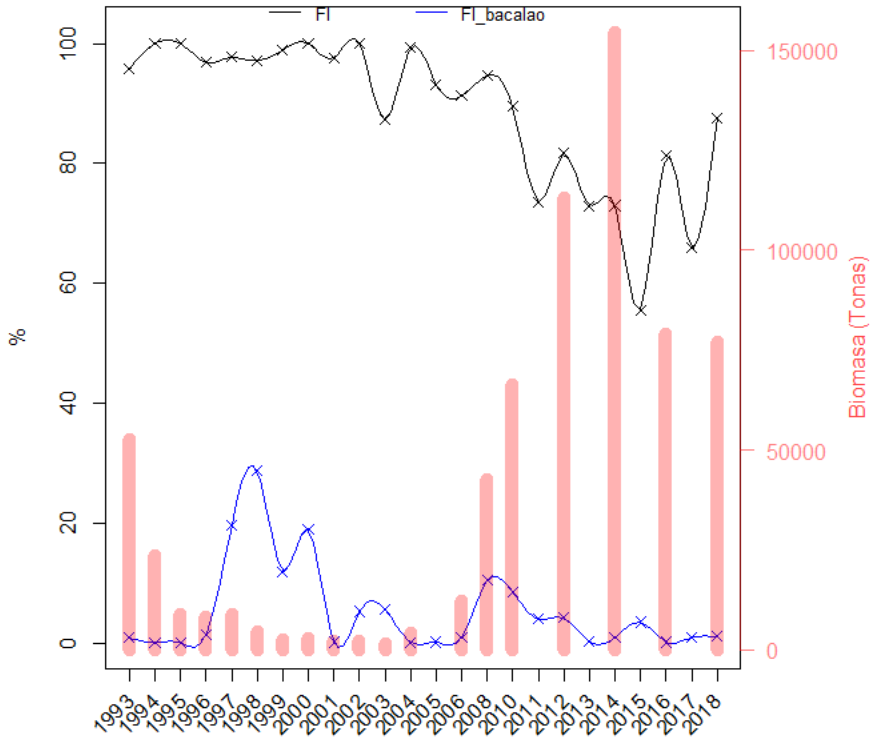


Figura 4.31: FI y $\text{FI}_{\text{bacalao}}$ para 1993-2018

La Figura 4.31 muestra los valores más elevados de $\text{FI}_{\text{bacalao}}$ en los años donde la biomasa del bacalao fue declinado a mínimos históricos, de finales de los años 90 a 2004; y sucede lo contrario a partir del 2011. Cabría pensar que la elevada predación sobre el bacalao que ejerce la propia especie ($\text{FI}_{\text{bacalao}}$ alto) llevó a la disminución de la biomasa.

La Figura 4.32 parece indicar que sucede lo dicho anteriormente, probablemente sería mejor hacerlo por grupo de tallas o edad, pero en principio no parece muy cierta la hipótesis anterior. Al principio de la serie (hasta 1996 aproximadamente) se ve que la mayor parte de la biomasa corresponde a individuos

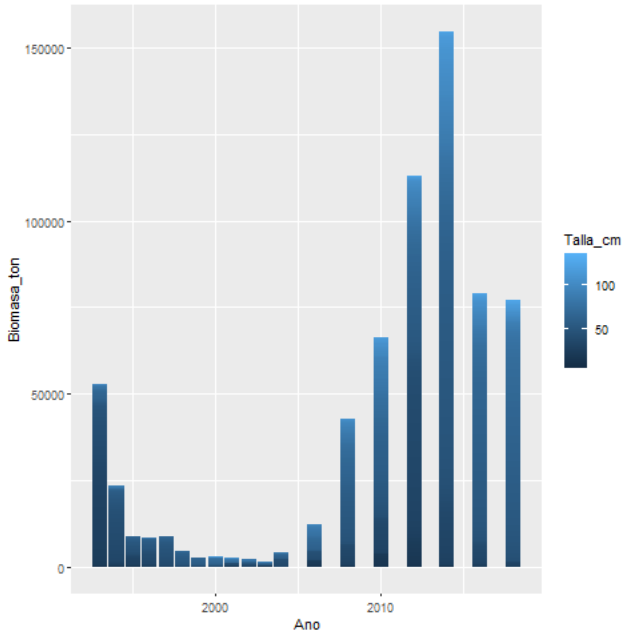


Figura 4.32: Biomasa por años y tallas

de tallas pequeñas, lo cual explica los valores bajos de $FI_{bacalao}$ en esos años, porque no pueden predarse cuando son del mismo tamaño. Es decir, la población estuvo constituída por la fracción de la población que representa la función de presa, y apenas hay fracción que ejerce de predadora. En los años siguiente sucede lo contrario, aumenta la fracción predadora, que realmente es la anterior población que ha crecido y ya emprenden el comportamiento predador de canibalismo porque su tamaño se lo permite. En la parte final de la serie temporal estudida, aun con un periodo sorprendente de recuperación del stock, observamos cierto equilibrio en la distribución poblacional en cuanto a composición de tallas, de ahí la mortalidad por predación digamos equilibrada.

Similares resultados son obtenidos haciendo el análisis a través del índice MWFI, como ofrece la Figura 4.33. Se aprecia que esta pauta de alimentación seguida por esta especie es reflejada de forma similar por $MWFI_{bacalao}$ y $FI_{bacalao}$, resultando un bioindicador de la situación de la población.

Las conclusiones que se pueden extraer de este análisis son las siguientes:

- La alimentación mediante canibalismo es habitual en el bacalao.
- Para que se produzca debe haber en la población tanto una fracción que resulte ser la presa, y otra fracción que ejerca de predador, y ello está definido claramente a lo largo del periodo vital comprendiendo rangos de tallas en cada caso.
- La observación y cuantificación de este hábito alimenticio ofrece una importante información indicadora de la situación demográfica de la población.

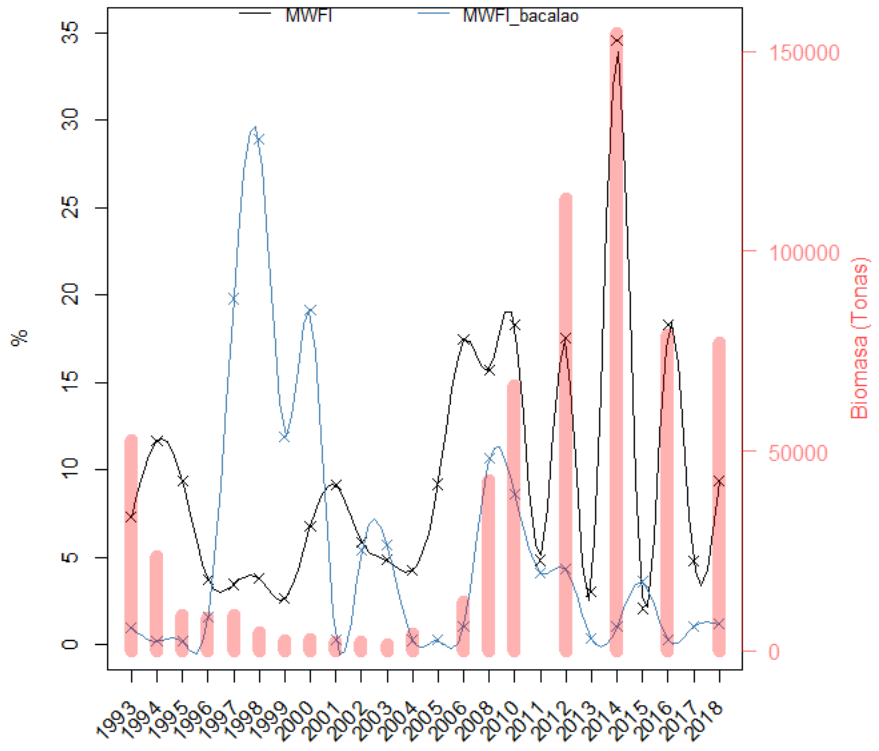


Figura 4.33: Mean Weight Fulness Index (MWFI) y MWFI_{cod} para 1993-2018.

Capítulo 5

Conclusiones

Los datos utilizados con información de los contenidos estomacales recogidos por el Instituto de Oceanografía Español han servido para analizar con mayor detalle el canibalismo dentro de la especie *G. morhua*. Para ello en primer lugar se ha hecho una revisión de diferentes técnicas estadísticas que se han utilizado para aunar en el tema: regresión, diagnósticos de normalidad e independencia, *clustering*, detección de atípicos, etc. Se ha cumplido el objetivo de generar conocimiento acerca de las relaciones tróficas del bacalao *G. morhua* con individuos de su misma especie, siendo estos los puntos más importantes:

- Los datos disponibles, resultantes del muestreo que se realiza, aún intentando la mayor y similar representatividad en función de todas las variables (temporal, talla, sexo, batimétrica, etc.), siempre presentan sesgos que van a influir. Por ello, el conocimiento de la información manejada contribuye al mejor desarrollo e interpretación.
- Las principales presas de alimentación en la alimentación del bacalao en Flemish Cap son: hypéridos, camarón boreal y gallineta. De la misma forma que son presas esenciales para otras especies allí distribuidas, por lo tanto son la base trófica fundamental del ecosistema de Flemish Cap.
- La dieta del bacalao varía con la talla del predador, la disponibilidad de presas y la profundidad de distribución, y presenta variaciones temporales y estacionales.
- El análisis de clústeres se puede perfilar:
  - El cambio ontogénico trófico: cambio de la dieta a medida que crece el individuo predador.
  - La alimentación predando sobre la propia especie *G. morhua* denota cuatro fases derivadas de la cantidad consumida y el tamaño de la presa.
  - El papel de presa por parte del bacalao está delimitado por el tamaño: una vez superada la talla crítica, la tasa de mortalidad por canibalismo prácticamente desaparece.
- Para que se produzca el canibalismo debe haber en la población tanto la fracción de tallas que actúa como presa como la que ejerce de predador.
- Esta pauta alimenticia sirve como bioindicador del reclutamiento y el análisis del tamaño de las presas ofrece una visión clara de las clases anuales más abundantes.

El estudio realizado mediante este trabajo debería ser continuado con la finalidad de estimar el consumo anual que ejerce el canibalismo sobre la presa, completando información sobre la componente de mortalidad natural para los modelos de evaluación. Para alcanzar tal estima se debe profundizar en el conocimiento de la actividad alimenticia a lo largo del año ya que la conocida disminución de la intensidad de alimentación durante ciertos períodos (en invierno, o previa a la puesta) deben ser cuantificados para obtener una estima anual más correcta.

Índice de figuras

Índice de cuadros

Bibliografía

[1] WILLIAM B. HAMILTON. *Place Names of Atlantic Canada*. University of Toronto Press, 1996. ISBN: 9780802075703. URL: http://www.jstor.org/stable/10.3138/9781442678507.

[2] Mark Kurlansky. *Cod. A biography of the fish that changed the world*. Charlotte Sheedy, 1997. ISBN: 9788412340174.

[3] Enrique de CÃ¡rdenas GonzÃ¡lez. «DinÃ¡mica de la poblaciÃ³n*delbacalaodeFlemishCap.Consideracionessobresuai* Tesis doct. 1995.

[4] Jenny Higgins. *Cod Moratorium in Newfoundland and Labrador*. 2009. URL: https://www.heritage.nf.ca/articles/economy/moratorium.php (visitado 16-05-2020).

[5] Doubleday WG. *Manual on Groundfish Surveys in the Northweast Atlantic*. English. North Atlantic Fisheries Organization. 203 págs.

[6] Instituto Oceanografico Espanol (IOE). *Protocolos de la Campana de Investigacion Pesquera en Flemish Cap*. Castellano. IOE. 51 págs.

[7] E. A. Nadaraya. «On Estimating Regression». En: *Theory of Probability & Its Applications* 9.1 (1964), págs. 141-142. DOI: 10.1137/1109020.

[8] Geoffrey S. Watson. «Smooth Regression Analysis». En: *SankhyÄ: The Indian Journal of Statistics, Series A (1961-2002)* 26.4 (1964), págs. 359-372. ISSN: 0581572X. (Visitado 25-08-2022).

[9] S. S. SHAPIRO y M. B. WILK. «An analysis of variance test for normality (complete samples)». En: *Biometrika* 52.3-4 (dic. de 1965), págs. 591-611. DOI: 10.1093/biomet/52.3-4.591.

[10] Carlos M. Jarque y Anil K. Bera. «A Test for Normality of Observations and Regression Residuals». En: *International Statistical Review / Revue Internationale de Statistique* 55.2 (1987), págs. 163-172. ISSN: 03067734, 17515823. URL: http://www.jstor.org/stable/1403192.

[11] G. E. P. Box y David A. Pierce. «Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models». En: *Journal of the American Statistical Association* 65.332 (1970), págs. 1509-1526. ISSN: 01621459. URL: http://www.jstor.org/stable/2284333.

[12] Stuart P. Lloyd. «Least squares quantization in PCM». En: *IEEE Trans. Inf. Theory* 28 (1982), págs. 129-136.

[13] E. W. Forgy. «Cluster analysis of multivariate data : efficiency versus interpretability of classifications». En: *Biometrics* 21 (1965), págs. 768-769.

[14] J. B. MacQueen. «Some Methods for Classification and Analysis of MultiVariate Observations». En: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. por L. M. Le Cam y J. Neyman. Vol. 1. University of California Press, 1967, págs. 281-297.

[15] J. A. Hartigan y M. A. Wong. «Algorithm AS 136: A K-Means Clustering Algorithm». En: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), págs. 100-108. ISSN: 00359254, 14679876. (Visitado 25-08-2022).

[16] David L. Davies y Donald W. Bouldin. «A Cluster Separation Measure». En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (1979), págs. 224-227. DOI: 10.1109/TPAMI.1979.4766909.

[17] D. M. Hawkins. *Identification of outliers*. Monographs on applied probability and statistics. London [u.a.]: Chapman y Hall, 1980. X, 188. ISBN: 041221900X.

[18] Keith Ord. «Outliers in statistical data : V. Barnett and T. Lewis, 1994, 3rd edition, (John Wiley & Sons, Chichester), 584 pp., [UK pound]55.00, ISBN 0-471-93094-6». En: *International Journal of Forecasting* 12.1 (mar. de 1996), págs. 175-176.

[19] D.S. Moore y G.P. McCabe. *Introduction to the Practice of Statistics*. Introduction to the Practice of Statistics. W.H. Freeman, 1999. ISBN: 9780716735021.

[20] Peter Rousseeuw y Mia Hubert. «Robust statistics for outlier detection». En: *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery* 1 (ene. de 2011), págs. 73-79. DOI: 10.1002/widm.2.

[21] R. Dennis Cook. «Detection of Influential Observation in Linear Regression». En: *Technometrics* 19.1 (1977), págs. 15-18. ISSN: 00401706. (Visitado 30-08-2022).

[22] Kenneth A Bollen y Robert W Jackman. «Regression diagnostics: An expository treatment of outliers and influential cases». En: *Sociological Methods & Research* 13.4 (1985), págs. 510-542.

[23] R.D. Cook y S. Weisberg. *Residuals and Influence in Regression*. Monographs on statistics and applied probability. Chapman y Hall, 1986.

[24] Myung Kim. «A cautionary note on the use of Cook's distance». En: *Communications for Statistical Applications and Methods* 24 (mayo de 2017), págs. 317-324. DOI: 10.5351/CSAM.2017.24.3.317.

[25] Kesar Singh y Minge Xie. «Bootlier-Plot: Bootstrap Based Outlier Detection Plot». En: *SankhyÄ: The Indian Journal of Statistics (2003-2007)* 65 (ene. de 2003), págs. 532-559. DOI: 10.2307/25053287.

[26] Bertrand Candelon y Norbert Metiu. *A distribution-free test for outliers*. eng. Bundesbank Discussion Paper 02/2013. Frankfurt a. M.: Deutsche Bundesbank, 2013.

[27] Joseph H. Silverman. «Lower bound for the canonical height on elliptic curves». En: *Duke Math. J.* 48.3 (sep. de 1981), págs. 633-648. DOI: 10.1215/S0012-7094-81-04834-1.

[28] B. Efron. «Bootstrap Methods: Another Look at the Jackknife». En: *Ann. Statist.* 7.1 (ene. de 1979), págs. 1-26. DOI: 10.1214/aos/1176344552.

[29] Leland Wilkinson. «Visualizing Outliers». En: 2016.

[30] MarÃa-JosÃ© RodrÃguez-Jaume y Rafael Mora CatalÃ¡ *AnÃ¡lisis de correspondencia*. 2001.

[31] Komala Sheshachala Srikanth. *solitude: An Implementation of Isolation Forest*.

[32] H. B. Mann y D. R. Whitney. «On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other». En: *Ann. Math. Statist.* 18.1 (mar. de 1947), págs. 50-60. DOI: 10.1214/aoms/1177730491. URL: https://doi.org/10.1214/aoms/1177730491.

[33] Anna Hart. «Mann-Whitney test is not just a test of medians: differences in spread can be important». En: *BMJ* 323.7309 (2001), págs. 391-393. ISSN: 0959-8138. DOI: 10.1136/bmj.323.7309.391. eprint: https://www.bmj.com/content/323/7309/391.full.pdf. URL: https://www.bmj.com/content/323/7309/391.

[34] G. E. P. Box y D. R. Cox. «An Analysis of Transformations». En: *Journal of the Royal Statistical Society. Series B (Methodological)* 26.2 (1964), págs. 211-252. ISSN: 00359246. URL: http://www.jstor.org/stable/2984418.

[35] W. N. Venables y B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[36] E. J. Hyslop. «Stomach contents analysisâa review of methods and their application». En: *Journal of Fish Biology* 17.4 (1980), págs. 411-429. ISSN: 10958649. DOI: 10.1111/j.1095-8649.1980.tb02775.x.

[37] B. L. Townhill y col. «Diets of the Barents Sea cod (*Gadus morhua*) from the 1930s to 2018». En: *Earth System Science Data* 13.3 (2021), págs. 1361-1370. DOI: 10.5194/essd-13-1361-2021.

[38] Sigbj Mehl y Knut Sunnanå. «Changes in growth of Northeast Arctic cod in relation to food consumption in 1984-1988». En: 1991.

[39] Webb D Keast A. «Mouth and Body Form Relative to Feeding Ecology in the Fish Fauna of a Small Lake, Lake Opinicon, Ontario». En: *Journal of the Fisheries Research Board of Canada* 23 (1966), págs. 1845-1874. DOI: 10.1139/f66-175.

[40] Webb D. «The effect of size on the fast-start performance of rainbow trout, Salmo Gardneri, and a consideration of piscivorous predator-prey interactions». En: *J. exp. Biol.* 65 (1976), págs. 157-177.

[41] Beamish FWH. «Swimming capacity». En: *In: Hoar WS, Randall Dj (eds) Fish physiology. Academic Press, New York* (1978), págs. 62-67.

[42] OA Popova. *The âpredator-preyârelationship among fish*. 1967.

[43] Larry A Nielsen y William F Schoch. «Errors in estimating mean weight and other statistics from mean length». En: *Transactions of the American Fisheries Society* 109.3 (1980), págs. 319-322.

[44] L. Persson. «A field experiment on the effects of interspecific competition from roach, Rutilus rutilus (L.), on age at maturity and gonad size in perch, Percafluviatilis L.» En: *Journal of Fish Biology* 37.6 (1990), págs. 899-906. DOI: https://doi.org/10.1111/j.1095-8649.1990.tb03593.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1095-8649.1990.tb03593.x.

[45] Francis Juanes, JA Buckel y FS Scharf. «Predatory behaviour and selectivity of a primary piscivore: Comparison of fish and non-fish prey». En: *Marine Ecology-progress Series - MAR ECOL-PROGR SER* 217 (jul. de 2001), págs. 157-165. DOI: 10.3354/meps217157.

[46] Rountree R Scharf F Juanes F. «Predator size-prey size relationships of marine fish predators: interspecific variation and effects of ontogeny and body size on trophic-niche breadth». En: *Marine Ecology-progress Series - MAR ECOL-PROGR SER* 208 (dic. de 2000), págs. 229-248. DOI: 10.3354/meps208229.

[47] William H. Kruskal y W. Allen Wallis. «Use of Ranks in One-Criterion Variance Analysis». En: *Journal of the American Statistical Association* 47.260 (1952), págs. 583-621. DOI: 10.1080/01621459.1952.10483441.

[48] H.I. Rhys. *Machine Learning with R, the tidyverse, and mlr*. Manning Publications, 2020. ISBN: 9781617296574. URL: https://books.google.es/books?id=BoeryQEACAAJ.

[49] S. Äyrämö y T. Kärkkäinen. «Introduction to partitioning-based clustering methods with a robust example». En: 2006.