

Trabajo Fin de Máster

**Análisis de la capacidad dispersiva y
filopatria de las poblaciones del género *Alosa*
en el noroeste de la península ibérica a través
de técnicas de aprendizaje automático.**

Alejandro Pico Calvo

Máster en técnicas estadísticas

Curso 2021-2022

Título en galego: Análise da capacidade dispersiva e filopatria das poboacións do xénero <i>Alosa</i> no noroeste da península ibérica a través de técnicas de aprendizaxe automático.
Título en español: Análisis de la capacidad dispersiva y filopatria de las poblaciones del género <i>Alosa</i> en el noroeste de la península ibérica a través de técnicas de aprendizaje automático.
English title: Analysis of the dispersal capacity and philopatry of populations of the genus <i>Alosa</i> in the northwest of the Iberian Peninsula using machine learning techniques.
Modalidad: Modalidad A
Autor/a: Alejandro Pico Calvo, Universidad de Santiago de Compostela
Director/a: Rosa María Crujeiras Casais, Universidad de Santiago de Compostela
Codirector/a: Fernando Cobo Gradín, Universidad de Santiago de Compostela
Breve resumen del trabajo: Análisis de la capacidad dispersiva y filopatria de las poblaciones de alosas del noroeste de la península ibérica a través del estudio de la composición microquímica de los otolitos de 141 adultos, 137 juveniles y la composición química de 17 ríos. Para ello, se asignará un río natal a los ejemplares a través de modelos de clasificación <i>machine learning</i>.

Doña Rosa María Crujeiras Casais y don Fernando Cobo Gradín de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Análisis de la capacidad dispersiva y filopatria de las poblaciones del género *Alosa* en el noroeste de la península ibérica a través de técnicas de aprendizaje automático.

Fue realizado bajo su dirección por don Alejandro Pico Calvo para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 29 de agosto de 2022.

La directora:

El codirector:

Doña Rosa María Crujeiras Casais

Don Fernando Cobo Gradín

El autor:

Don Alejandro Pico Calvo

Agradecimientos

En primer lugar, a todo el equipo de investigación del Laboratorio de Hidrobiología. A Fernando por darme la oportunidad de iniciar mi carrera investigadora, confiar en mí e incluirme en los proyectos del equipo de investigación desde que era estudiante de grado. A Rufino, Sandra y Chimpi por hacerme sentir como parte de una familia y ayudarme siempre que lo he necesitado. Querría agradecer especialmente a Nachón el haber sido como un hermano mayor y guía para mí, estando siempre que lo he necesitado, del que he aprendido enormemente y sin el cual no habría sido posible desarrollar este trabajo.

A mi familia y mi pareja Andrea por ser un gran apoyo a lo largo de esta aventura que ha sido el Máster de Técnicas Estadísticas, ayudarme siempre que lo he necesitado y darme fuerzas para seguir siempre adelante en los momentos más duros.

A mis amigos y especialmente a Fikru y Óscar, sin los cuales estos años de máster habrían sido muy duros, fueron un apoyo diario necesario. Me llevo una amistad que atesoraré toda la vida.

Finalmente, mi eterno agradecimiento a Rosa, directora de este trabajo, por ser una increíble docente impartiendo las asignaturas del máster, una excelente tutora a la hora de guiarme en el transcurso de este trabajo y siempre estar dispuesta a ayudar con independencia del mes, día y hora a la que le llegasen mis correos. Siempre positiva aportando nuevas ideas y perspectivas, sin ella no habría sido posible la consecución de este estudio.

Índice general

Resumen en español	XII
English abstract	XIII
Introducción	15
1.1. Contexto biológico	15
1.1.1. Especies objeto de estudio.....	15
1.1.2. Problemática biológica	16
1.2. Objetivos	17
1.3. Resumen de procedimientos.....	18
1.4. Área de estudio y origen de los datos	21
1.5. Análisis de los otolitos	22
Metodología estadística.....	23
2.1. Estimación de parámetros y simulación de juveniles.....	23
2.1.1. Justificación.....	23
2.1.2 Selección de metodología.....	24
2.1.3. Modelos de regresión lineal	25
2.2. Análisis exploratorio de la microquímica de los juveniles.....	28
2.2.1. Diagramas de dispersión de los parámetros microquímicos	28
2.2.2 Análisis de similitud entre juveniles de distinto origen.....	29
2.3. Selección de ríos y parámetros para los diferentes análisis.....	30
2.3.1. Selección de ríos.....	30
2.3.2. Selección de parámetros químicos	31
2.4. Métodos de clasificación	31
2.4.1. Métodos basados en árboles de decisión	32
2.4.2. <i>Boosting</i>	33
2.4.3. <i>Bagging – random forest</i>	36
2.4.4. QDA	38
2.5. Entrenamiento, evaluación y comparación de los modelos.....	39
2.5.1. Entrenamiento y partición de los datos	39
2.5.2. Evaluación y comparación de modelos	39
2.5.3. Ensamblado de modelos.....	40
2.6. Clasificación de los ejemplares adultos.....	41
Resultados	41
3.1. Evaluación y comparación de los modelos	41
3.1.1. Estudio de la filopatria	42

3.1.2. Estudio de la dispersión en el medio marino.....	52
3.1.3. Selección de modelos y ejemplares para la clasificación.....	62
3.2. Clasificación de los ejemplares adultos.....	62
3.2.1. Ejemplares capturados en el medio fluvial (estudio de la filopatria)	63
3.2.2. Ejemplares capturados en el medio marino (estudio de la dispersión).....	65
Discusión y conclusiones	67
4.1. Rendimiento y comparación de los modelos.....	67
4.2. La importancia del ratio isotópico.....	67
4.3. Origen de los ejemplares capturados en el medio fluvial.....	68
4.3.1. Los ríos Miño y Ulla	68
4.3.2. Los ríos Lérez y Tambre	69
4.3.3. El río Mondego	69
4.3.4. El comportamiento filopátrico	70
4.4. Análisis de la dispersión en el medio marino.....	71
4.5. Limitaciones del trabajo y propuestas	71
Anexo	73
Modelos: stochastic gradient boosting	73
Modelos seleccionados para el estudio de la filopatria	73
Modelos seleccionados para el estudio de la dispersión	76
Modelos: random forest	79
Modelos seleccionados para el estudio de la filopatria	79
Modelos seleccionados para el estudio de la dispersión	82
Referencias bibliográficas	85
Apéndices.....	88
Apéndice de figuras.....	88
Apéndice de tablas	91

Resumen en español

Alosa alosa (Linnaeus, 1758) y *Alosa fallax* (Lacépède, 1803) son dos especies de peces anádromos pertenecientes a la familia Clupeidae. Como el resto de las especies diádromas¹, sus poblaciones en los ríos europeos están en declive y su escasez se ha convertido en un problema emergente.

Los datos sobre muchas características biológicas del ciclo vital de estas especies y especialmente los correspondientes a la fase juvenil y a su desarrollo en el medio marino y costero son aún escasos. En la península ibérica apenas se dispone de datos científicos sobre sus poblaciones. En Galicia, existen estudios sobre sus poblaciones en los ríos Ulla y Miño, sin embargo, se desconoce si estas especies pueden residir en otros ríos gallegos.

La microquímica de los otolitos de los peces diádromos guarda una fuerte correlación con la microquímica del agua de su río de nacimiento. En el estudio que presentamos se analizó la composición microquímica de los otolitos de 141 adultos, repartidos entre ambas especies, capturados en el mar y en los ríos Ulla y Miño y se comparó con la de los otolitos de 137 juveniles, repartidos entre ambas especies, procedentes de diversos ríos del noroeste peninsular y Francia. De esta forma, a través de dicha correlación se pudo estudiar el carácter filopátrico de los ejemplares, especificar su río de nacimiento y establecer su rango de dispersión durante la fase marina. Así, mediante la aplicación de metodologías *machine learning*, aportamos nueva información del rango dispersivo de estas especies y de las interacciones entre poblaciones.

¹ Especies de peces que migran y alternan entre medios fluviales y marinos.

English abstract

Alosa alosa (Linnaeus, 1758) and *Alosa fallax* (Lacépède, 1803) are two species of anadromous fish belonging to the family Clupeidae. Like all other diadromous species, their populations in European rivers are declining and their scarcity has become an emerging problem.

Data on many biological characteristics of the life cycle of these species and especially those corresponding to the juvenile phase and their development in the marine and coastal environment are still scarce. In the Iberian Peninsula, there are hardly any scientific data available on their populations. In Galicia, there are studies on their populations in the Ulla and Miño rivers; however, it is not known whether these species may reside in other Galician rivers.

The microchemistry of the otoliths of diadromous fish is strongly correlated with the microchemistry of the water of their river of birth. In the present study, the microchemical composition of the otoliths of 141 adults, distributed between both species, caught at sea and in the Ulla and Miño rivers was analysed and compared with that of the otoliths of 137 juveniles, distributed between both species, from several rivers in the northwest of the Iberian Peninsula and France. In this way, through this correlation it was possible to study the philopatric character of the specimens, specify their river of birth and establish their dispersal range during the marine phase. Thus, through the application of machine learning methodologies, we provide new information on the dispersal range of these species and the interactions between populations.

Capítulo 1

Introducción

Este primer capítulo se divide en cinco apartados, centrándose el primero de estos, en presentar las especies analizadas en el trabajo y exponer las causas que han motivado la elección de este tema para la realización del trabajo de fin de máster. Seguidamente, se enuncian los objetivos que se pretenden alcanzar al culminar el estudio y se lleva a cabo una breve descripción de las principales metodologías utilizadas para la consecución de dichos objetivos.

Por otro lado, en el cuarto apartado, se especifica la procedencia de las muestras analizadas y la región geográfica en la que se encuadra el estudio.

Finalmente, y para concluir el capítulo, se lleva a cabo una síntesis centrada en la metodología utilizada para el tratamiento del material biológico (los otolitos) cuyo fin persigue extraer la información necesaria para la culminación del proyecto.

1.1. Contexto biológico

A continuación, se realiza una breve presentación de las especies objetivo del estudio, así como de la situación ecológica en la que se encuentran sus poblaciones haciendo hincapié en la región geográfica del noroeste peninsular en la que se encuadra el trabajo.

1.1.1. Especies objeto de estudio

Bajo el término genérico de alosas, se han descrito 16 especies pertenecientes a la familia de los Clupeidos, en la misma en la que se incluyen la sardina [*Sardina pilchardus* (Walbaum, 1792)] o el arenque *Clupea harengus* Linneus, 1758, entre otros. En Europa están presentes en simpatria dos especies de alosas: el sábalo [*Alosa alosa* (Linnaeus, 1758)] y la saboga [*Alosa fallax* (Lacépède, 1803)] (figura 1).

Ambas especies son migradoras anádromas, presentan ciclos de vida complejos en los que la reproducción y los primeros estadios de vida se desarrollan en hábitats de agua dulce, mientras que el crecimiento y la maduración se completan en hábitats oceánicos. Así, los juveniles migran al mar durante su primer año de vida (Taverny 1991; Taverny y Elie 2001; Nachón 2017). Durante su fase de crecimiento, las alosas se congregan en los hábitats costeros y realizan movimientos de ida y vuelta entre el mar y el estuario, especialmente la saboga durante sus dos primeros años de vida (Taverny 1991; Taverny y Elie 2001; Nachón 2017). Los hábitats utilizados por los subadultos de ambas especies se encuentran cerca de las desembocaduras de los ríos con las zonas de desove más importantes (Taverny 1991; Taverny y Elie 2001). La fase de crecimiento se completa en el mar hasta que alcanzan la madurez sexual (generalmente entre 3 y 6 años de edad)

y migran de vuelta a los ríos para desovar (Aprahamian *et al.* 2003a, 2003b; Baglinière *et al.* 2003). Como muchos peces anádromos, las alosas tienden a regresar a su río natal para desovar (presentan *homing*), con diversos grados de fidelidad entre las distintas poblaciones, aunque pocos individuos se desvían para desovar en un río diferente al natal (Martin *et al.* 2015; Randon *et al.* 2018).



Figura 1: Ejemplares adultos de *A. alosa* arriba y *A. fallax* abajo. Fuente: Modificado de Nachón (2017).

1.1.2. Problemática biológica

En el medio continental gallego existen algunas de las pocas poblaciones de alosas de la península ibérica (ríos Ulla y Miño) (Nachón 2017). Las poblaciones han sufrido un drástico declive a lo largo del siglo XX. En la actualidad, a pesar de que parecen ser estables y recurrentes en el tiempo, siguen en grave riesgo debido a que las presiones antrópicas, aquellas originadas a causa de la actividad humana, no han disminuido. Así pues, estas especies presentan una serie de valores culturales y etnográficos en esta comunidad autónoma. Además, estas especies constituyen un componente original de la biodiversidad cuyas particularidades biológicas y ecológicas (diversidad de los medios que frecuenta, la teórica correspondencia stock-río, la ausencia de comportamiento de salto y de poblaciones de repoblación), las convierten en indicadores privilegiados de la calidad biológica y de la conectividad de los ecosistemas acuáticos (Baglinière *et al.* 2003). Por ello, a pesar de estos intereses económicos, científicos y patrimoniales que suscitan y del acusado declive de sus poblaciones en las últimas décadas, sorprende la escasa atención que han recibido en Galicia en materia de gestión e investigación. Tan sólo muy recientemente se han desarrollado dos tesis doctorales, una en el río Ulla y otra en el río Miño, en los inicios de la década de 2010 (Mota, 2014; Nachón, 2017), que han intentado paliar y revertir parcialmente esta situación, aportando datos muy interesantes a nivel científico y a nivel de gestión y protección. Sin embargo, esos trabajos se centran casi exclusivamente en el estudio de la fase continental de ambas especies, por lo que persiste una gran laguna de

conocimiento sobre la fase marina, de importancia vital para el futuro de estas especies. Se desconoce en gran medida su distribución en el medio marino, qué hábitats frecuentan, su ecología trófica y aspectos tan relevantes como la dispersión en el medio marino y si existe conexión o mezcla de individuos procedentes de poblaciones continentales adyacentes (Nachón *et al.* 2016; Nachón 2017). El conocimiento de los aspectos biológicos y ecológicos durante la fase marina es indispensable para establecer políticas y planes de gestión eficaces para conservarlas y, por extensión, para conservar la biodiversidad.

En los ríos Ulla y Miño ya existen zonas de especial protección y están incluidas en la Red Natura 2000, siendo precisamente la presencia de estas especies uno de los criterios para designar algunas de esas zonas o áreas protegidas. En el medio marino también existen zonas protegidas, inmediatamente adyacentes a la costa, que están incluidas en la Red Natura 2000, siendo alguno de los criterios la presunta presencia de ambas especies de alosa. Sin embargo, aunque se dispone de abundante información sobre la utilización de los hábitats de agua dulce, sólo se dispone de datos fragmentarios sobre la forma en que los peces anádromos utilizan el medio oceánico, con una resolución espacial deficiente. Es conocida la presencia de ambas especies en el medio costero gallego, a partir de las capturas accidentales en pesquerías de interés comercial (Nachón *et al.* 2016), pero se desconocen las zonas exactas de captura, así como todo lo relativo a la biología y ecología de estas especies en el medio marino. Además, se sabe poco sobre la migración oceánica específica de las poblaciones o la mezcla de los peces procedentes de diferentes ríos (Nachón *et al.* 2016; 2020). Supuestamente algunas de las capturas incidentales en el mar de ambas especies ocurren a varios cientos de kilómetros de sus probables ríos de origen (Nachón *et al.* 2016). Se desconoce si la dispersión se ve favorecida por la mezcla en el mar de poblaciones de diferente origen natal o si se dispersan en distintos hábitats marinos (Martin *et al.* 2015; Nachón *et al.* 2020).

Por otro lado, Galicia presenta una colección de ríos cuyas condiciones geológicas y ambientales son compatibles con la existencia de poblaciones de alosa. Sin embargo, tan solo existen estudios poblacionales en los ríos Ulla y Miño (Nachón 2017). De esta forma, existen ríos que podrían albergar poblaciones de alosa que todavía no se han estudiado, siendo el alto coste en tiempo y recursos uno de los principales condicionantes.

1.2. Objetivos

En concordancia con lo expuesto anteriormente, se hace preciso identificar la correspondencia entre los stocks o poblaciones del medio marino y las poblaciones del medio continental a través del estudio de la microquímica de los otolitos de las alosas. La determinación individual del área natal es clave para la gestión, ya que permite identificar el stock al que pertenece el individuo en

caso de mezcla de poblaciones (King y Roche 2008). Este aspecto sería totalmente novedoso y constituiría una herramienta potente para la gestión, dado que permitirá supervisar de manera integral las poblaciones de alosas, en todos los medios ocupados por las mismas. El conocimiento del estado de las subpoblaciones de alosas del medio marino permitirá valorar el estado de conservación del hábitat que ocupan, así como la importancia de conservar un hábitat que acoge unas especies con tales particularidades ecológicas y biológicas. Además, poder deducir el río de origen de las alosas, nos permitirá cuantificar la tasa de *homing* de las mismas en los ríos gallegos, así como obtener datos de presencia en ríos donde todavía no se ha constatado la presencia de poblaciones asentadas. De forma específica, el presente proyecto persigue los siguientes 2 objetivos:

1. Estudiar la capacidad dispersiva de las poblaciones marinas de alosas gallegas y costas aledañas (Asturias, Cantabria y Portugal) y analizar así potenciales interacciones entre poblaciones que podrían indicar la existencia de una dinámica metapoblacional.
2. Estudiar el grado de filopatria (*homing*) de las alosas en los ríos gallegos y explorar la presencia de poblaciones en ríos que potencialmente pueden albergar poblaciones.

Con el fin de mostrar los resultados relativos a la consecución de dichos objetivos en el presente estudio, este se estructuró de forma que en primera instancia y como parte de este primer capítulo, se muestra un resumen de los procedimientos metodológicos empleados en el proyecto, así como una presentación del área de estudio y el origen de los datos. Finalmente, se hace un breve resumen de la metodología utilizada para el análisis de las muestras biológicas.

En el segundo capítulo, ahondamos en la metodología estadística aplicada para el estudio de los datos obtenidos de las muestras biológicas haciendo hincapié en la explicación de los modelos de clasificación empleados.

El tercer y cuarto capítulo recopilan respectivamente los resultados obtenidos tras el análisis de los datos y la interpretación que se lleva a cabo de estos, así como las conclusiones obtenidas a raíz de dichas interpretaciones.

1.3. Resumen de procedimientos

En este apartado, por un lado, se define el material biológico (otolitos) empleado para la realización del estudio, así como la teoría científica basada en la correlación de parámetros

biológicos en la que se centra el trabajo. Por otro lado, se realiza una breve revisión de los procedimientos estadísticos empleados en para analizar este tipo de datos.

Los otolitos son estructuras aragoníticas que se encuentran por pares en el oído interno de los peces óseos y que presentan un crecimiento continuo en forma de capas que se superponen (similar al crecimiento de los trocos de los árboles) a lo largo de la vida del individuo. La concentración de determinados elementos en las diferentes capas de crecimiento presenta una correlación significativa (coeficiente de correlación de Spearman) respecto a la concentración de dichos elementos en el agua ambiental, especialmente en agua dulce (Campana y Thorrold 2001; Walther y Thorrold 2006; Morais y Daverat 2016 en Nachón *et al.* 2020). Como consecuencia, en caso de que existan ríos natales que se diferencian significativamente en su composición microquímica, los núcleos de los otolitos pertenecientes a peces nacidos en cada uno de ellos registrarán las firmas únicas de cada área natal particular, permitiendo la identificación retrospectiva del origen natal (Campana y Thorrold 2001; Walther *et al.* 2008; Morais y Daverat, 2016 en Nachón *et al.* 2020). Así, las propiedades únicas de estas estructuras calcificadas las hacen especialmente útiles para desentrañar las características del ciclo vital, los patrones de movimiento de los peces diádromos, las capacidades de dispersión, los patrones de conectividad de las poblaciones y el origen natal de los peces (Nachón 2017).

En este estudio nos basamos en el análisis de la composición microquímica de los otolitos de ejemplares del género *Alosa* para establecer el origen natal de individuos adultos capturados tanto en el medio marino como en el fluvial con el fin de determinar sus capacidades dispersivas y el grado de filopatria. Para ello, exploraremos el rendimiento de diferentes metodologías de clasificación automática comprendidas en el campo del aprendizaje estadístico supervisado. El principal objetivo de estos clasificadores es la construcción de modelos, a través de los cuales, se harán predicciones.

Existe un amplio abanico de métodos de clasificación *machine learning* cuyo rendimiento varía en función del escenario en el que se utilice. Así pues, si bien existen metodologías ampliamente utilizadas por su robustez y flexibilidad en diversas situaciones, no existe una única metodología válida ni óptima para todos los casos. De esta forma, se vuelve de vital importancia comprender los mecanismos que subyacen a cada uno de estos métodos y saber escoger el más adecuado. En nuestro caso, nos topamos ante un problema de clasificación en el que nos valdremos de tan solo tres variables explicativas para determinar el río de origen de los ejemplares de *alosa*. Atendiendo a esto, se optó por el empleo de la metodología *boosting*, consistente en la combinación de árboles de decisión, para llevar a cabo la labor de clasificación. Los clasificadores basados en árboles de decisión pueden emplearse, entre otras cosas, para la asignación del río de origen de especies anádromas en función de la composición microquímica de sus otolitos

(Guinand *et al.* 2002). El *boosting* es un método de aprendizaje lento que consiste en la combinación de árboles de decisión pequeños que, *a priori*, serían malos predictores para dar lugar a un mejor predictor. Ya que tan solo disponemos de tres variables para realizar nuestras predicciones, en caso de crear un árbol de decisión, este gozaría de poca profundidad y es por ello que esta metodología es especialmente adecuada para este tipo de situaciones en la que los árboles poco profundos se muestran como predictores relativamente débiles.

Por otro lado, el rendimiento de esta metodología se comparará con otros métodos de clasificación automáticos generales también empleados para la asignación de especies anádromas a sus regiones de origen en función de la microquímica de otolitos y ríos (Smoliński *et al.* 2020). Los métodos contemplados fueron el *bagging*-bosques aleatorios (*random forest*), el análisis discriminante cuadrático (QDA) y el clasificador bayesiano ingenuo (*Naive Bayes*). Por último, cabe destacar que, debido a la falta de ejemplares juveniles de ciertos ríos contemplados, a lo largo de este estudio, se llevará a cabo una estimación de la microquímica teórica de los otolitos de los juveniles de estos ríos, así como un trabajo de simulación para la obtención de un stock de juveniles de los diversos medios fluviales con los que se entrenará y comprobará el rendimiento de los diferentes métodos de clasificación considerados. Con este objetivo, se emplearán datos auxiliares de la microquímica otolítica de juveniles de ríos franceses con los que aumentar la robustez de la estimación.

La totalidad del análisis estadístico llevado a cabo en el estudio se realizó mediante el uso del lenguaje de programación R. En concreto, se elaboró un código original a través del software R 4.1.2 y una serie de paquetes² que facilitaron el análisis estadístico de los datos, así como su representación gráfica. Todos los análisis realizados sobre los datos fueron llevados a cabo por el autor de esta memoria en el Laboratorio de Hidrobiología de la USC dirigido por el doctor Fernando Cobo Gradín.

² caret, doParallel, MASS, plyr, dplyr, tidyr, tibble, nortest, normtest, exactRankTests, lmttest, scales, ggplot2, corrplot y RColorBrewer.

1.4. Área de estudio y origen de los datos

Tal y como se puede observar en la figura 2, los datos referentes a la composición microquímica elemental e isotópica de los otolitos y ríos analizados en este estudio, proceden de diversos proyectos de índole europea (DiadES) y nacional (1MARDEALOSAS) centrados en el estudio de las especies diádromas y, en concreto, en el estudio de las alosas como es el caso de 1MARDEALOSAS. Este último ha sido un proyecto centrado en el estudio de las poblaciones de alosas en las costas gallegas en el que he tenido la oportunidad de participar desde su inicio e inspiró la conceptualización y elaboración del presente trabajo.

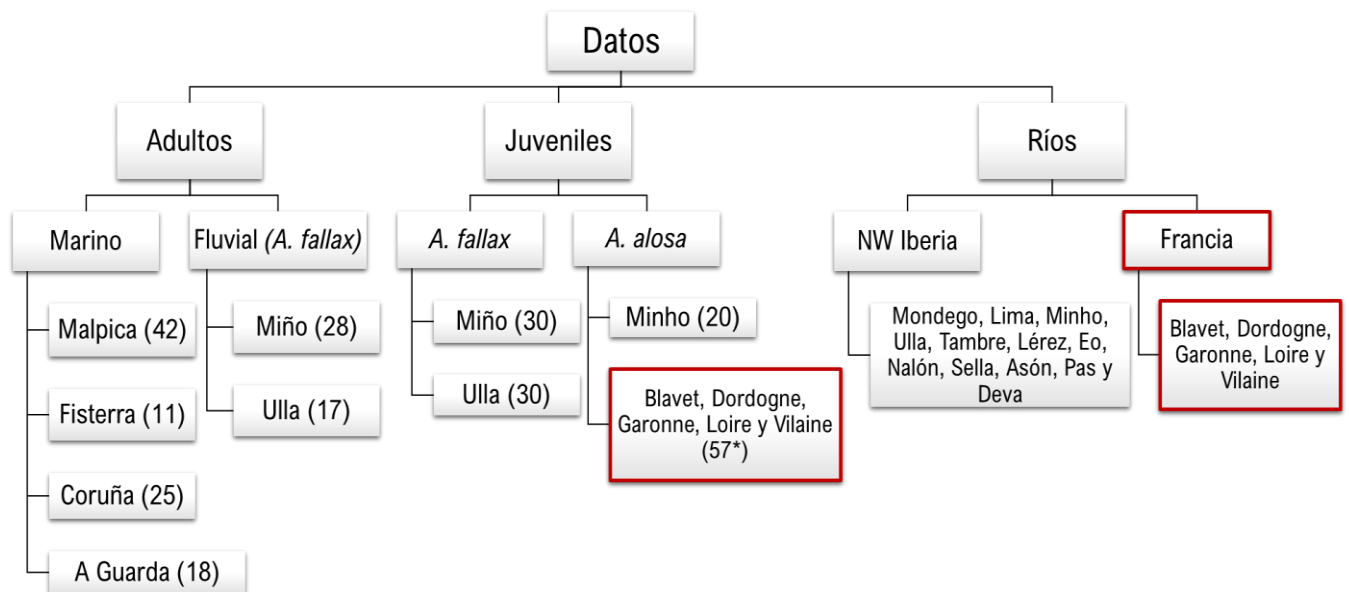


Figura 2: Esquema ilustrativo del material analizado para la realización del estudio. Los datos contorneados de rojo fueron empleados como información auxiliar relevante en el proceso de estimación de diversos parámetros.

Por un lado, la composición microquímica del agua de los ríos considerados en el proyecto, procede de muestreos llevados a cabo por diversos grupos de investigación, siguiendo un protocolo común en el que el agua fue almacenada en botellas estancas especializadas cuyo material impide la contaminación de la muestra y conservadas en neveras hasta su análisis. Los ríos gallegos seleccionados son los que presentan condiciones favorables para la existencia de poblaciones de alosas, por tanto, son candidatos potenciales a presentar poblaciones no detectadas todavía (como excepción tenemos los ríos Miño y Ulla, cuyas poblaciones se encuentran bien estudiadas).

Por otro lado, si bien una importante proporción de los datos de la composición microquímica de los otolitos provienen de estudios previos, los ejemplares procedentes del proyecto 1MARDEALOSAS han sido procesados y analizados con el propósito de realizar el presente

estudio. De este modo, el procedimiento aplicado en el análisis de estos ejemplares, que se mostrará a continuación, se puede extrapolar como la metodología aplicada al resto de especímenes utilizados para la elaboración del presente estudio.



Figura 3: Área de estudio considerada para la elaboración del proyecto correspondiente a la región noroeste de la península ibérica. Los ríos analizados aparecen resaltados en diferentes coloraciones. Fuente: elaboración propia.

1.5. Análisis de los otolitos

El análisis de la composición química del otolito requiere adoptar una serie de precauciones en el transcurso de su preparación con el fin de evitar una eventual contaminación que pudiese alterar su composición elemental (Proctor y Thresher 1998). En consecuencia, es preciso seguir varios pasos de descontaminación y preparación. Para ahondar en este aspecto metodológico, se recomienda la lectura del protocolo desarrollado por Tabouret (2009).

Una vez se han preparado las muestras, se someten a un proceso de ablación láser. Esta técnica consiste en focalizar una gran cantidad de energía luminosa en la superficie del otolito que produce un aerosol (con la composición microquímica del otolito) que es transportado a través de un gas vector hasta un espectrofotómetro de masas donde es analizado. Para ver un desarrollo más profundo de esta técnica, se recomienda la lectura de Martín (2013).

Capítulo 2

Metodología estadística

El presente capítulo se divide en seis secciones diferentes. En primer lugar, se aborda el cálculo de los parámetros biológicos faltantes y necesarios para la realización del estudio con los que se obtendrá una colección de datos sintéticos que permitan representar a los ejemplares juveniles nacidos en todos los ríos contemplados en el trabajo. A continuación, se muestran y comparan las composiciones microquímicas de los otolitos de los juveniles nacidos en los diferentes ríos. Posteriormente, se seleccionan los ríos y parámetros escogidos para la consecución de los objetivos del estudio. Tras esta selección previa, se presentan y definen los métodos de clasificación empleados para la ajuste de los modelos de clasificación. Del mismo modo, se explica el proceso llevado a cabo para el entrenamiento, validación y comparación de los modelos. Finalmente, se detalla tratamiento para la clasificación de los ejemplares adultos.

2.1. Estimación de parámetros y simulación de juveniles

2.1.1. Justificación

Tal y como se comentó anteriormente, las poblaciones de alosas europeas han experimentado un acusado declive en los últimos tiempos, motivo por el cual la obtención de ejemplares para su estudio se torna en un laborioso trabajo que consume una gran cantidad de recursos. Por otro lado, al ser especies cuyas poblaciones se intentan proteger, existe una profunda reticencia a retirarlas de su hábitat ya que se estaría incidiendo en su declive. Como resultado, la cantidad de muestras disponibles para la investigación de estas especies es muy limitada, de hecho, y aunque pareciese que el tamaño muestral empleado en el presente estudio se torna pequeño, es común encontrar en la literatura estudios basados en tamaños muestrales menores.

Centrándonos en el escenario del presente estudio, si bien disponemos muestras de juveniles de las poblaciones de alosas conocidas en los ríos Ulla y Miño, carecemos de juveniles procedentes de los ríos gallegos de donde se especula que pudiesen existir poblaciones todavía no estudiadas. De la misma forma, carecemos de datos de los ríos extracomunitarios que incluimos en el presente estudio. Sin embargo, conocer la composición microquímica de los juveniles es de vital importancia para la posterior clasificación de los ejemplares adultos. De esta forma, parece imprescindible hallar la manera de inferir la composición microquímica de los otolitos de los juveniles de las especie de *A. alosa* y *A. fallax* en los ríos de los que tan solo se tiene muestras de la composición microquímica de sus aguas.

Si bien como se ha mencionado, carecemos de muestras de juveniles de ciertos ríos objeto de estudio, se dispone de muestras de juveniles de una colección de ríos franceses donde estas

especies presentan poblaciones asentadas. Estos datos auxiliares se emplearon con el fin de inferir los valores de los parámetros necesarios para llevar a cabo el estudio.

Asimismo, una vez estimados los parámetros necesarios, se hace imprescindible la obtención de un tamaño muestral asequible y representativo de todos los ríos objeto de estudio a fin de poder entrenar de forma correcta los diferentes modelos de clasificación automática con el propósito de alcanzar el mejor rendimiento posible. En las siguientes secciones, se lleva a cabo un proceso de selección de metodología y estimación de los parámetros microquímicos propios de los juveniles de cada uno de los ríos con el fin de obtener una colección de datos sintéticos que emplear, junto a datos reales que ya poseemos, para el entrenamiento de los modelos de clasificación.

2.1.2 Selección de metodología

En estadística, las pruebas aplicadas a las muestras con el fin de obtener información a nivel poblacional a partir de las mismas se pueden dividir en paramétricas y no paramétricas. La diferencia fundamental entre ellas radica en que, mientras que las pruebas paramétricas precisan de unas condiciones más restrictivas para ser empleadas, las no paramétricas presentan unas condiciones menos restrictivas, más flexibles. Sin embargo, la ventaja de las pruebas paramétricas sobre las no paramétricas radica en que las primeras presentan mayor potencia estadística que las segundas en términos de contrastes de hipótesis y ofrecen mejores estimaciones siempre y cuando el modelo se encuentre bien especificado (Sedgwick 2015).

De esta forma, ante una variable aleatoria X con función de distribución de probabilidad F , i.e. $F(x) = P(X \leq x)$, decimos que esa variable X sigue un modelo paramétrico si su F asociada pertenece a una familia de distribuciones indexada por un parámetro θ de dimensión finita:

$$X \sim F \in F_{\theta} = \{F_{\theta}; \theta \in \Theta \subseteq \mathbb{R}^K\}.$$

Por otro lado, en base a la relación de dependencia existente entre la microquímica de los otolitos de los juveniles del género *Alosa* y la microquímica del agua de sus ríos de nacimiento se planteó la posibilidad de establecer un modelo de regresión lineal (como se describirá en la siguiente sección) con el fin de estimar la composición microquímica de los otolitos de los posibles juveniles procedentes de los ríos carentes de muestras de ejemplares. Si bien existen una serie de hipótesis básicas que deben contrastarse con el fin de validar el modelo de regresión lineal, este fue escogido, *a priori*, al existir diversas publicaciones que validan el uso de metodologías paramétricas en el análisis de las abundancias de estos elementos en los otolitos de las especies diádromas (Whalter y Torrold 2008; Wells *et al.* 2003).

Cabe destacar que, pese a existir estudios que se valen de la composición microquímica de los otolitos para el análisis de la filopatria en el género *Alosa*, (Martin *et al.* 2015; Randon *et al.* 2018) en ninguno de ellos se menciona la posibilidad de diferencias significativas en los coeficientes de asimilación de los elementos en los otolitos en función de la especie, tendiendo muchas veces a suponer que se comportan de forma muy similar por ser especies gemelas. En vista a esto, se procedió a realizar una prueba no paramétrica de localización entre las concentraciones elementales de los otolitos de ambas especies. En concreto, se aplicó la prueba de Mann-Whitney-Wilcoxon para comparar las poblaciones de ambas especies en el río Miño:

Siendo X_i e Y_j las concentraciones elementales de un ejemplar de *A. alosa* y *A. fallax* respectivamente:

- $X_1, \dots, X_m$ i. i. d. $X \sim F_x$, F_x continua
- $Y_1, \dots, Y_n$ i. i. d. $Y \sim F_y$, F_y continua
- $X_1, \dots, X_m$ e $Y_1, \dots, Y_n$ independientes

Para $i = 1, \dots, m$ y $j = 1, \dots, n$ se define $Z_{ij} = \mathbb{I}(X_i < Y_j)$ y entonces se construye el estadístico:

$$U_{m,n} = \sum_{i=1}^m \sum_{j=1}^n Z_{ij}$$

Z_{ij} contabiliza el número de Y_j 's mayores que algún X_i y $U_{m,n}$ el número de veces que las Y_j son mayores que algún X_i .

Este estadístico se empleará posteriormente en el subapartado 5.2 de este mismo capítulo con el fin de comprobar si las diferencias en el rendimiento de los diferentes modelos a la hora de clasificar de forma correcta las alosas adultas en su río natal son significativas.

2.1.3. Modelos de regresión lineal

Los modelos de regresión lineal representan la dependencia de una variable Y (variable dependiente, variable respuesta) con respecto a otra variable X (variable independiente, variable explicativa). La regresión se suele formalizar como la media condicionada de la variable respuesta en función del valor que tome la variable explicativa: $m(x) = \mathbb{E}(Y/X = x)$. Por tanto, podemos descomponer la variable respuesta en función del resultado de X (a través de la media condicionada), más un error de media cero: $Y = m(X) + \varepsilon$ donde ε es el error y verifica $\mathbb{E}(\varepsilon/X = x) = 0$.

En nuestro caso, diseñamos modelos de regresión lineal univariante para cada uno de los parámetros químicos objeto de estudio del mismo modo realizado en la literatura consultada

(Whalter y Torrold 2008; Wells *et al.* 2003). Como la concentración de Sr y Ba de los otolitos depende principalmente de la concentración de estos elementos en el agua (Bath *et al.* 2000; Farrell y Campana 1996) se establecieron modelos de regresión lineal univariante para estos elementos, considerando en cada caso como variable explicativa la concentración de Sr, Ba y la relación $\text{Sr}^{87}/\text{Sr}^{86}$ del agua y como variable respuesta la concentración de dichos parámetros en los otolitos de los juveniles de ambas especies.

Del mismo modo que la utilización de las abundancias elementales es un recurso útil en este tipo de análisis, la relación entre el isótopo Sr^{87} y el isótopo Sr^{86} ha demostrado ser un valioso indicador de la filopatria de peces diádromos (Gahagan *et al.* 2012). La ventaja que ofrece este parámetro respecto a las abundancias elementales reside en que, a diferencia de estas, la relación isotópica no presenta un coeficiente de partición entre el otolito y el ambiente, es decir, existe una relación 1:1 entre ambos medios (Whalter *et al.* 2008). Así pues, si bien no es necesaria la realización de un modelo de regresión en este caso, se decidió determinarlo para el conjunto de ambas especies con el fin de corroborar este hecho.

2.1.3.1. Hipótesis básicas de los modelos de regresión lineal

Con el fin de poder emplear de forma fiable los modelos lineales obtenidos, se precisa el cumplimiento de una serie de hipótesis básicas que deben ser corroboradas:

- **Homocedasticidad.** La varianza del error es la misma cualquiera que sea el valor de las variable explicativa: $\text{Var}(\varepsilon/X = x) = \sigma^2$ para todo x .
- **Normalidad.** El error tiene distribución normal: $\varepsilon \in N(0, \sigma^2)$.
- **Independencia.** Las variables aleatorias que representan los errores $\varepsilon_1, \dots, \varepsilon_n$ son mutuamente independientes, entendiéndose que vamos a obtener una muestra de n observaciones bajo el modelo de regresión. Esta suposición dice que los n errores (no observados) serían mutuamente independientes.
- **Linealidad.** La media de la variable respuesta toma un valor inicial cuando la variable explicativa vale cero y además dicha media crece una cantidad fija cada vez que la variable explicativa aumenta en una unidad.

Las hipótesis de homocedasticidad, normalidad e independencia se refieren a los errores de la regresión, pero como no observamos los errores, debemos aproximarlos a través de los residuos. Aunque los residuos no coincidan uno a uno con los correspondientes errores, el comportamiento global que muestra el conjunto de los residuos seguirá reflejando el comportamiento que siguen los errores y, en definitiva, el posible incumplimiento de las hipótesis (Cox y Snell 1968).

2.1.3.2. Pruebas estadísticas para la validación de los modelos de regresión lineal

Existen múltiples pruebas estadísticas que permiten contrastar las hipótesis de los modelos de regresión lineal. De esta forma, suponiendo independencia de los datos, en el presente trabajo se emplearon las siguientes pruebas:

- **Prueba de Shapiro-Wilk:** Prueba empleada para el contraste de normalidad. Bajo la hipótesis nula (H_0), una muestra $x_1, \dots, x_n$ proviene de una población normalmente distribuida.
- **Prueba de Harrison-McCabe:** Prueba de heteroscedasticidad basada en los residuos de mínimos cuadrados ordinarios. Bajo H_0 existe homogeneidad de la varianza en todo el rango de la variable.
- **Prueba de Bowman:** Este test no paramétrico se utiliza para detectar la no dependencia lineal entre una variable respuesta y las explicativas. Bajo H_0 la variable respuesta depende linealmente de los regresores.

2.1.3.3. Estimación de parámetros y simulación de juveniles

A través los modelos de regresión establecidos en el apartado anterior, se estimaron los valores de Sr, Ba y $\text{Sr}^{87}/\text{Sr}^{86}$ que, de forma teórica, presentan los juveniles de los diferentes ríos de los que carecemos de muestras biológicas. El cálculo de estas nuevas observaciones se realizó de la forma tradicional, sustituyendo en la recta de regresión el valor de x , y calculando el valor $\hat{Y}$.

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Siendo $\hat{\beta}_0$ y $\hat{\beta}_1$ los parámetros estimados de la recta de regresión.

Tras realizar las estimaciones pertinentes, se pudo obtener una muestra de datos sintéticos con los que simular juveniles de los ríos de interés y poder así conformar una muestra representativa de los diferentes ríos. Para ello, sencillamente se generaron valores aleatorios para cada parámetro provenientes de una distribución $N(\mu_i, \sigma_i)$, donde $i = 1, \dots, n$, representa cada uno de los cursos fluviales. Las estimaciones de la concentración de los parámetros en cada río y el error estándar asociado a cada valor estimado, fueron los parámetros empleados para la generación de los datos sintéticos.

Como resultado del proceso de simulación, encontramos la forma de obtener muestras válidas para el entrenamiento de los clasificadores automáticos que nos permitan dictaminar de forma eficaz el lugar de origen de las alosas adultas. Sin embargo, y siendo esto explicado de forma más extensa en el capítulo final, sería necesario validar los resultados valiéndolos de muestras reales para el entrenamiento de los modelos de clasificación.

2.2. Análisis exploratorio de la microquímica de los juveniles

2.2.1. Diagramas de dispersión de los parámetros microquímicos

La composición microquímica del agua de los ríos es otorgada por el sustrato a través del cual estos discurren a lo largo de todo su recorrido. Por consiguiente, la concentración de determinados elementos químicos y la relación entre estos dota a cada curso fluvial de una identidad que permite diferenciar a los ríos entre sí.

En base a este hecho y como consecuencia directa de la relación existente entre la composición microquímica de los ríos y la composición microquímica de los otolitos de las alosas, se procedió a elaborar una serie de diagramas de dispersión, a partir del stock de juveniles obtenido, que permitiesen dilucidar de forma clara la identidad de cada una de las cuencas fluviales.

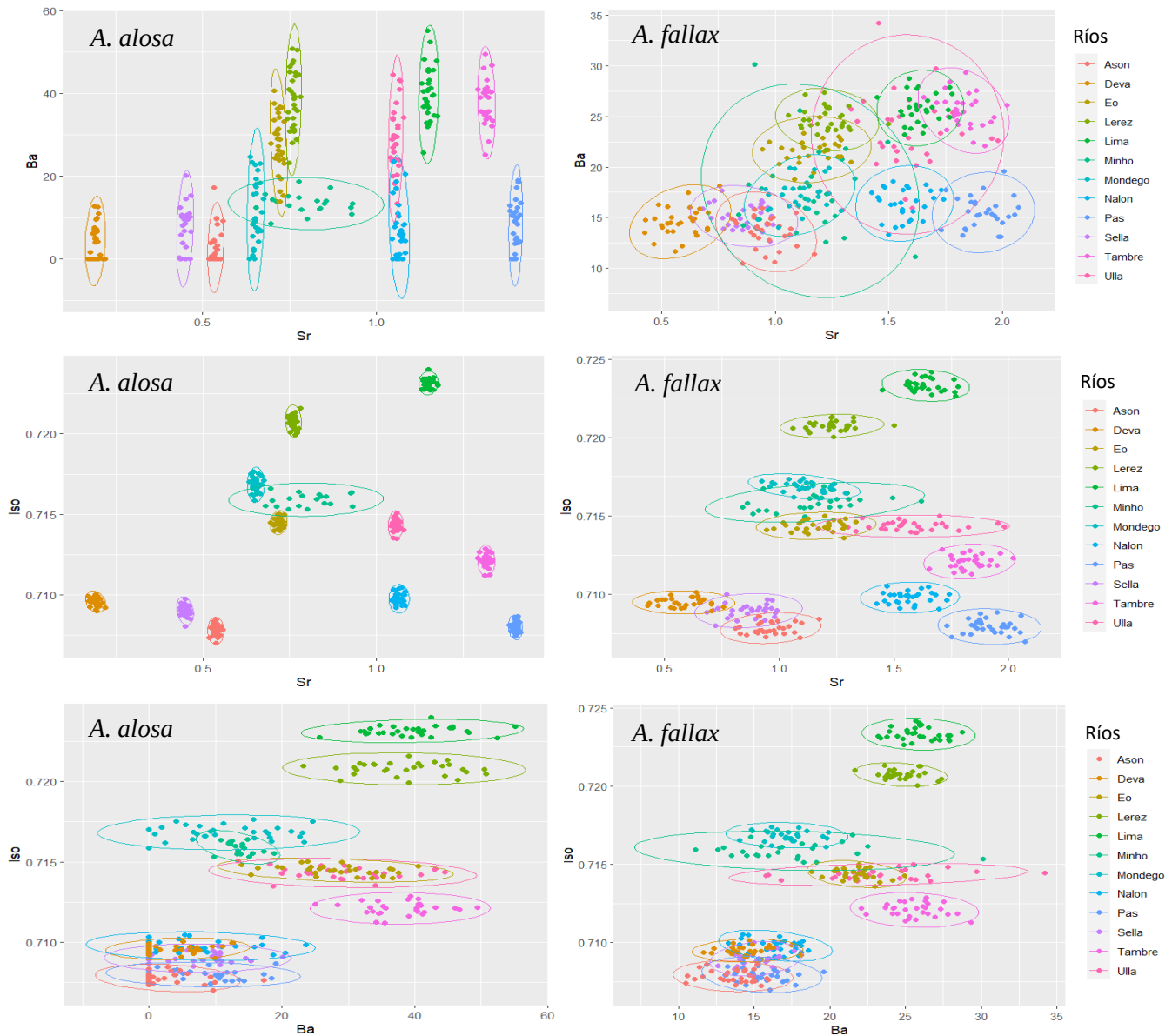


Figura 4: Concentraciones de bario (Ba), estroncio (Sr) y Sr87/Sr86 (Iso) correspondientes a la microquímica de los otolitos de los juveniles pertenecientes a cada uno de los ríos estudiados.

Observando la figura 4, nos percatamos de que el ratio isotópico del Sr (Iso) es un parámetro que permite una gran discriminación entre los diferentes cursos fluviales. Por otro lado, y en relación con las concentraciones elementales, parece que el Sr caracteriza la identidad de cada río de mejor forma que el Ba, siendo los valores de concentración de este último los que parecen tener una mayor variabilidad en cada río.

2.2.2 Análisis de similitud entre juveniles de distinto origen

Si bien la composición microquímica del agua de los ríos otorga una marca identificatoria a cada río y cuenca hidrológica (Randon *et al.* 2018), existe la posibilidad de que cuencas que se asientan sobre sustratos geológicos semejantes, presenten ríos cuya composición sea muy similar. Como consecuencia, es posible que la composición microquímica de los otolitos de los juveniles de ríos diferentes se aproximen en gran medida, lo que podría dificultar su correcta clasificación.

Con el fin de analizar en que grado dista la composición microquímica de los otolitos pertenecientes a los juveniles de los diferentes ríos, se optó por calcular la distancia de Canberra. Esta medida, es una versión ponderada de la distancia Manhattan, introducida y refinada en 1967 por Lance, Williams y Adkins. A menudo se utiliza para los datos dispersos alrededor de un origen, ya que es muy sensible para valores próximos a cero, además, es insensible a cambios de escala (Mirzaei *et al.* 2017).

De esta forma, sean $x'_i = (x_{i1}, \dots, x_{ip})$ y $x'_j = (x_{j1}, \dots, x_{jp})$ las observaciones de dos ríos (l, k) resultado de medir p variables $X_1, \dots, X_p$ sobre ellos, la distancia de Canberra es:

$$\delta_c(l, k) = \sum_{i=1}^p \frac{|x_{li} - x_{ki}|}{|x_{li}| + |x_{ki}|}$$

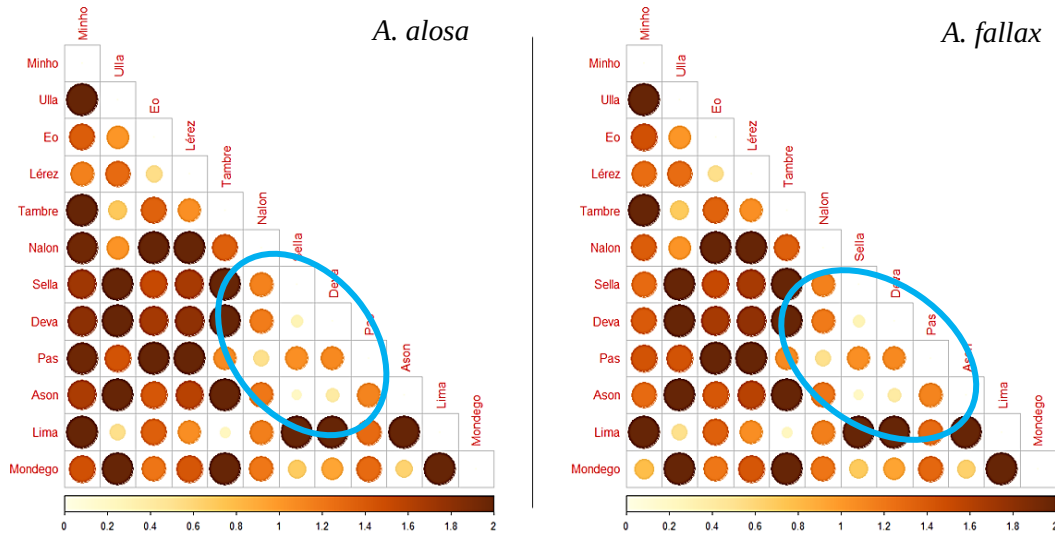


Figura 5: Matriz de distancias entre la composición microquímica de los diferentes ríos contemplados en el estudio. La distancia fue calculada en función a las concentraciones de estroncio (Sr), bario (Ba) y el ratio isotópico de los isótopos del Sr.

Al contemplar la figura 5 observamos que los ríos de la cornisa cantábrica presentan, en general, un mayor parecido entre sí (elipsis azul). Esto se debe a que la región geográfica por la que discurren estos ríos de la vertiente cantábrica está formada por material calcáreo que dota de una mayor homogeneidad química a los cursos fluviales que discurren sobre él.

2.3. Selección de ríos y parámetros para los diferentes análisis

2.3.1. Selección de ríos

Los ejemplares de *A. alosa* y *A. fallax* presentan una marcada tendencia a regresar a sus ríos de origen para reproducirse (Martin *et al.* 2015; Randon *et al.* 2018). No obstante, existen evidencia de que, con relativa frecuencia, ejemplares que se encuentran realizando la migración con el fin de reproducirse, se adentran en ríos próximos a su río de origen, existiendo de esta forma un comportamiento filopátrico “parcial” ya que, si bien los ejemplares son capaces de regresar a la zona costera próxima a su río de origen, no necesariamente se introducen en el mismo río (Nachón *et al.*, 2020). De esta forma, existen casos de ejemplares reproductores del río Miño que ingresan en el Ulla y viceversa (Nachón, 2016). Más aún, se especula que existen casos de alosas naturales de las cuencas gallegas que remontan los ríos Mondego y Lima, pertenecientes a Portugal. Por otro lado, los ríos gallegos Lérez y Tambre, presentan las condiciones adecuadas para albergar poblaciones de alosas, aunque se desconoce su presencia o ausencia en dichos ríos.

A consecuencia de lo expuesto, se decidió que para el estudio de la filopatria se analizarían los ríos Mondego, Lima, Lérez, Tambre, Miño y Ulla debido a su proximidad. El motivo por el que los ríos de la cornisa cantábrica no se contemplan en este estudio se debe, principalmente, a que

se encuentran lo suficientemente distanciados del Miño y Ulla como para suponer que los casos de ingesión de los ejemplares de origen gallego en los ríos del norte peninsular, es muy infrecuente. (Randon *et al.* 2018)

En contraste, para el análisis de la dispersión en el medio marino, sí que se tienen en consideración la totalidad de los ríos analizados, ya que ambas especies de alosa presentan rangos de dispersión lo suficientemente amplios como para comprender el área de estudio al completo (Martin *et al.* 2015). De esta forma, es plausible la idea de capturar en las costas de Galicia ejemplares procedentes de ríos portugueses y de la vertiente cántabra.

2.3.2. Selección de parámetros químicos

A consecuencia de los diversos orígenes de las muestras, podemos clasificar las alosas adultas en dos grupos atendiendo a la información disponible. Por un lado, de los ejemplares adultos capturados en río, se poseen datos de la concentración de Ba y Sr. Valiéndonos de estos ejemplares calculamos el grado de filopatria que presentan ambas especies. Por otro lado, disponemos de datos de abundancia del Ba y Sr, así como de la relación isotópica $\text{Sr}^{87}/\text{Sr}^{86}$. A través de estos datos determinamos el grado de dispersión durante la fase marina. (Martin *et al.* 2015; Randon *et al.* 2017).

A diferencia de lo que ocurre con los ejemplares capturados durante la fase marina, no tenemos datos de la relación isotópica de los ejemplares reproductores capturados en los ríos. Así pues, los modelos de clasificación automática diseñados para los ejemplares en fase reproductora y los ejemplares de la fase marina difieren en que los segundos, a diferencia de los primeros, contemplan la relación isotópica como parámetro para la clasificación.

2.4. Métodos de clasificación

En la presente sección se abordan las diferentes metodologías de clasificación, basadas en el *machine learning*, contempladas para llevar a cabo la asignación a su río de origen de los ejemplares de alosa adultos capturados tanto en el medio fluvial como en el medio marino. En concreto, se realizará una breve introducción de dos métodos de clasificación basados en árboles de decisión y de dos métodos pertenecientes a la familia de los métodos generadores. Además, se ofrece una explicación a grandes rasgos del funcionamiento de cada uno de los algoritmos considerados.

Los diferentes algoritmos tenidos en cuenta en la consecución del estudio tienen en común que, para su ajuste y entrenamiento, se emplearon como variables de entrada las concentraciones elementales del Sr y Ba, así como la especie a la que pertenece cada uno de los ejemplares. Como variable objetivo y salida del modelo obtenemos el río de origen de la alosa adulta. Cabe destacar

que, como ya se comentó en apartados anteriores, para el estudio del rango dispersivo en el medio marino contamos, como variable de entrada extra, con la relación isotópica $\text{Sr}^{87}/\text{Sr}^{86}$.

Gracias a los métodos de clasificación contemplados en este apartado, seremos capaces de ajustar modelos que permitan dilucidar la región de origen de las alosas adultas. Como consecuencia, podremos indagar en el grado de filopatria presente en las poblaciones del noroeste peninsular, así como en el rango dispersivo que ambas especies presentan en esta región, siendo estos aspectos clave para la consecución de los objetivos del trabajo.

Cabe destacar que los subapartados correspondientes a la presente sección, son una síntesis de los conceptos e ideas impartidas en la asignatura de Aprendizaje Estadístico del Máster de Técnicas Estadísticas impartida por los profesores Rubén Fernández Casal y Manuel Oviedo³ de la fuente en el curso 2021/2022 y que han sido adaptadas, particularizadas y completadas con otras fuentes/referencias al presente caso de estudio.

2.4.1. Métodos basados en árboles de decisión

El árbol de decisión nació en la década de 1970 como un método alternativo para los métodos estadísticos que dependen de las hipótesis de normalidad y linealidad de las variables. La idea base de este método subyace en la fragmentación del espacio predictor en regiones tan simples que permitan la representación del proceso como un árbol binario. De esta forma, partimos de un nodo inicial que representa a toda la muestra. Este nodo se divide en dos nuevos nodos que dividen la muestra en dos subgrupos. Este proceso se repite hasta llegar a las "hojas" del árbol, que son los nodos terminales utilizados para hacer predicciones por medio de la media en un problema de regresión o, en nuestro caso, la categoría modal en un problema de clasificación. (Fernández *et al.* 2021)

Los nodos se dividen dando respuesta a una pregunta dicotómica sobre una variable predictora seleccionada de forma previa (figura 6). A través de este procedimiento, se busca que los nodos terminales sean homogéneos respecto a la variable respuesta. En el caso de los árboles de clasificación, obtendríamos nodos puros cuando en los nodos terminales tan solo se observarían elementos de una clase. Pese a que siempre es posible la obtención de nodos puros, en la mayoría de las ocasiones no sería conveniente, ya que esto supondría la construcción de árboles muy profundos que se traducirían en modelos excesivamente complejos. La desventaja de estos modelos radica en su difícil interpretación y en un posible problema de sobreajuste. Así pues, lo ideal sería hallar un equilibrio entre la complejidad del modelo y la pureza de las hojas (Fernández *et al.* 2021).

³ Esta información puede ser consultada en los apuntes facilitados por los profesores y disponibles en https://rubenfcasal.github.io/aprendizaje_estadistico

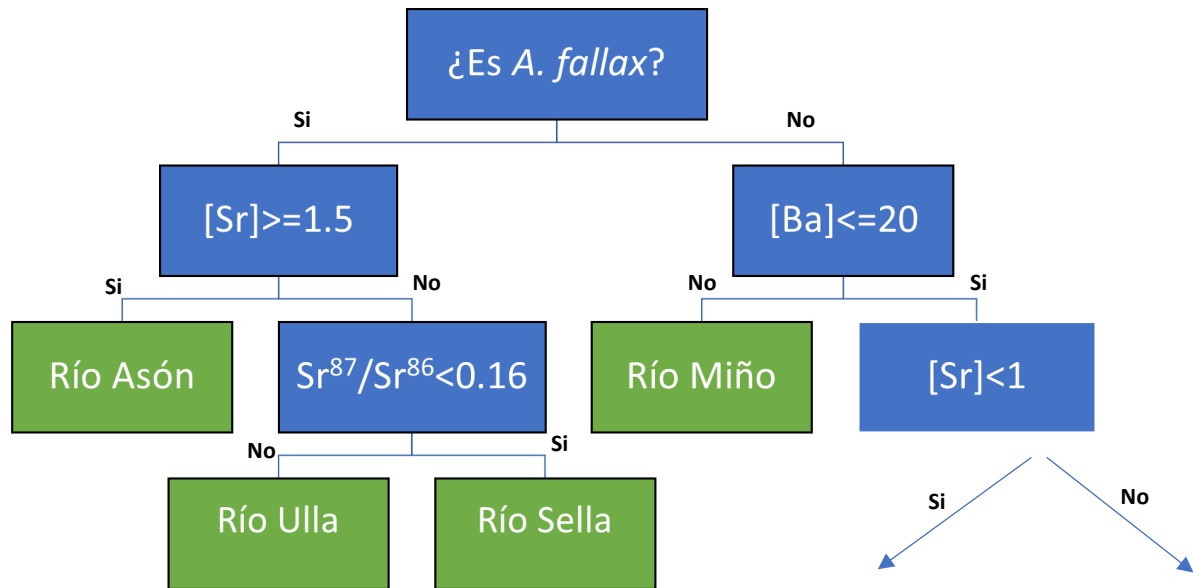


Figura 6: Ejemplo de árbol de decisión basado en la temática del presente estudio. Fuente: elaboración propia.

Esta metodología, si bien no depende de las hipótesis de normalidad y linealidad, no goza de una gran capacidad predictiva. No obstante, se trata de una técnica básica del aprendizaje estadístico, ya que constituye los cimientos de métodos eficaces tales como el *boosting*, donde se usan árboles de decisión poco profundos o el *bagging*, metodología que emplea árboles sin podar. En el presente estudio se utilizaron ambas metodologías basadas en árboles de decisión, las cuales se presentan a continuación, con el fin de alcanzar los objetivos estipulados. (Fernández *et al.* 2021)

2.4.2. Boosting

2.4.2.1. Conceptualización

La metodología *boosting* se basa en la combinación de árboles de decisión poco profundos, malos predictores, con el fin de conseguir un predictor de mayor potencia. La idea original fue desarrollada por Valiant (1984), sin embargo, no fue hasta que Freund y Schapire (1996) presentaron el algoritmo *AdaBoost*, que esta metodología se popularizó entre la comunidad científica. Este algoritmo no combina árboles independientes, sino que estos se van generando en una secuencia en la que cada árbol depende del anterior. Se utiliza siempre el mismo conjunto de datos (de entrenamiento), pero a estos datos se les otorga pesos en cada iteración que dependen de lo que ha ocurrido en la iteración anterior. Se les da más peso a las observaciones mal clasificadas para que en sucesivas iteraciones se clasifiquen bien. Finalmente, la combinación de los árboles se hace mediante una suma ponderada de las clasificaciones realizadas. (Fernández *et al.* 2021)

Posteriormente, Friedman solventó varios problemas del *AdaBoost* profundizando en los métodos iterativos de descenso de gradientes y formulando el método *gradient boosting machine*, el cual, se basa en encontrar un modelo aditivo que minimice una función de pérdida combinando árboles de clasificación poco profundos (clasificadores débiles). (Fernández *et al.* 2021)

2.4.2.2. Pseudocódigo general

Existen múltiples algoritmos *boosting* propuestos por diferentes investigadores, sin embargo, en nuestro caso se expone el algoritmo SAMME (*Stagewise Additive Modeling using a Multi-class Exponential loss function*), el cual se encuentra enfocado a problemas de clasificación multiclase (Zhu *et al.* 2009), como es el caso de estudio de este trabajo.

Suponiendo una submuestra de entrenamiento $(x_1, y_1), \dots, (x_n, y_n)$ donde $x_i \in \mathbb{R}^p$ es el vector de variables explicativas del dato i -ésimo e y_i la variable respuesta cualitativa que puede tomar valores $1, \dots, K$ siendo K el número de clases, se debe ajustar un clasificador $Y(x)$ al que cuando se le introduzcan los valores de x , asigne la observación a una clase c (de las clases $1, \dots, K$) de la siguiente forma:

1. Establecer pesos iniciales para las observaciones $w_i = \frac{1}{n}, i = 1, \dots, n$.

2. Repetir de $b = 1$ a $b = B$:

(a) Ajustar un clasificador $T^{(b)}(x)$.

(b) Calcular una medida de error

$$err^{(b)} = \sum_{i=1}^n w_i \mathbb{I}(c_i \neq T^{(b)}(x_i)) / \sum_{i=1}^n w_i.$$

(c) Calcular

$$\alpha^{(b)} = \log \left(\frac{1 - err^{(b)}}{err^{(b)}} \right) + \log(K - 1).$$

(d) Restablecer

$$w_i \leftarrow w_i \cdot \exp \left(\alpha^{(b)} \cdot \mathbb{I}(y_i \neq T^{(b)}(x_i)) \right), \quad i = 1, \dots, n.$$

3. Salida

$$Y(x) = \arg \max_{k=1, \dots, K} \sum_{b=1}^B \alpha^{(b)} \cdot \mathbb{I}(T^{(b)}(x) = k).$$

2.4.2.3. Hiperparámetros

Los hiperparámetros de un modelo son valores que deben ser prefijados para “correr” los algoritmos. Son valores que a menudo no se extraen de los datos de entrenamiento, por lo que

normalmente son facilitados por el analista de datos. El valor óptimo de un hiperparámetro no puede conocerse *a priori* para un problema determinado, debido a esto, estos no aparecen reflejados en el pseudocódigo. De forma excepcional, el hiperparámetro número de árboles (B) sí que aparece reflejado en el pseudocódigo al indicar el número de iteraciones realizadas en uno de los pasos del algoritmo.

Esta metodología depende, principalmente, de tres hiperparámetros susceptibles a ser seleccionados de forma óptima:

- **Número de árboles (B):** En cada iteración, el objetivo es ajustar de forma óptima el gradiente (en nuestro caso, los residuos), pero este enfoque *greedy* no garantiza el óptimo global y puede dar lugar a sobreajustes.
- **Parámetro de regularización (λ):** Regula la velocidad de aprendizaje o *learning rate*. Cuanto más pequeño sea el valor de λ más lento va a ser el proceso de aprendizaje y serán necesarias más iteraciones, lo cual incrementa los tiempos de cómputo. No obstante, los valores pequeños de λ evitan el problema de sobreajuste.
- **Número de divisiones del árbol (d):** Limita la profundidad de los árboles de clasificación empleados para construir el modelo, si su valor es de 1, tendremos un modelo aditivo.

Con el fin de seleccionar los modelos más precisos, entendiendo como precisión el porcentaje de predicciones correctas respecto al total de predicciones, los anteriores hiperparámetros fueron seleccionados de entre una colección de valores o rejilla mediante un proceso de validación cruzada. De forma específica, se realizaron 5 repeticiones de validación cruzada de 10 pliegues en los datos de entrenamiento, utilizando un conjunto diferente de pliegues para cada validación cruzada. Para más detalles en la selección de los hiperparámetros, se puede consultar el Anexo.

2.4.2.4. Stochastic gradient boosting

En 1999 el propio Friedman, basándose en la técnica *bagging* desarrollada por Breiman, refinó el algoritmo que él mismo había formulado, estableciendo así el método *stochastic gradient boosting*. El único cambio respecto al *gradient boosting machine*, reside en que en cada iteración del modelo no se utiliza la totalidad de la muestra de entrenamiento, sino que se escoge una submuestra de forma aleatoria (70% de las observaciones generalmente). La principal ventaja de este método respecto a su predecesor radica en la reducción de la varianza y de los tiempos de cómputo. (Fernández *et al.* 2021)

El *stochastic gradient boosting* es una metodología muy popular debido a su elevado rendimiento, es por esto por lo que en el presente trabajo la empleamos como clasificador *boosting*.

2.4.3. *Bagging – random forest*

2.4.3.1. Conceptualización

El *bagging* o *bootstrap aggregating* es una metodología combinada desarrollada por Breiman en 1996 basada en la utilización del remuestreo *bootstrap* junto con un modelo de regresión o de clasificación como es nuestro caso. Consiste en promediar predicciones generadas a partir de distintas muestras de entrenamiento. De esta forma, se entrena un modelo para cada muestra de entrenamiento, por lo que obtendremos tantos modelo y predicciones como muestras. (Fernández *et al.* 2021)

La principal ventaja de este método es que permite reducir drásticamente la varianza y estimar el error de predicción de forma directa, evitando así el uso de una muestra de prueba o aplicar validación cruzada. No obstante, se hace preciso el empleo de múltiples muestras de entrenamiento. Debido a que en la práctica esto es difícilmente alcanzable, se emplea la metodología de remuestreo *bootstrap* para, a través de una única muestra de entrenamiento, obtener múltiples muestras que serán empleadas para el entrenamiento de modelos los diferentes modelos cuyas predicciones serán promediadas. (Fernández *et al.* 2021)

En el caso particular de los modelos de clasificación, en realidad en lugar de trabajar con la media, se debe sustituir esta por la categoría modal y emplear el criterio del “voto mayoritario”, en el cual, todos los modelos presentan el mismo peso y cada modelo aporta un voto. Asimismo, la proporción de votos de cada categoría es una estimación de su probabilidad. Fernández *et al.* 2021)

2.4.3.2. Pseudocódigo general

Suponiendo una submuestra de entrenamiento $S = (x_1, y_1), \dots, (x_n, y_n)$ donde $x_i \in \mathbb{R}^p$ es el vector de variables explicativas del dato i -ésimo e y_i la variable respuesta cualitativa que puede tomar valores $1, \dots, K$ siendo K el número de clases, se debe ajustar un clasificador $Y(x)$ al que cuando se le introduzcan los valores de x , asigne la observación a una clase c de la siguiente forma:

1. *Repetir de $b = 1$ a $b = B$:*

(a) $S_b =$ muestra *bootstrap* de S .

(b) Ajustar un clasificador $T^{(b)}(x)$ con S_b .

2. Salida

$$Y(x) = \arg \max_{k=1,\dots,K} \sum_{b=1}^B \mathbb{I}(T^{(b)}(x) = k).$$

2.4.3.3. Hiperparámetros

Esta metodología depende, principalmente, de tres hiperparámetros susceptibles a ser seleccionados de forma óptima:

- **Número de árboles (B):** Se generarán tantos árboles como muestra *bootstrap* y predicciones, se puede contemplar como el número de iteraciones.
- **Número de predictores seleccionados (m):** Es el número de predictores escogidos al azar en cada división. Se recomienda comenzar con $m = \sqrt{p}$ predictores y optimizar desde ahí en los problemas de clasificación.
- **Número mínimo de observaciones en el nodo final.**

El proceso de selección de los hiperparámetros se llevó a cabo de igual forma que el expuesto en el apartado del *boosting*.

2.4.3.4. Random forest

En la construcción de los múltiples modelos, puede ocurrir que una buena proporción de estos presente una estructura muy similar pese a la aleatoriedad introducida por el remuestreo *bootstrap*. Esta condición de semejanza estructural está presente especialmente en la parte alta de los árboles, cuando existen variables que, debido a su relevancia en la clasificación, se van a encontrar siempre en los primeros puntos de corte. Este fenómeno de correlación entre árboles fue solventado por Breiman en 2001 al modificar el algoritmo original. Esta modificación consistió en seleccionar al azar $q < p$ predictores candidatos para realizar los cortes antes de realizar dichos cortes. A esta metodología Breiman la denominó *random forest* y es más eficiente que el *bagging* pese a necesitar un mayor número de árboles, ya que la construcción de los mismos es más rápida al contemplar un menor número de predictores para su generación. (Fernández *et al.* 2021)

Debido a la mayor eficiencia del *random forest* sobre el *bagging* original, los empleamos como clasificador en el presente estudio. No obstante, al ser el *bagging* un caso determinado de *random forest* donde $p = q$, lo contemplamos como posible clasificador al ser el número de predictores

un hiperparámetro seleccionado por validación cruzada e incluir como opción que se tome igual número de predictores como variables hay.

2.4.4. QDA

El QDA es un método de clasificación perteneciente a la familia de los clasificadores denominados modelos generadores, los cuales son modelos en los que la clave reside en la distribución de las clases por separado de un conjunto de datos y los algoritmos de aprendizaje tienden a modelar los patrones/distribución subyacentes de los puntos de datos. Estos modelos utilizan la intuición de la probabilidad conjunta teórica, creando escenarios en los que una variable explicativa (X), o una colección de variables explicativas en el caso multidimensional, y una respuesta (Y) existen al mismo tiempo.

En el caso del QDA, las probabilidades *a posteriori* de un individuo de pertenecer a un determinado grupo son calculadas en base a la distribución conjunta de (X, Y) y al teorema de Bayes (Fernández *et al.* 2021):

$$P(Y = k|X = x) = \frac{P(Y = k)f_k(x)}{\sum_{l=1}^k P(Y = l)f_l(x)}$$

donde $f_k(x)$ es la función de densidad del vector aleatorio $X = (X_1, \dots, X_n)$ de observaciones de la muestra pertenecientes a la clase k .

En el análisis discriminante cuadrático, asumiendo que $X|Y = k \sim N(\mu_k, \Sigma_k)$, se obtienen las siguiente funciones discriminantes cuadráticas:

$$-\frac{1}{2}(x - \mu_k)^t \sum_K^{-1} (x - \mu_k) - \frac{1}{2} \log(|\Sigma_k|) + \log(P(Y = k)).$$

Esta metodología es recomendable tan solo cuando el tamaño muestral es muy superior al número de predictores empleados. En nuestro caso, tan solo presentamos cuatro predictores por lo que es un método plausible para la clasificación.

2.4.4.1. Clasificador bayesiano ingenuo (Naive Bayes)

Este modelo, al igual que el QDA, en la familia de los modelos generadores. Sin embargo, el modelo *Naive Bayes* asume que las variables predictoras son independientes, simplificando de esta forma el caso anterior pese a que, en la práctica, esta suposición difícilmente se cumpla. De hecho, este método se puede ver como un caso determinado de QDA con matrices Σ_k diagonales. A raíz de la simplicidad del modelo, se puede trabajar con un tamaño muestral moderadamente menor que en el anterior caso y, en muchas ocasiones, obtener un alto rendimiento. (Fernández *et al.* 2021)

Cabe destacar que, tanto para este método como para el análisis cuadrático discriminante, se hace necesaria la asunción de normalidad de los datos. Es decir, a diferencia de los clasificadores basados en los árboles de decisión, en este caso la distribución de las variables predictoras es un factor a tener en cuenta.

2.5. Entrenamiento, evaluación y comparación de los modelos

Una vez especificados las diferentes metodologías contempladas en el estudio para la clasificación de los ejemplares a su río natal, se hace imprescindible la obtención de modelos de clasificación eficaces basados en dichas metodologías para la consecución de nuestros objetivos.

Así pues, en el presente apartado nos centramos en aquellos aspectos relacionados con el ajuste, y evaluación de los modelos contemplando diferentes escenarios en función de la mayor o menor disponibilidad de muestras. Tras esto, se comparará el rendimiento de los diferentes modelos con el fin de escoger aquel con el que llevar a cabo la clasificación final de las alosas adultas.

2.5.1. Entrenamiento y partición de los datos

Los modelos de clasificación necesitan que se les proporcione una muestra de entrenamiento con el fin de que el algoritmo en el que se basa el modelo encuentre patrones en los datos y generar así un modelo de clasificación que reconozca dichos patrones y asigne correctamente el, en este caso, río de origen del ejemplar. No obstante, para poder cuantificar la capacidad predictiva de los modelos, es necesario ofrecerle al modelo una muestra formada por observaciones de las que, si bien se conoce su variable respuesta, estas no han participado en el ajuste (entrenamiento) del modelo. Es por esto, que del stock total de juveniles de los que se dispone para la realización del estudio, se crearon 2 subconjuntos muestrales; una muestra de entrenamiento (70% de las observaciones), con el fin de ajustar los modelos, y una muestra de test (30% de las observaciones), con el fin de cuantificar su capacidad predictiva.

2.5.2. Evaluación y comparación de modelos

Tal y como se comentaba en el apartado anterior, el fin de un modelo es predecir, con la mayor exactitud posible, la variable respuesta de una observación que dicho modelo “no ha visto” con anterioridad. Con el fin de evaluar la precisión o *accuracy* de los modelos empleamos la muestra de test. No obstante, de forma previa al empleo de esta partición de prueba, se llevó a cabo un proceso iterativo de validación cruzada similar al llevado a cabo para la selección de los hiperparámetros de los modelos. La idea que subyace a la realización de este proceso es la de ajustar y evaluar el modelo múltiples veces (5 repeticiones) con distintos subconjuntos (10 subconjuntos o pliegues) creados a partir de los datos de entrenamiento, obteniendo en cada repetición una estimación del error. El promedio de todas las estimaciones (50 estimaciones por

modelo) tiende a converger en el valor real del error de test que cuantifica las veces que, en nuestro caso, el río asignado al juvenil por el modelo no coincide con el río al que pertenece de verdad.

Posteriormente, se calculó la precisión o *accuracy* de los modelos empleando la muestra de test y se comparó esta medición con la obtenida en el proceso de remuestreo anterior.

Por último, los valores de precisión o *accuracy* obtenidos por remuestreo de cada uno de los diferentes modelos fueron comparados dos a dos a través de la prueba de Mann-Whitney-Wilcoxon para muestras pareadas.

2.5.2.1. Tamaños muestrales

El hecho de poder simular la composición microquímica de los otolitos de ejemplares juveniles de los diferentes ríos, nos permite trabajar con diferentes tamaños muestrales. Debido a que las poblaciones de alosas se encuentran en acusado declive y rara vez se puede disponer de un amplio tamaño muestral de juveniles, se hace interesante explorar el comportamiento de los modelos ante diferentes tamaños muestrales, ya que modelos eficaces con bajos tamaños muestrales serían de suma utilidad para la comunidad científica dedicada al estudio de especies que, o bien sus poblaciones se encuentran en peligro, o bien son difíciles de capturar. Con este fin, para cada una de las metodologías de clasificación propuestas, se manejaron tres tamaños muestrales diferentes por río para entrenar los modelos y evaluar su precisión: $n_1 = 10$, $n_2 = 30$, $n_3 = 60$. Si bien se tratan de tamaños muestrales relativamente bajos, muy rara vez se podrá contar con mayores tamaños muestrales de estas especies provenientes de un mismo lugar, por lo que son adecuados para el estudio.

2.5.3. Ensamblado de modelos

El proceso de ensamblado de modelos o *model ensembling* consiste en la combinación de las predicciones de distintos modelos con el objetivo de mejorar las predicciones finales (Amat 2018). Esta estrategia se basa en la asunción de que, distintos modelos entrenados independientemente, emplean distintos aspectos de los datos para realizar las predicciones, es decir, cada uno es capaz de identificar parte de la “verdad” pero no toda ella. Combinando la perspectiva de cada uno de ellos, se obtiene una descripción más detallada de la verdadera estructura subyacente en los datos (Amat 2018).

A colación de lo expuesto, se procedió a combinar las predicciones de los diferentes tipos de modelos contemplados en el estudio con el fin de comprobar la existencia de una mejoría en la calidad predictiva. Debido a que los modelos ajustados son de clasificación, la combinación de las predicciones se llevó a cabo a través de la categoría modal.

2.6. Clasificación de los ejemplares adultos

Tras llevar a cabo el ajuste y validación de los modelos de clasificación con los que asignar el río natal de los ejemplares de alosa adultos, se escogieron los modelos que presentaron los mejores resultados en relación con la calidad predictiva con el fin de llevar a cabo la clasificación de los ejemplares adultos capturados en río para el análisis de la filopatria y, por otro lado, la clasificación de los ejemplares capturados en el mar para el estudio de capacidad de dispersión en el medio marino.

A mayores, una vez clasificados los ejemplares adultos en sus respectivos ríos de origen, se calcularon las probabilidades marginales de asignación de cada uno de los individuos a los diferentes ríos. A partir de estos datos se cuantificó el número de individuos cuya asignación a un río específico superaba un nivel mínimo de probabilidades preestablecido. Se establecieron 3 umbrales de probabilidad mínima de asignación, siendo estos del 80%, 65% y 50%. Aquellos ejemplares que no alcanzasen dichos umbrales fueron clasificados como de origen desconocido.

Capítulo 3

Resultados

En el siguiente capítulo se recogen los principales resultados obtenidos tras el ajuste de los diferentes modelos de clasificación y posterior clasificación de los ejemplares de alosa adultos en sus respectivos ríos de origen. En líneas generales, el capítulo se divide en dos grandes apartados, estando el primero de estos centrado en la comparación y selección de los modelos de clasificación ajustados, mientras que en el segundo abordamos la asignación a un río de origen determinado de los ejemplares adultos de *A. alosa* y *A. fallax* capturados tanto en el medio marino, para el estudio la capacidad dispersiva, como en el medio fluvial, para el estudio de la filopatria o *homing*.

3.1. Evaluación y comparación de los modelos

En el presente apartado se sintetizan los resultados obtenidos del rendimiento alcanzado por los diferentes modelos de clasificación entrenados con diversos tamaños muestrales. Dicho rendimiento se evalúa en función de la precisión obtenida. Este parámetro aparece reflejado en una serie de gráficos en los que encontramos la precisión media de cada modelo, la dispersión de la precisión obtenida en varias realizaciones con diferentes submuestras y la comparación entre el rendimiento de los modelos al ser evaluados a través de la muestra de entrenamiento o la de test. Asimismo, las diferencias entre las precisiones de los modelos son sometidas a la prueba de

prueba de Mann-Whitney-Wilcoxon para muestras pareadas para esclarecer la existencia de diferencias significativas en el rendimiento de los modelos.

Finalmente, tras realizar la evaluación y comparación de los modelos, existe un subapartado final en el que se determina el modelo de clasificación con el que se llevará a cabo la asignación del río de origen de las alosas adultas.

3.1.1. Estudio de la filopatria

- $n_1 = 10$

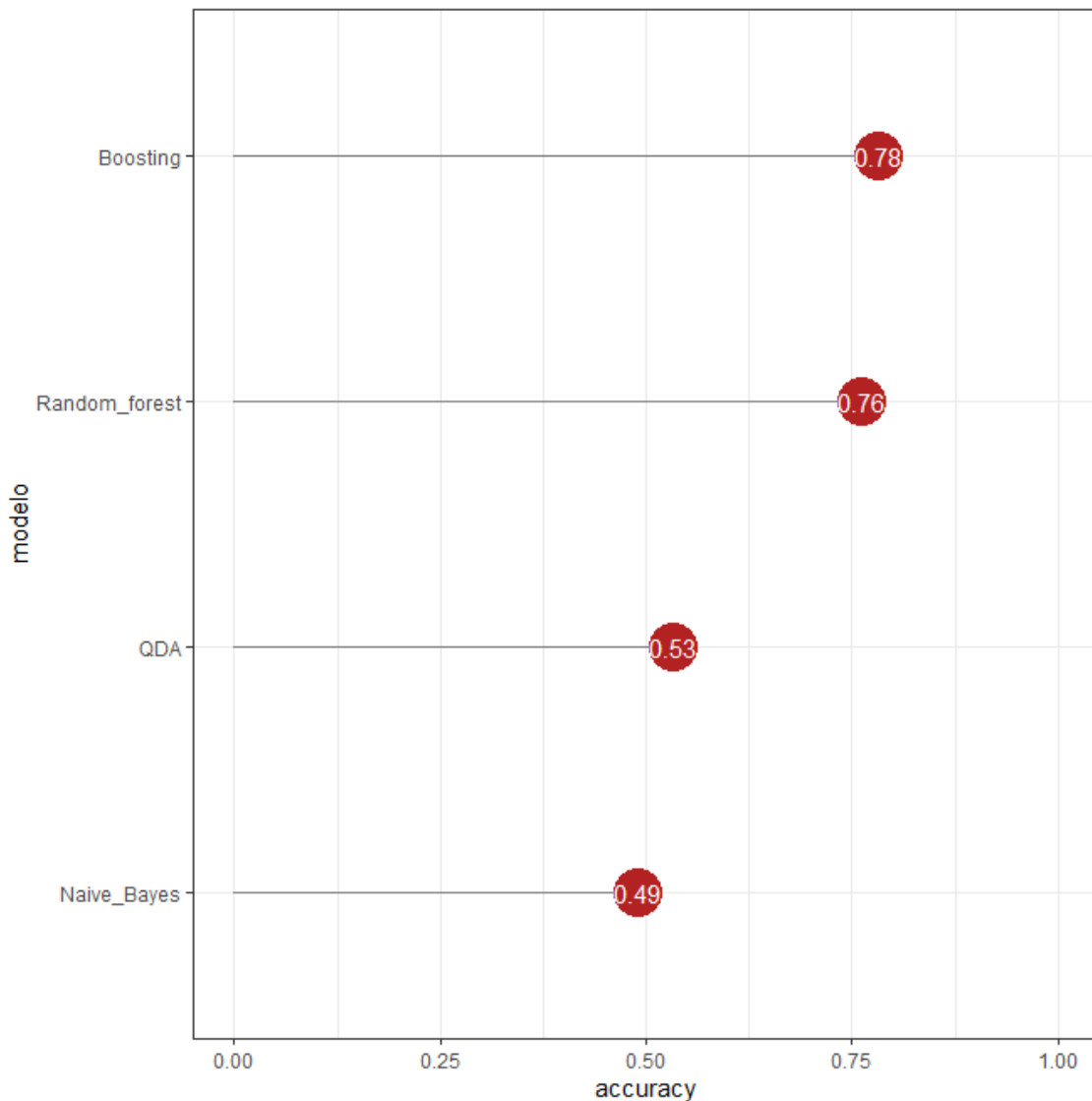


Figura 7: Precisión o *accuracy* medio de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_1 = 10$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.

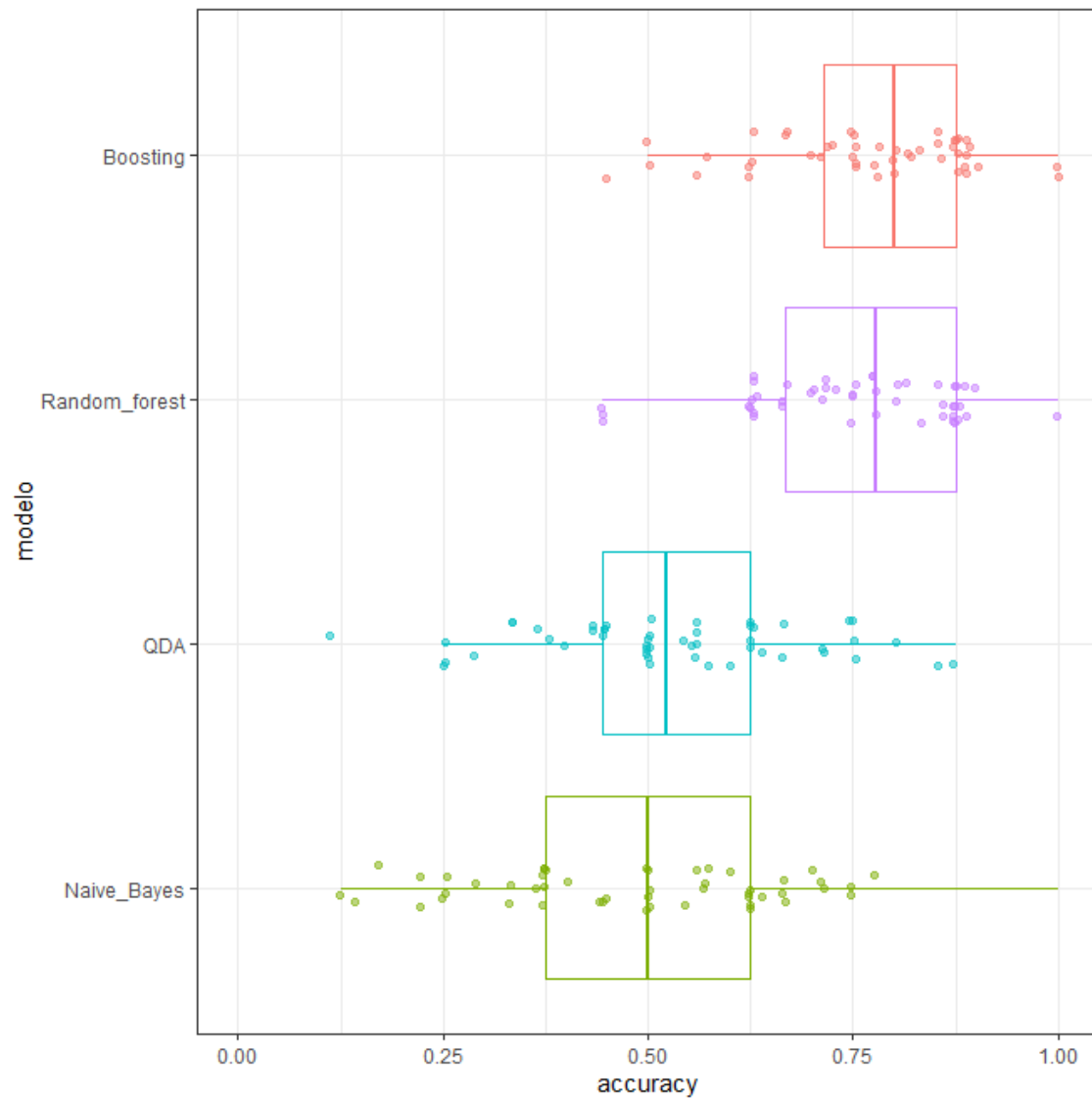


Figura 8: Precisión o *accuracy* de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_1 = 10$). Cada punto es una realización del proceso de validación cruzada.

Tal y como podemos observar en la figura 7, los modelos basados en árboles de decisión alcanzan un poder predictivo superior al alcanzado por los modelos generadores. De entre los primeros, el *boosting* presenta mejores resultados seguido muy de cerca del *random forest*.

Además, si nos fijamos en la figura 8 observaremos que la dispersión de los valores obtenidos del modelo *boosting* es menor que la del resto de modelos, siendo el *Naive Bayes*, el modelo en el que se observa una mayor dispersión de los resultados.

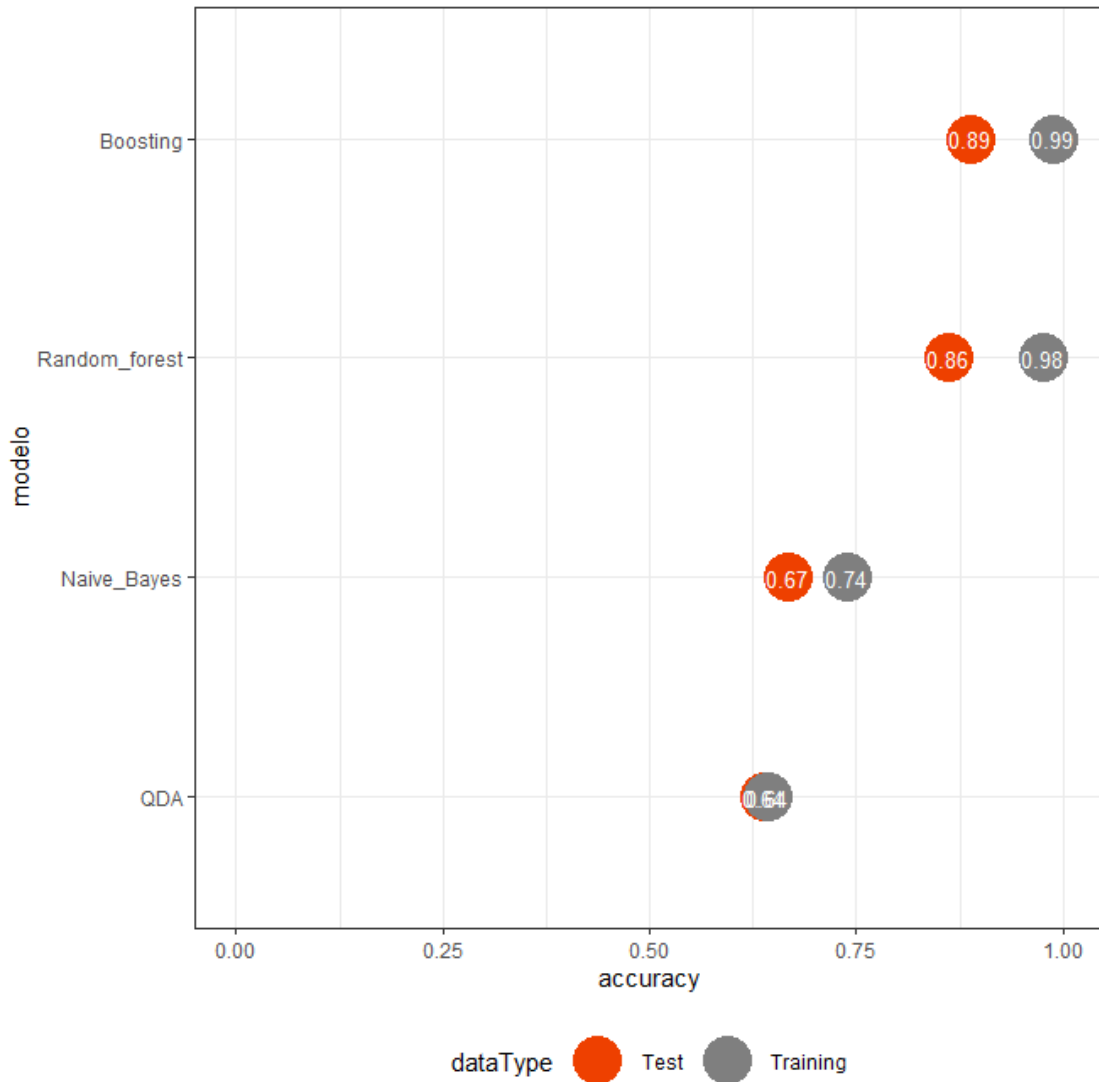


Figura 9: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la filopatria ($n_1 = 10$).

En la figura 9 se puede observar cómo los modelos *boosting*, *random forest* y *Naive Bayes* presentan un mejor comportamiento ante la muestra de entrenamiento respecto a la de test. Esto es una consecuencia lógica al estar comparando observaciones que el modelo ya ha “visto” anteriormente con observaciones nuevas para el mismo.

Tabla 1: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la filopatria ($n_1 = 10$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.

Modelo A	Modelo B	p-valor
Naive Bayes	Boosting	6.80e-08
QDA	Boosting	1.37e-07
QDA	Naive Bayes	9.01e-08
Random forest	Boosting	5.36e-01
Random forest	Naive Bayes	1.37e-07
Random forest	QDA	1.78e-07

Existen diferencias significativas entre todos los modelos al comparar su precisión (tabla 1) con la excepción del *random forest* y el *boosting*. De esta forma, los modelos basados en árboles de decisión son significativamente superiores a los generadores en este escenario.

Al ensamblar los modelos QDA y *Naive Bayes*, se consiguió una precisión promedio del 67%, mejorando los resultados obtenidos de forma individual por cada uno de los métodos generadores, pero no alcanzando los valores de precisión del *random forest* o el *boosting*. Asimismo, al combinar los cuatro modelos de clasificación, se aumentó la precisión media hasta alcanzar el valor del 86%.

▪ $n_2 = 30$

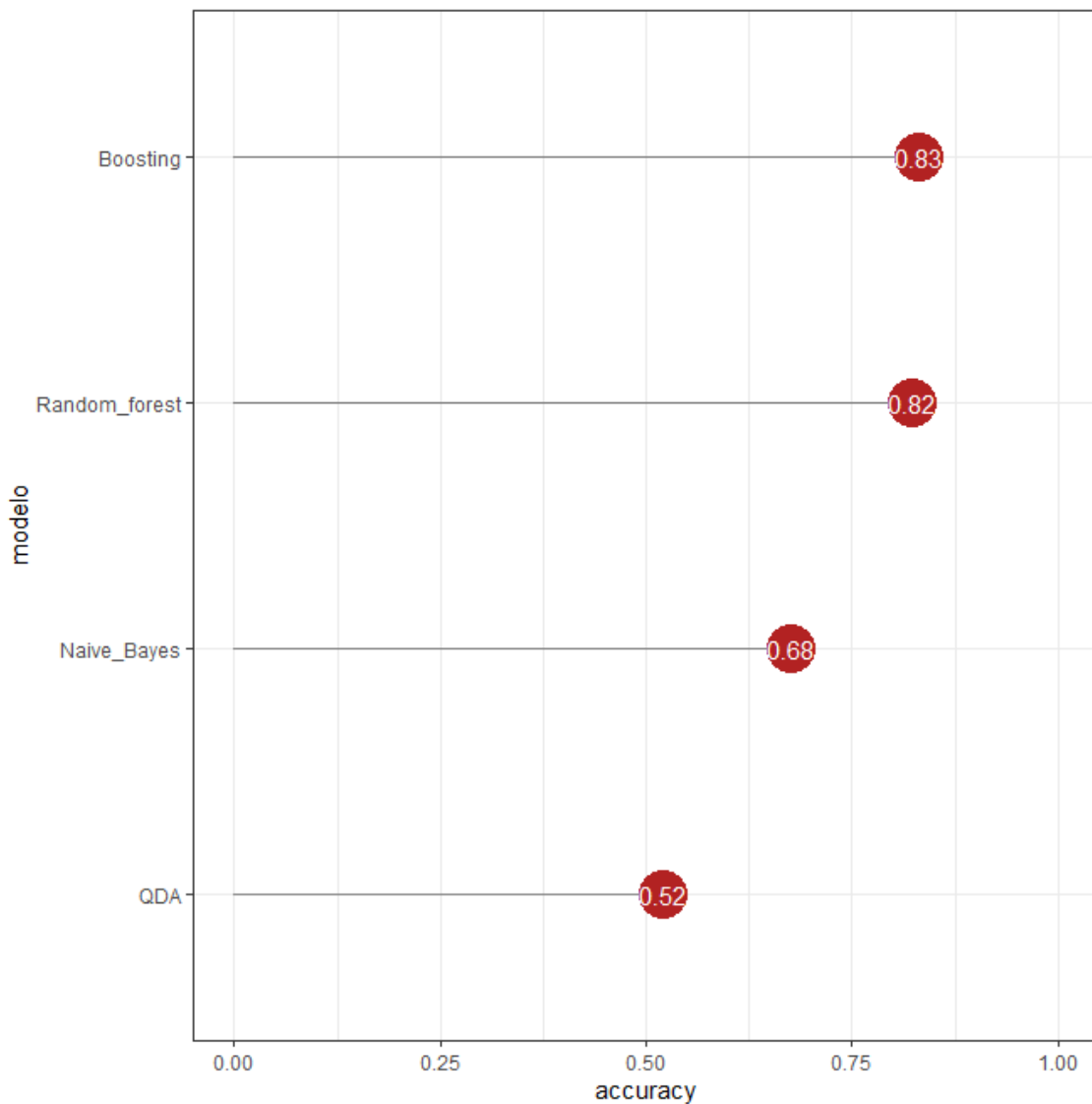


Figura 10: Precisión o *accuracy* medio de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_2 = 30$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.

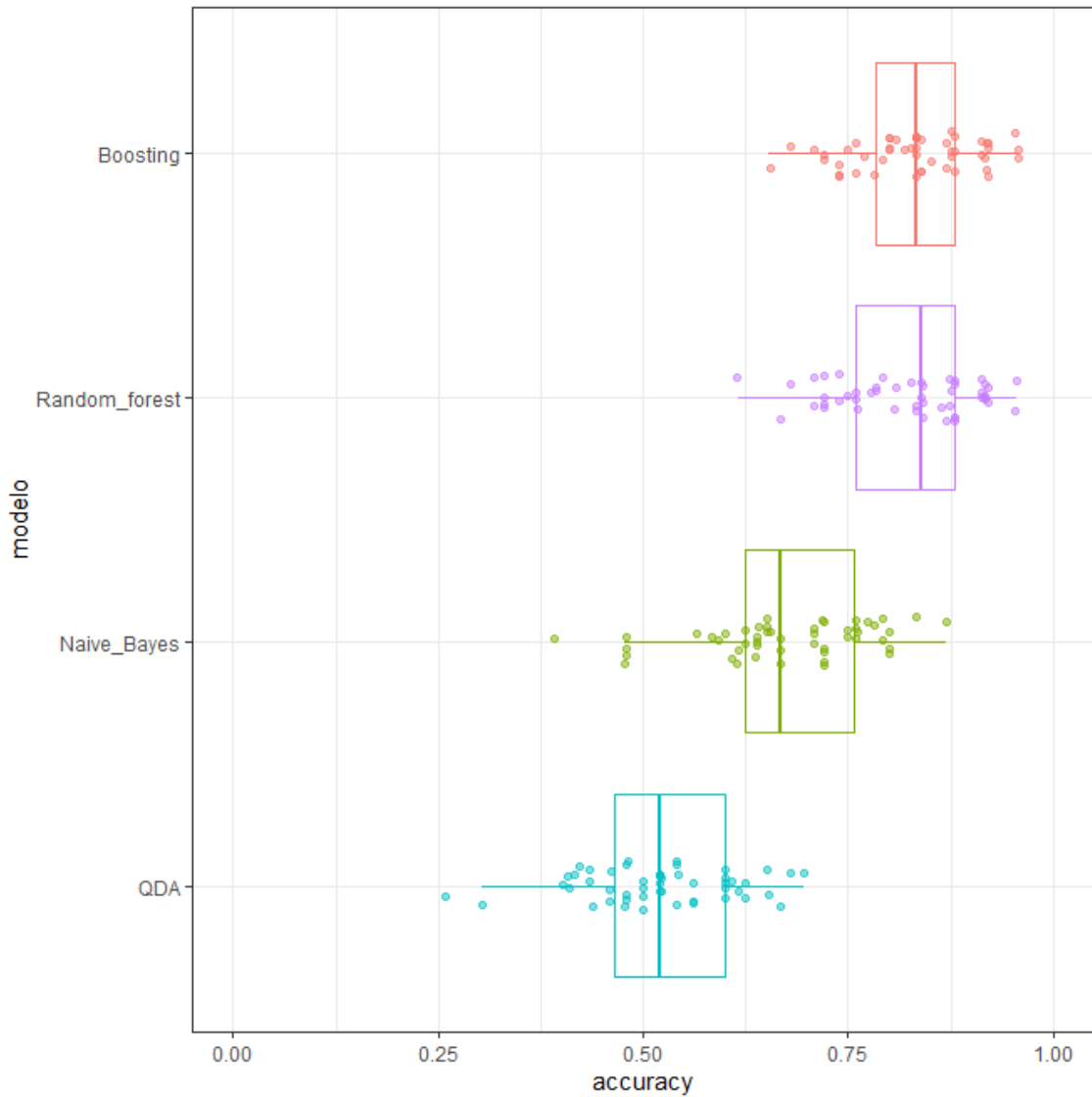


Figura 11: Precisión o *accuracy* de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_2 = 30$). Cada punto es una realización del proceso de validación cruzada.

Tal y como podemos observar en las figura 10 al aumentar el tamaño muestral de 10 a 30 ejemplares por río, la precisión media alcanzada en general aumenta. El modelo donde se registra un mayor aumento de la precisión media es el *Naive Bayes*, en contraste, el modelo QDA incluso llega a perder un 1% de calidad predictiva. No obstante, la relación de eficacia en la predicción entre modelos basado en árboles de decisión y modelos generadores sigue la misma pauta que la observada con el tamaño muestral menor.

En relación con la dispersión de las observaciones (figura 11), se observa una menor variabilidad en todos los modelos al compararlos con los ajustados a un tamaño muestral menor.

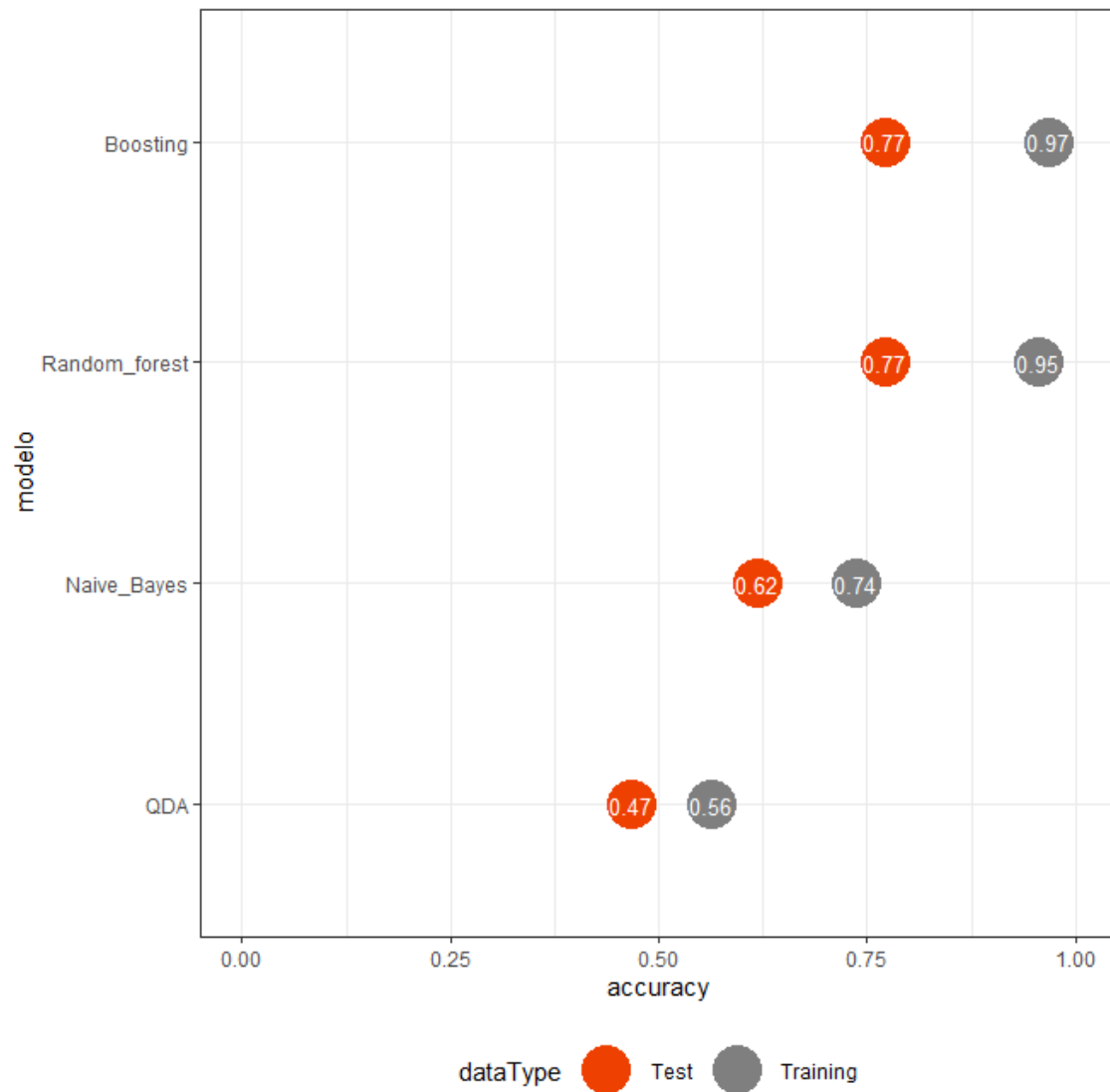


Figura 12: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la filopatria ($n_2 = 30$).

En la figura 12 podemos contemplar cómo, al igual que ocurría antes, los modelos presentan un mejor comportamiento predictivo al emplear la muestra de entrenamiento respecto a la de test. En este caso, el modelo QDA también presenta esta tendencia.

Tabla 2: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la filopatria ($n_2 = 30$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.

Modelo A	Modelo B	p-valor
Naive Bayes	Boosting	1.05e-08
QDA	Boosting	4.63e-09
QDA	Naive Bayes	3.31e-07
Random forest	Boosting	3.59e-01
Random forest	Naive Bayes	2.88e-08
Random forest	QDA	5.70e-09

Como en el caso anterior, existen diferencias significativas entre todos los modelos al comparar su precisión (tabla 2) con excepción del *random forest* y el *boosting*. De esta forma, los modelos basados en árboles de decisión son significativamente superiores a los generadores en este escenario.

Al combinar mediante el ensamblaje de modelos, los modelos QDA y *Naive Bayes*, se consiguió una precisión promedio del 61%, siendo este resultado peor que el alcanzado de forma individual por el modelo *Naive Bayes*. Asimismo, al combinar los cuatro modelos de clasificación, se alcanzó una precisión del 75%, siendo este un resultado peor que los obtenidos de forma individual por el *boosting* y *random forest*.

▪ $n_3 = 60$

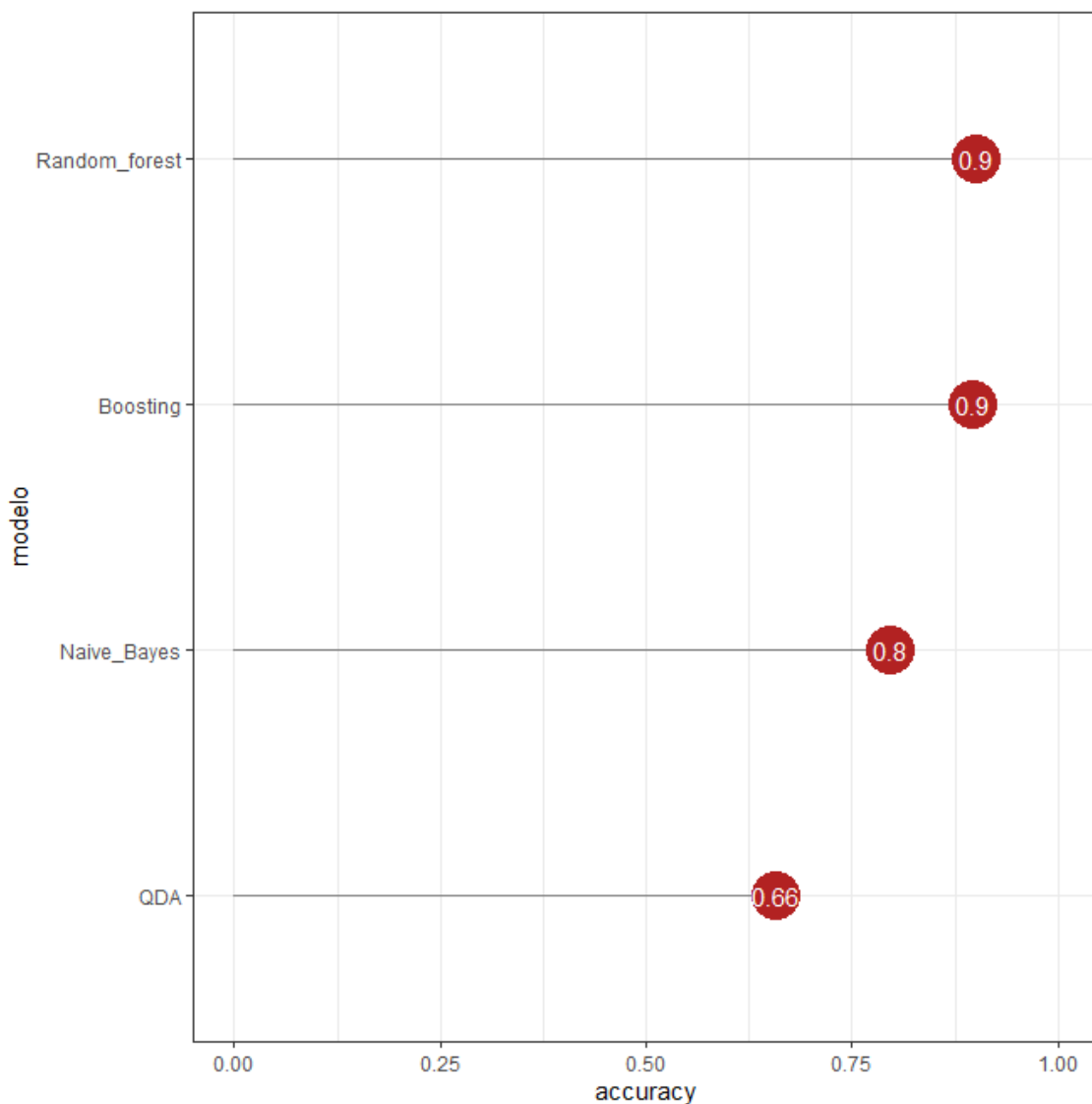


Figura 13: Precisión o *accuracy* medio de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_3 = 60$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.

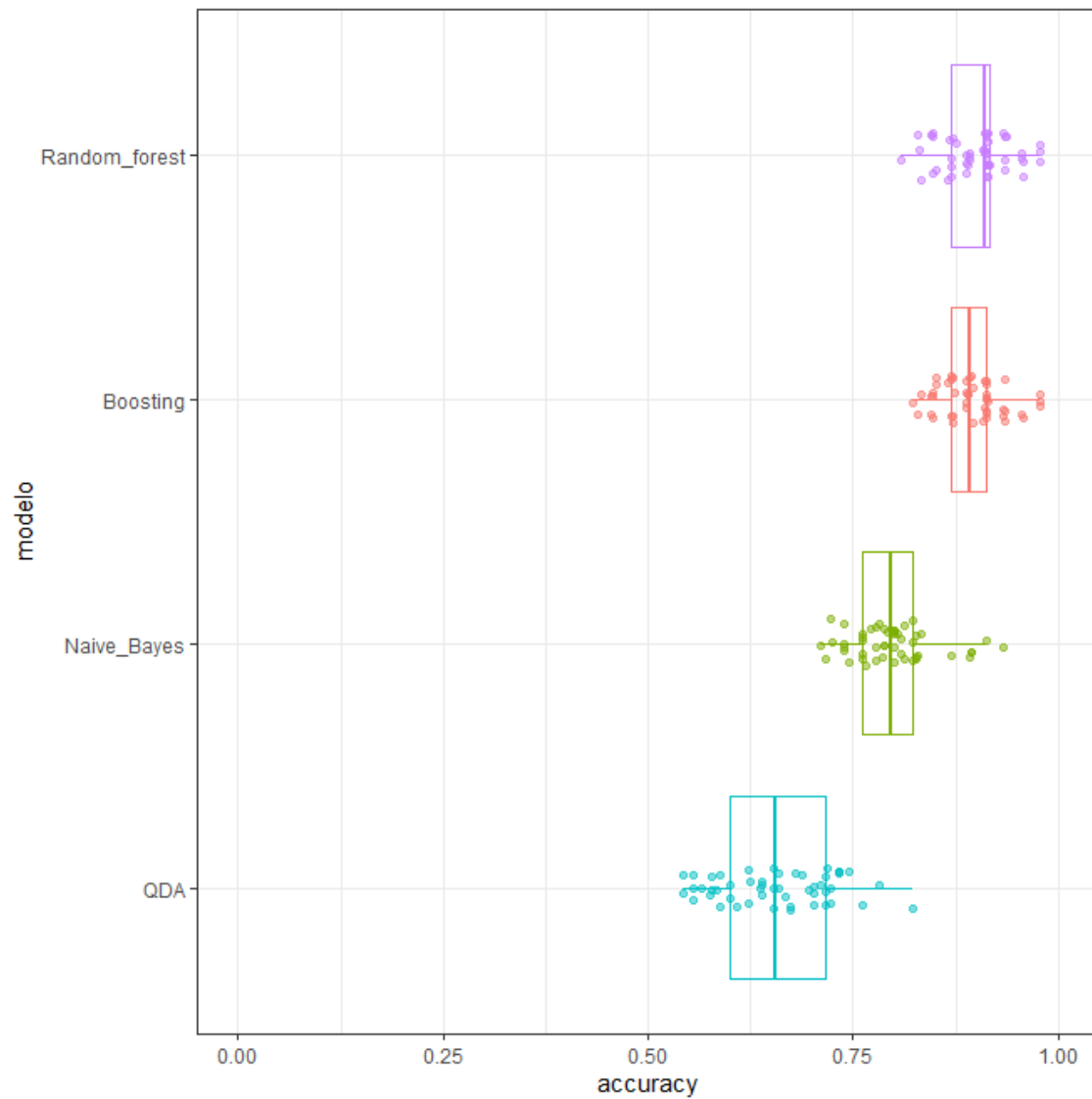


Figura 14: Precisión o *accuracy* de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_3 = 60$). Cada punto es una realización del proceso de validación cruzada.

En las figuras 13 y 14 podemos ver como se repite la tendencia a aumentar de la precisión media y disminuir la dispersión de las observaciones conforme se aumenta el tamaño muestral. En este caso, El modelo *boosting* y *random forest* alcanzan la misma precisión media. Además, al contrario de lo que sucedía en el caso anterior, la precisión media del modelo QDA también se ve aumentada pese a continuar siendo el modelo con peor calidad predictiva y que mayor dispersión de los datos presenta.

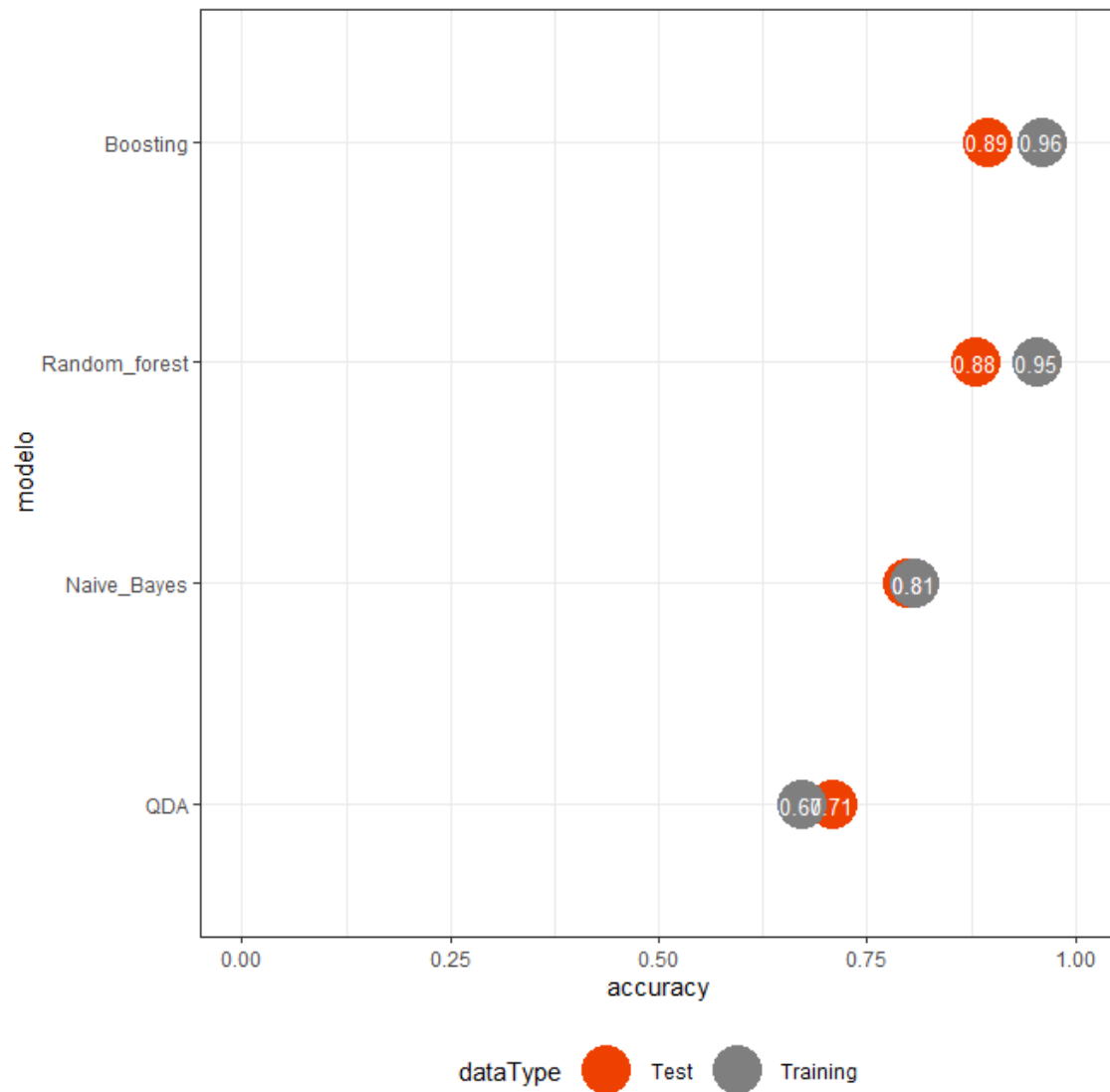


Figura 15: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la filopatria ($n_3 = 60$).

Al aumentar el tamaño muestral a 60 juveniles por río, se puede apreciar cómo, si bien el comportamiento general de los modelos sigue siendo mejor ante la muestra de entrenamiento, los resultados tienden a acercarse en mayor medida a los obtenidos con la muestra de test (figura 15). Cabe destacar que el modelo que peor calidad predictiva presenta también es el único que parece mostrar un mejor comportamiento ante la muestra de test respecto a la de entrenamiento.

Tabla 3: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la filopatria ($n_3 = 60$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.

Modelo A	Modelo B	p-valor
Naive Bayes	Boosting	7.22e-09
QDA	Boosting	4.62e-09
QDA	Naive Bayes	6.42e-09
Random forest	Boosting	3.08e-01
Random forest	Naive Bayes	6.42e-09
Random forest	QDA	4.62e-09

Como en el caso anterior, exceptuando la diferencia entre el *random forest* y el *boosting*, existen diferencias significativas entre todos los modelos al comparar su precisión (tabla 3). De esta forma, los modelos basados en árboles de decisión son significativamente superiores a los generadores en este escenario y en los anteriores. Por otro lado, si bien los modelos *boosting* y *random forest* presentaron una precisión media del del 90%, el modelo *boosting* parece tener una calidad predictiva ligeramente superior al emplear la muestra de test.

Al ensamblar los modelos QDA y *Naive Bayes*, se consiguió una precisión promedio del 79%, siendo este resultado ligeramente inferior al alcanzado de forma individual por el modelo *Naive Bayes*. Asimismo, al combinar los cuatro modelos de clasificación la precisión media alcanzó el valor aproximado de 90%, no mejorando este resultado al obtenido por parte de los modelos basados en árboles de decisión.

3.1.2. Estudio de la dispersión en el medio marino

- $n_1 = 10$

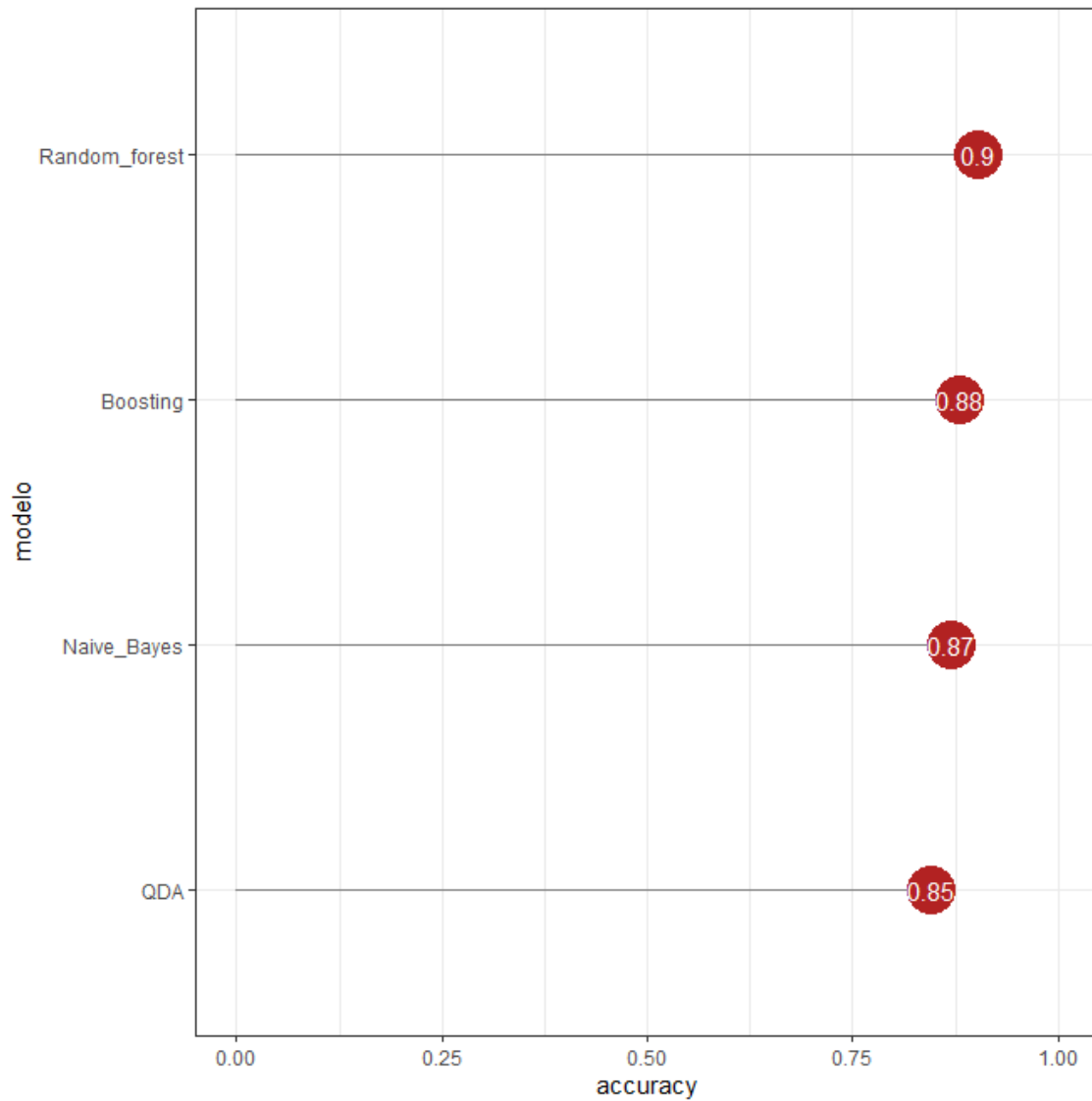


Figura 16: Precisión o *accuracy* medio de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_1 = 10$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.

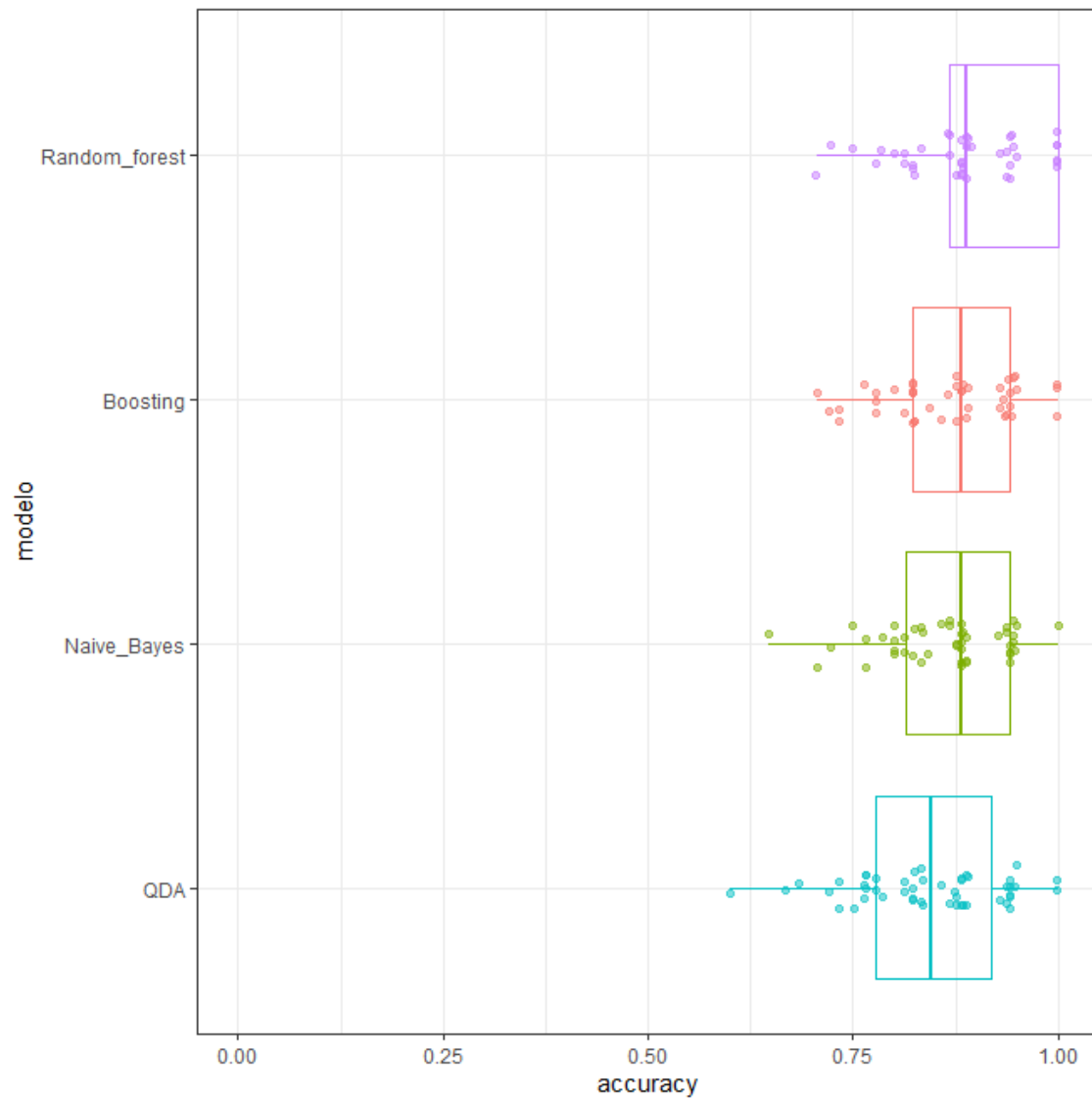


Figura 17: Precisión o *accuracy* de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_1 = 10$). Cada punto es una realización del proceso de validación cruzada.

Al igual que ocurre en los modelos para el estudio de la filopatria, podemos observar en la figura 16 que los modelos basados en árboles de decisión alcanzan un poder predictivo superior al alcanzado por los modelos generadores. De entre los primeros, el *boosting* presenta mejores resultados seguido muy de cerca del *random forest*. No obstante, el resto de los modelos muestran un rendimiento aceptable y próximo al *boosting*.

Por otro lado, la dispersión de las observaciones generadas por validación cruzada es parecida entre los diferentes modelos (figura 17).

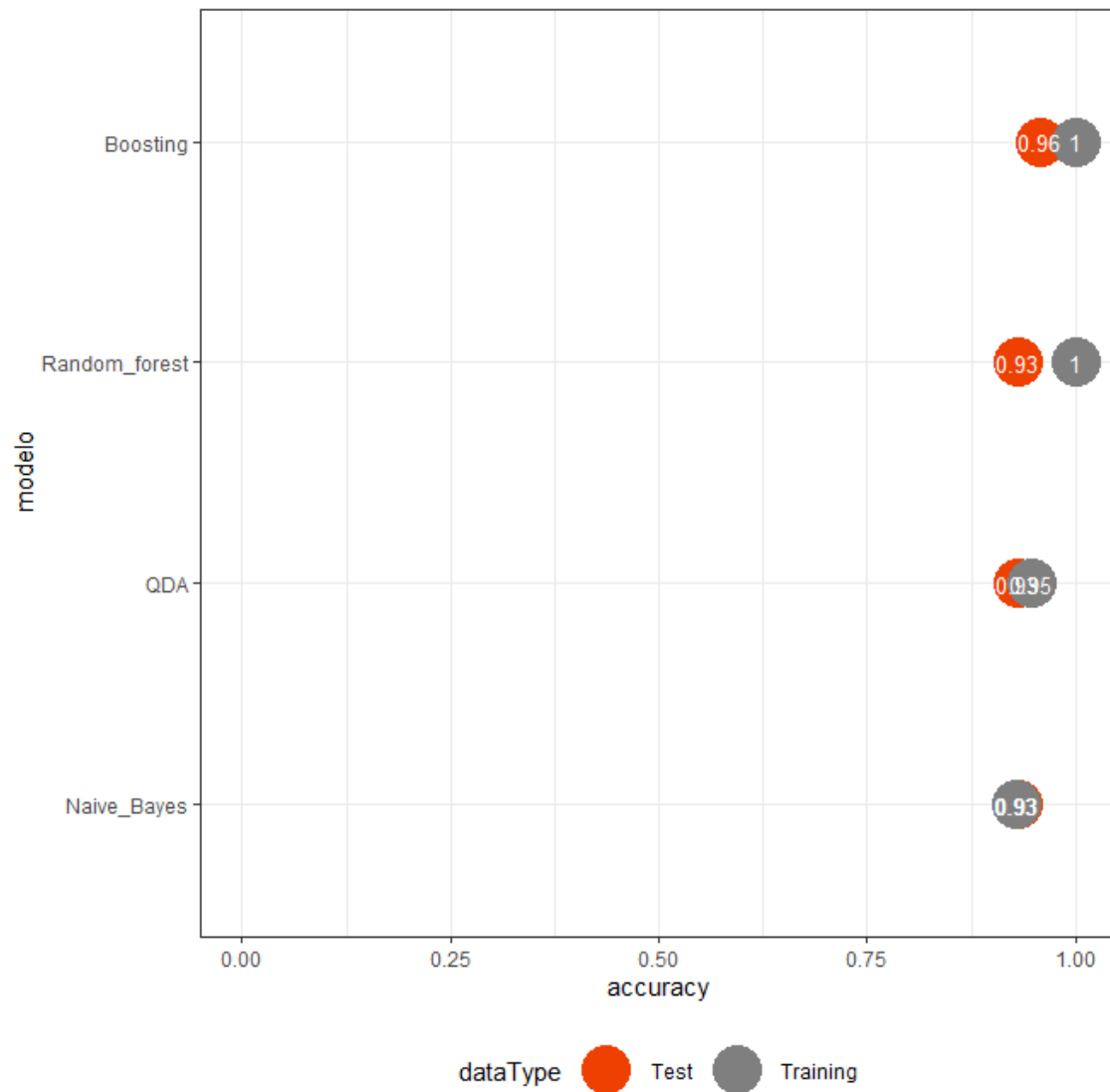


Figura 18: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la dispersión ($n_1 = 10$).

En la figura 18 podemos contemplar cómo, al igual que ocurría en el caso de la filopatria, los modelos, en general, presentan un mejor comportamiento predictivo al emplear la muestra de entrenamiento respecto a la de test. Sin embargo, la precisión obtenida empleando ambas muestras es muy próxima entre sí pese a emplear un tamaño muestral reducido. Este hecho se acentúa en el caso de los métodos generadores que prácticamente tienen la misma calidad predictiva con independencia de la partición empleada para la predicción.

Tabla 4: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la dispersión ($n_1 = 10$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.

Modelo A	Modelo B	p-valor
Naive Bayes	Boosting	3.49e-01
QDA	Boosting	7.75e-02
QDA	Naive Bayes	1.58e-01
Random forest	Boosting	1.03e-01
Random forest	Naive Bayes	1.20e-01
Random forest	QDA	7.06e-04

Si nos fijamos en la tabla 4 podemos comprobar que tan solo existen diferencias significativas en la calidad predictiva entre el modelo de *random forest* y los modelos generados QDA y *Naive Bayes*.

Al combinar mediante *model ensambling* los modelos QDA y *Naive Bayes*, se consiguió una precisión promedio del 93%, siendo este resultado superior al alcanzado de forma individual por los métodos basados en árboles de decisión. Asimismo, al combinar los cuatro modelos de clasificación la precisión media aumentó hasta el 94%, siendo este resultado superior al alcanzado de forma individual por cada uno de los modelos y muy similar al conseguido al combinar los modelos generadores.

- $n_2 = 30$

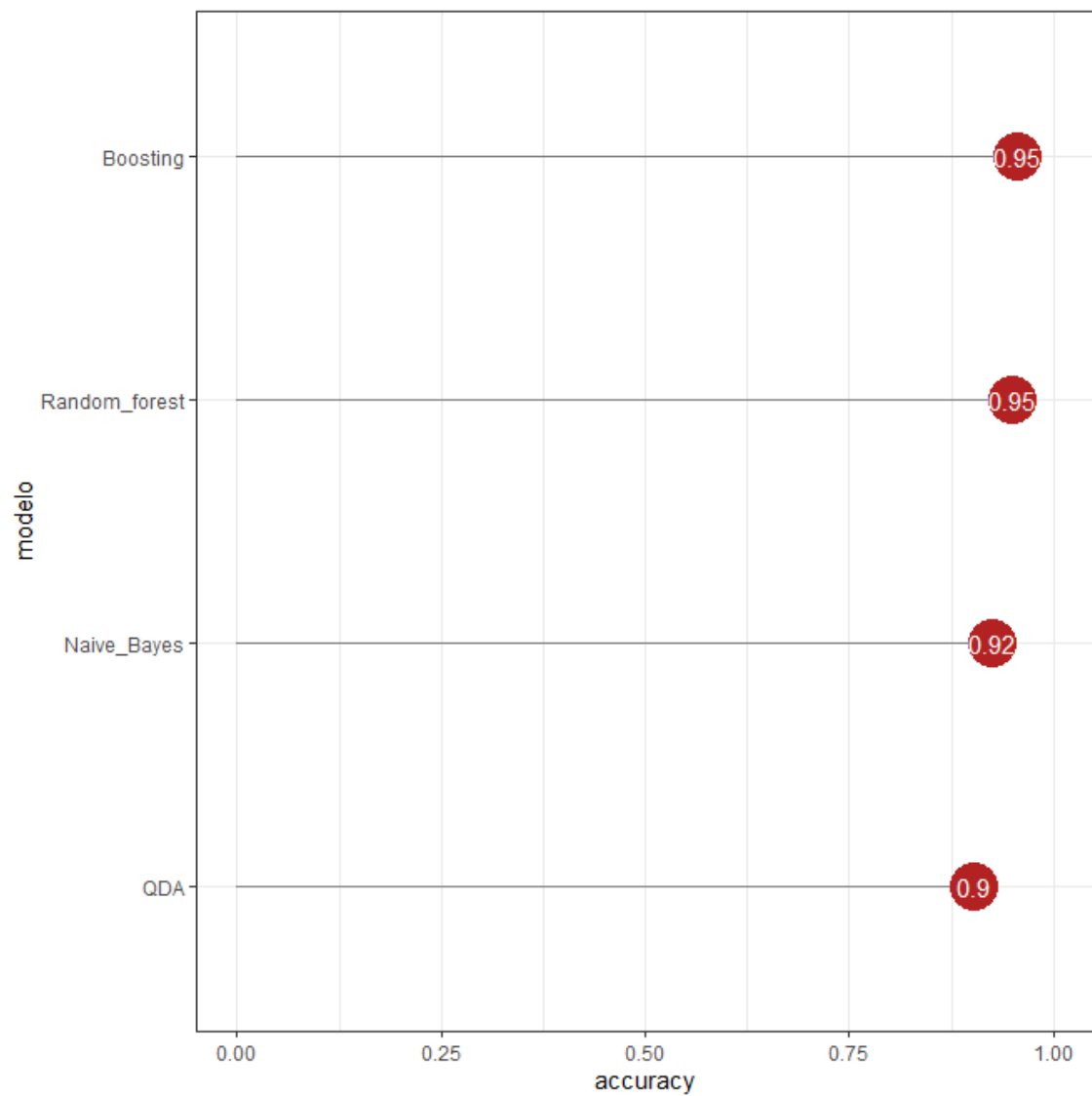


Figura 19: Precisión o *accuracy* medio de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_2 = 30$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.

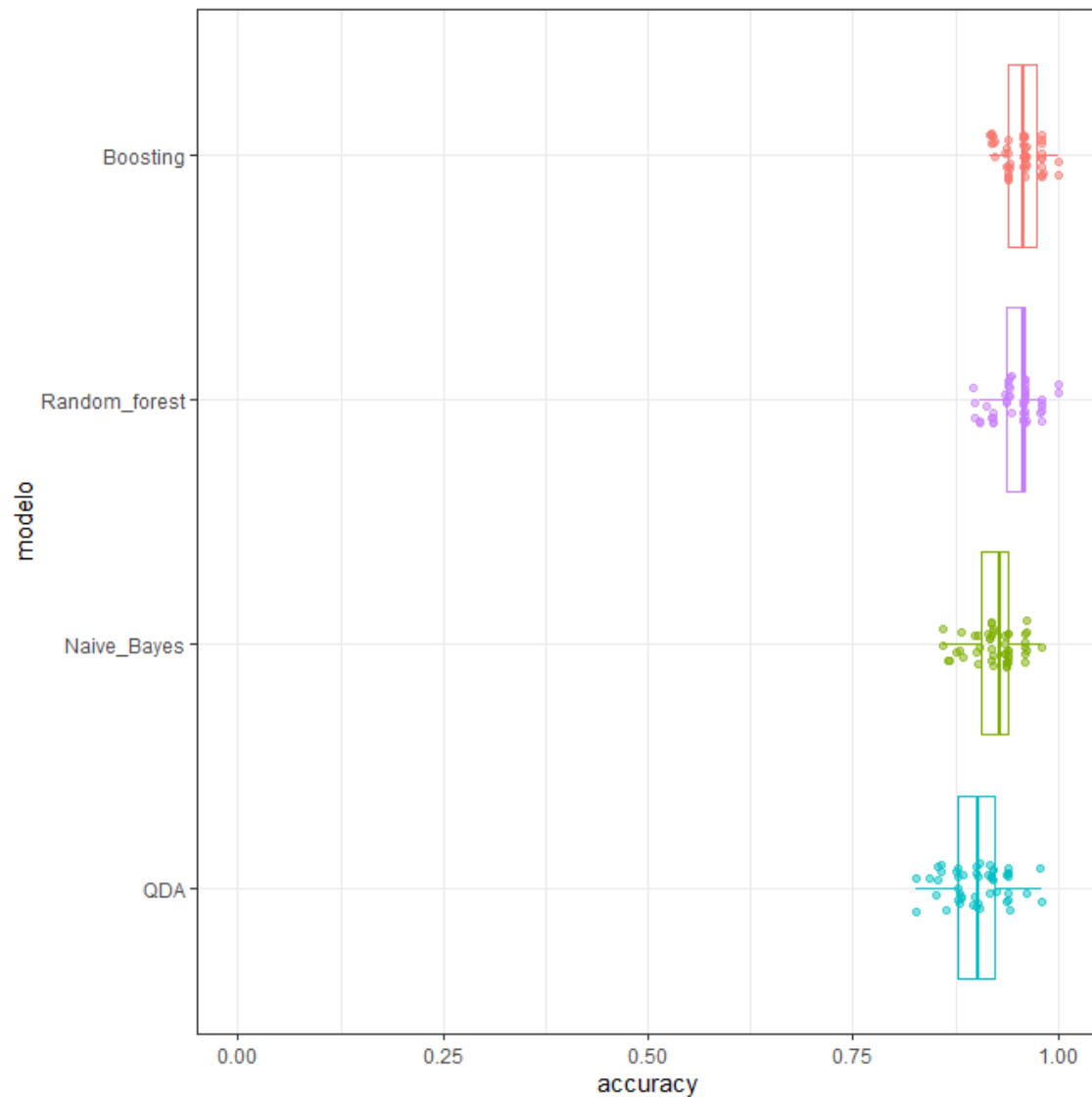


Figura 20: Precisión o *accuracy* de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_2 = 30$). Cada punto es una realización del proceso de validación cruzada.

Tal y como podemos observar en las figura 19 al aumentar el tamaño muestral de 10 a 30 ejemplares por río, la precisión media alcanzada en general aumenta hasta valores mínimos del 90%. No obstante, la relación de eficacia en la predicción entre modelos basado en árboles de decisión y modelos generadores sigue la misma pauta que la observada con el tamaño muestral menor.

En relación con la dispersión de las observaciones (figura 20), se sigue la tendencia registrada en los modelos del estudio de la filopatria, a mayor tamaño muestral menor variabilidad en los resultados.

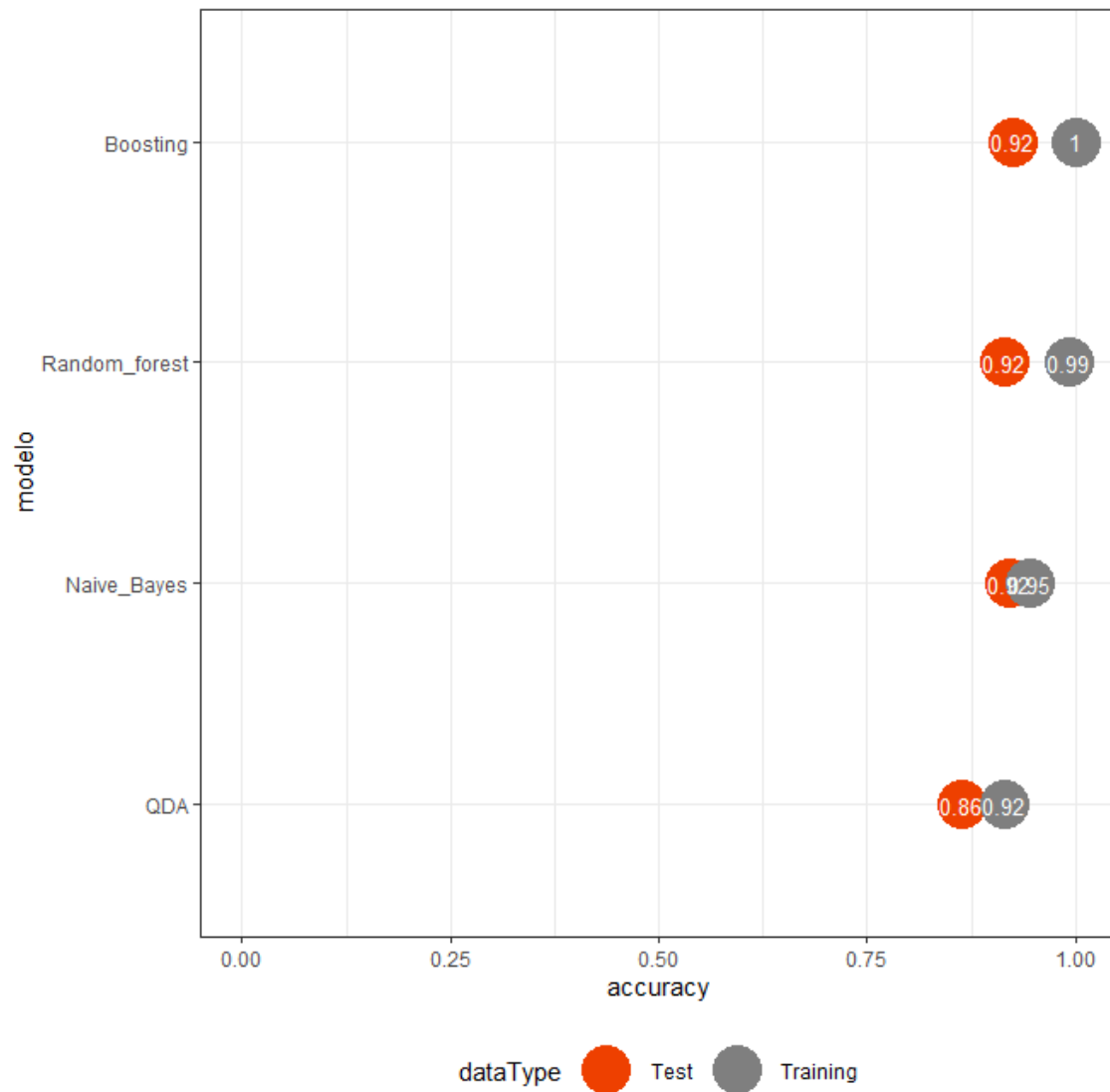


Figura 21: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la dispersión ($n_2 = 30$).

La calidad predictiva de los modelos continúa siendo mayor al emplear la muestra de entrenamiento en comparación con la test (figura 21). Sin embargo, con independencia de la muestra empleada, todos los modelos muestran una calidad predictiva aceptable.

Tabla 5: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la dispersión ($n_2 = 30$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.

Modelo A	Modelo B	p-valor
Naive Bayes	Boosting	1.85e-06
QDA	Boosting	3.61e-08
QDA	Naive Bayes	1.55e-03
Random forest	Boosting	1.69e-01
Random forest	Naive Bayes	1.85e-06
Random forest	QDA	3.98e-08

En la tabla 5 se aprecia como, en contraste de lo ocurrido con el primer tamaño muestral, en este caso todos los modelos muestran diferencias significativas entre sí en lo que respecta a la precisión alcanzada con la única excepción encontrada al comprar el *random forest* con el *boosting*, donde no hay diferencias significativas.

En este escenario, al combinar los modelos generadores QDA y *Naive Bayes*, se alcanza la misma precisión que al combinar los cuatro modelos de clasificación, siendo este valor del 92%. Así pues, no se consigue una mejoría notoria al combinar los distintos modelos de clasificación.

▪ $n_3 = 60$

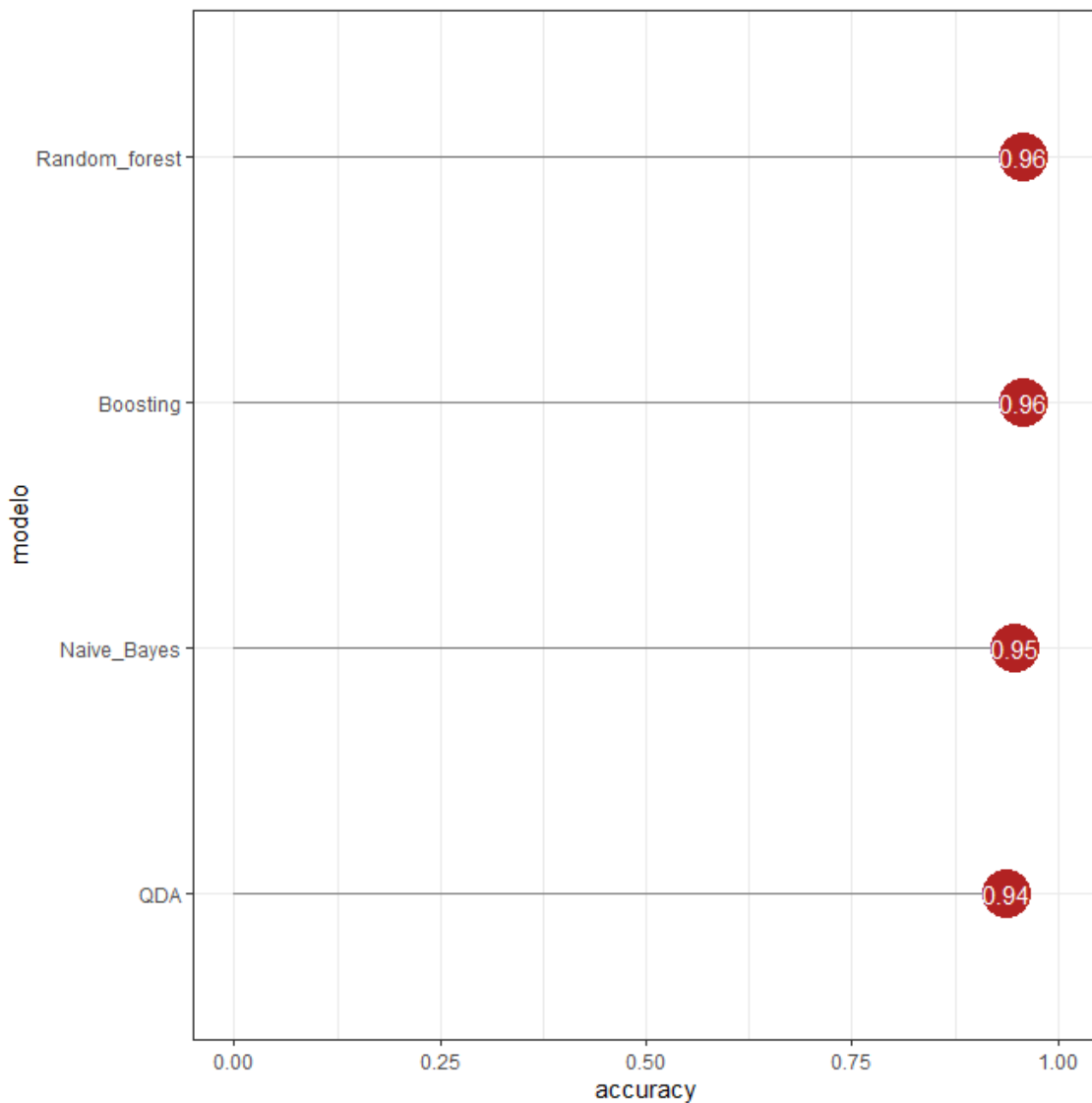


Figura 22: Precisión o *accuracy* medio de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_3 = 60$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.

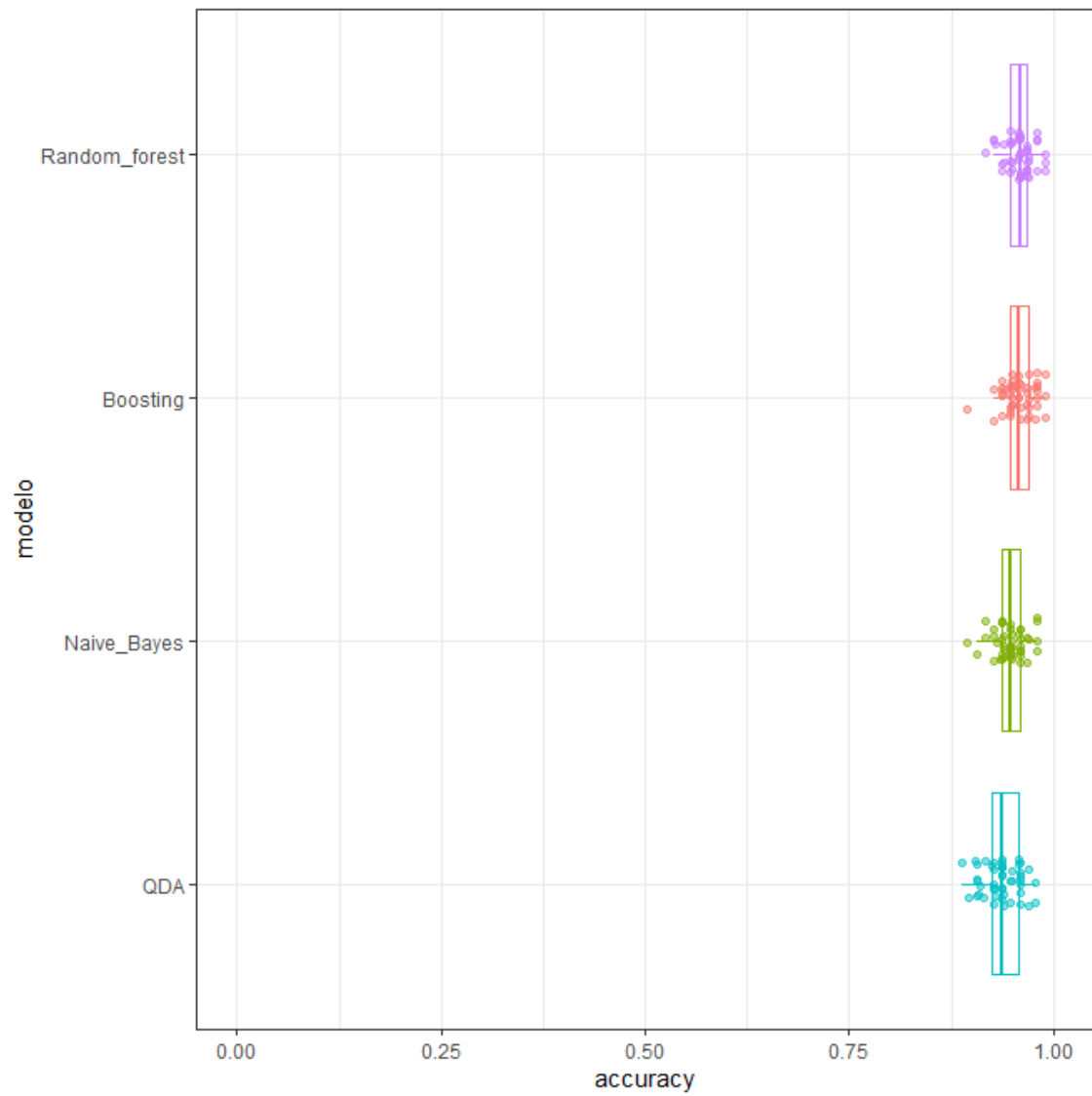


Figura 23: Precisión o *accuracy* de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_3 = 60$). Cada punto es una realización del proceso de validación cruzada.

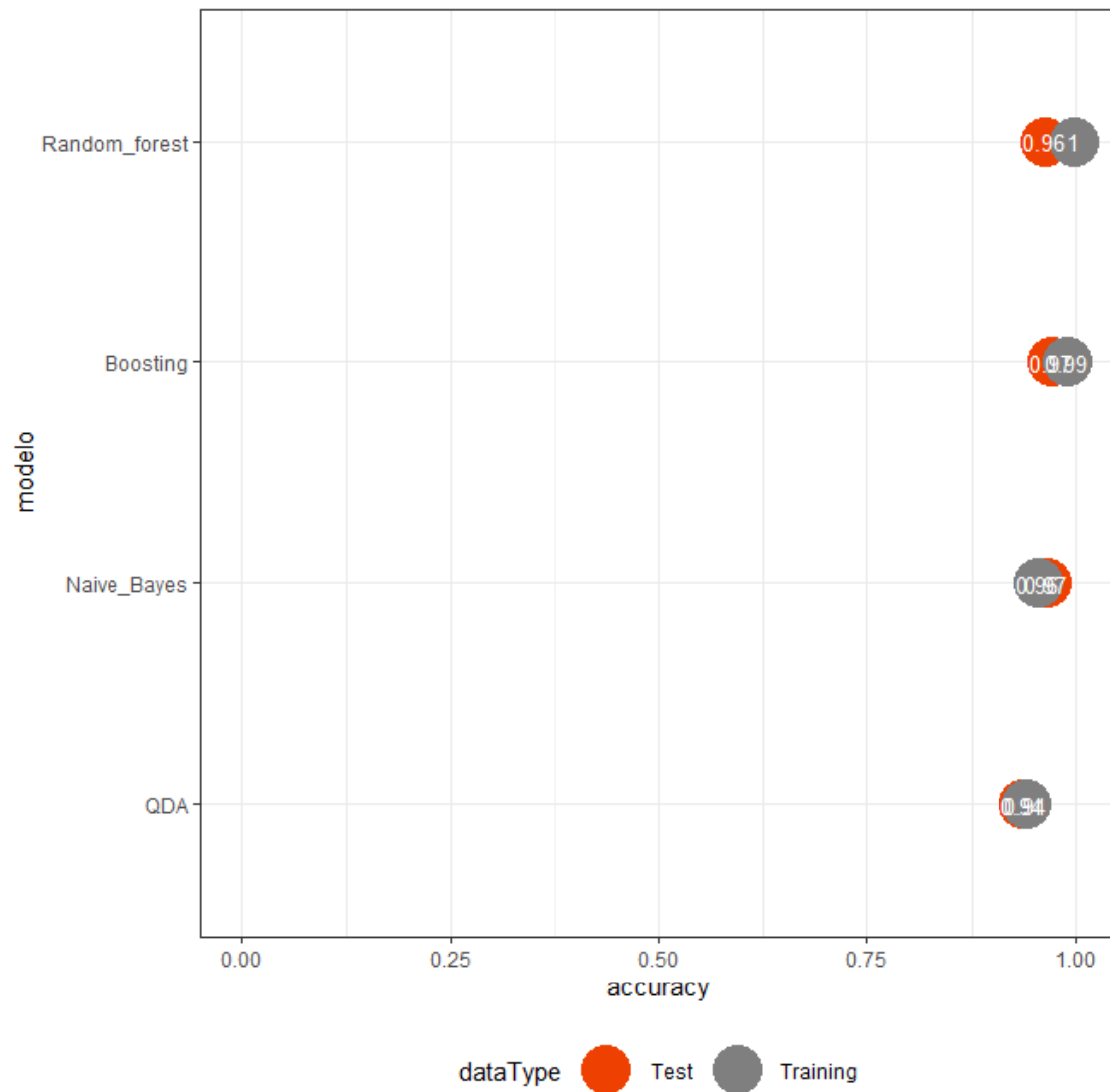


Figura 24: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la dispersión ($n_3 = 60$).

Tabla 6: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la dispersión ($n_3 = 60$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.

Modelo A	Modelo B	p-valor
Naive Bayes	Boosting	1.10e-03
QDA	Boosting	1.48e-05
QDA	Naive Bayes	8.87e-03
Random forest	Boosting	6.20e-01
Random forest	Naive Bayes	3.21e-04
Random forest	QDA	5.29e-06

Al aumentar el tamaño muestral de al doble respecto al caso anterior, las tendencias observadas a lo largo de este apartado continúan. La precisión media aumenta ligeramente en todos los modelos (figura 22), así como el descenso en la dispersión de los resultados obtenidos por validación cruzada (figura 23). Además, la calidad predictiva entre las muestras de entrenamiento y test son muy similares (figura 24). A mayores, y al igual que ocurría con el tamaño muestral de 30 ejemplares, existen diferencias significativas entre las diferentes modelos de clasificación excepto entre los modelos basados en árboles de decisión (tabla 6).

Al combinar mediante model ensambling los modelos QDA y *Naive Bayes*, se consiguió una precisión promedio del 96%, siendo este resultado muy similar al alcanzado de forma individual por los métodos basados en árboles de decisión. Asimismo, al combinar los cuatro modelos de clasificación la precisión media aumentó hasta el 97%, un resultado muy similar también al alcanzado por los modelos basados en árboles de decisión.

3.1.3. Selección de modelos y ejemplares para la clasificación

A la luz de los resultados obtenidos, se decidió emplear un modelo *boosting* con un tamaño muestral por río de 60 ejemplares para la clasificación de las alosas adultas capturadas tanto en el medio fluvial como marino para el análisis del grado de filopatria y de la capacidad dispersiva durante la fase marina respectivamente.

El modelo *boosting* presentó en la mayoría de los escenarios un mejor rendimiento que los otros modelos de clasificación. No obstante, las diferencias de rendimiento entre el *random forest* y el *boosting* no fueron significativas en ninguno de los escenarios propuestos, por lo que el *random forest* sería un modelo de clasificación igualmente válido.

Finalmente, y con el fin de evitar en la medida de lo posible clasificaciones erróneas, en el estudio de la dispersión en el medio marino, únicamente se clasificaron aquellos ejemplares de ambas especies que presentaran una probabilidad mínima de asignación a su río de origen de un 80%. No obstante, debido a la baja proporción de ejemplares clasificados de forma exitosa siguiendo este criterio en el estudio de la filopatria, se optó por clasificar la totalidad de individuos.

3.2. Clasificación de los ejemplares adultos

Tras seleccionar el modelo con el que llevar a cabo la clasificación de los ejemplares adultos capturados en los diferentes medios, en el presente apartado se recopilan los resultados obtenidos de dicha clasificación. Por un lado, se han integrado dos gráficos en los que se puede observar el número de ejemplares asignado a cada uno de los ríos contemplados en el estudio de la filopatria, por un lado, y en el estudio de la dispersión por el otro.

Por otro lado, dos gráficos correspondientes al estudio de la filopatria y al de la dispersión, como en el caso anterior, donde aparecen reflejadas las probabilidades de pertenecer a cada uno de los ríos analizados en función de tres umbrales mínimos de probabilidad de asignación a un río determinado.

3.2.1. Ejemplares capturados en el medio fluvial (estudio de la filopatria)

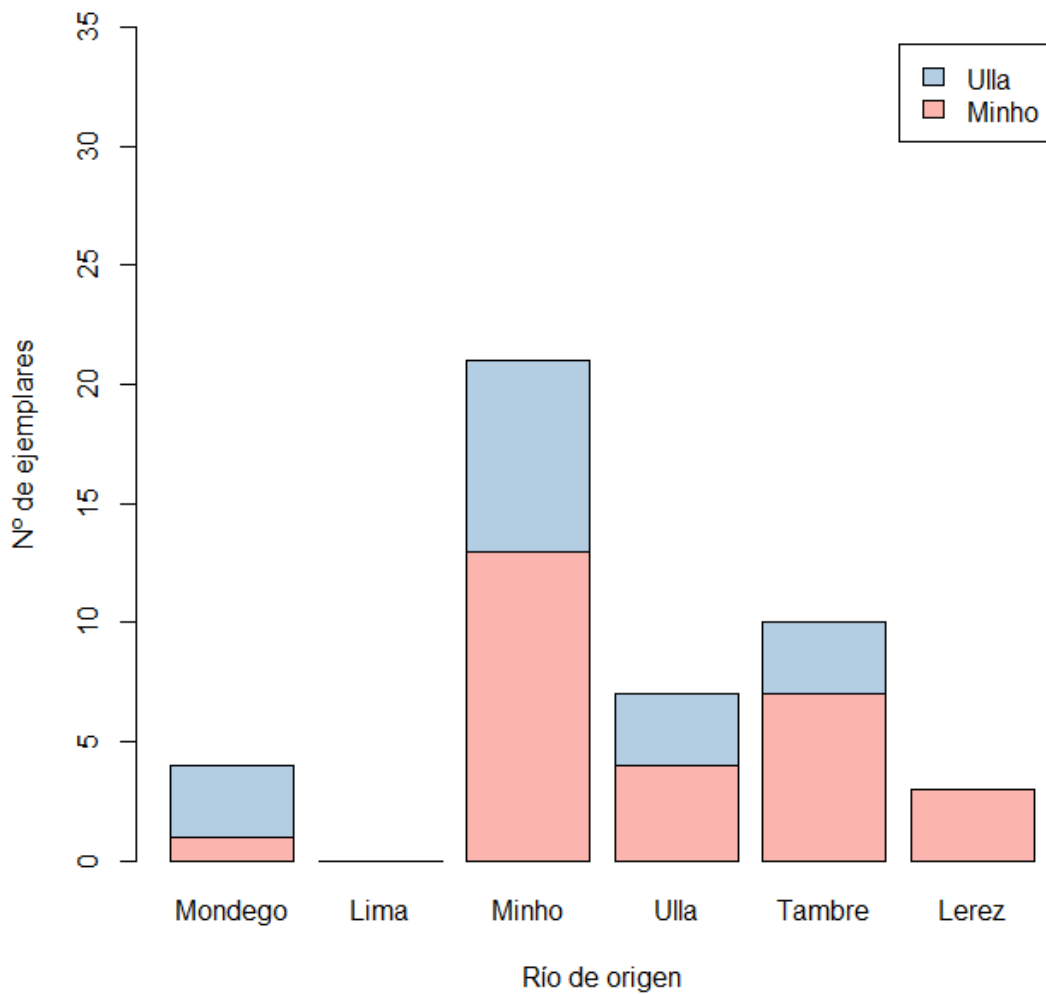


Figura 25: Clasificación de los ejemplares adultos capturados en el río Ulla (azul) y el río Miño (rojo), en sus respectivos ríos de origen.

Tabla 7: Número de ejemplares de diferentes orígenes capturados en los ríos Ulla y Miño.

	Mondego	Lima	Minho	Ulla	Tambre	Lérez
Minho	1	0	13	4	7	3
Ulla	2	0	8	3	3	0

Tal y como se puede observar en la figura 25 y tabla 7, el mayor número de ejemplares reproductores capturados en los ríos Ulla y Miño, corresponden a individuos nacidos en el río Miño. Además, se aprecia la existencia en estos ríos de ejemplares nacidos en los ríos Lárez y Tambre, lugares en los que no se ha constatado la presencia de poblaciones de estas especies.

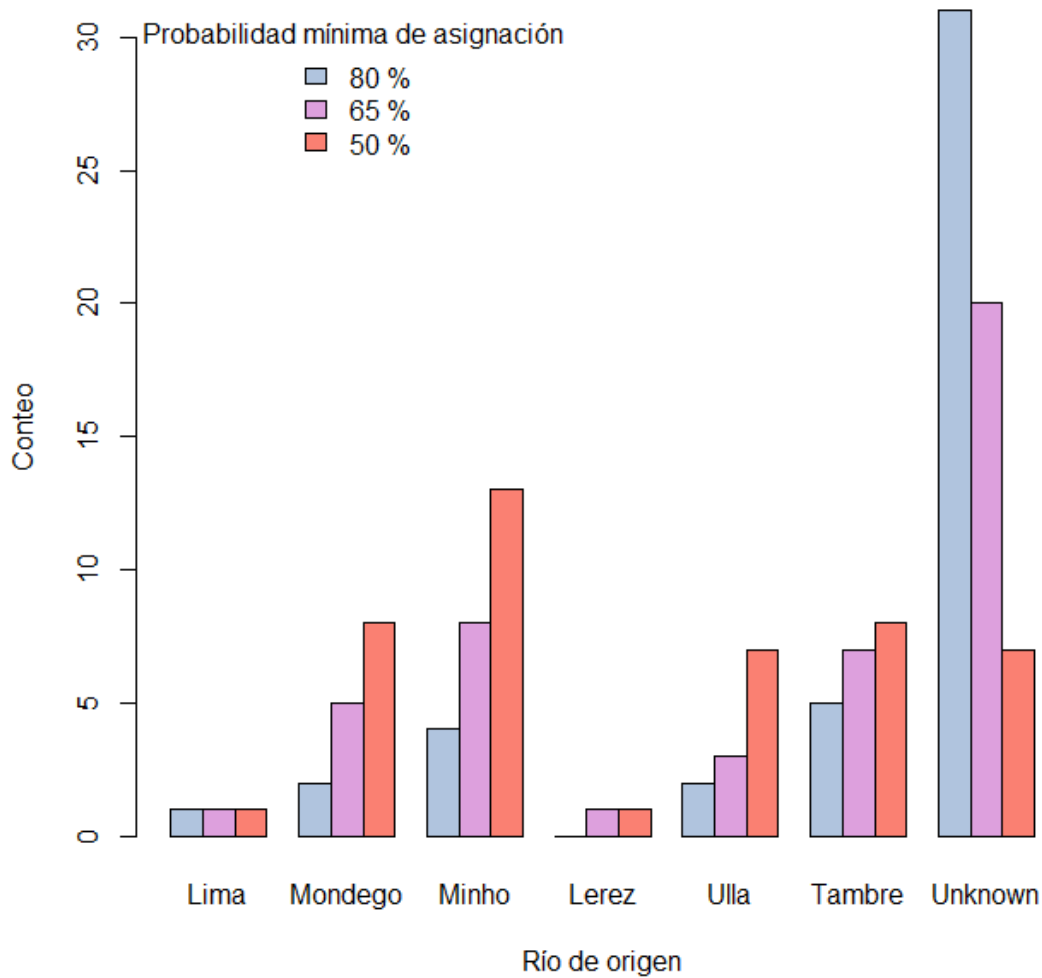


Figura 26: Número de ejemplares capturados en medio fluvial y asignados a su río de origen en función de los diferentes umbrales de probabilidades mínimas de asignación a un río específico. Aquellos ejemplares que no alcancen un umbral fueron clasificados como de origen desconocido (*unknown*).

Tabla 8: Número de ejemplares capturados en el medio fluvial asignados a su río de origen en función a los diferentes umbrales de probabilidad mínima de asignación.

	Lima	Mondego	Minho	Ulla	Tambre	Lárez	Unknown
80%	1	2	4	0	2	5	31
65%	1	5	8	1	3	7	20
50%	1	8	13	1	7	8	7

Al observar la figura 26 y tabla 8, podemos ver como la mayoría de los ejemplares presentan una probabilidad de asignación a su río de origen menor al 80%. A medida que esta restricción se relaja el número de ejemplares clasificados aumenta de forma considerable para la mayoría de los ríos.

3.2.2. Ejemplares capturados en el medio marino (estudio de la dispersión)

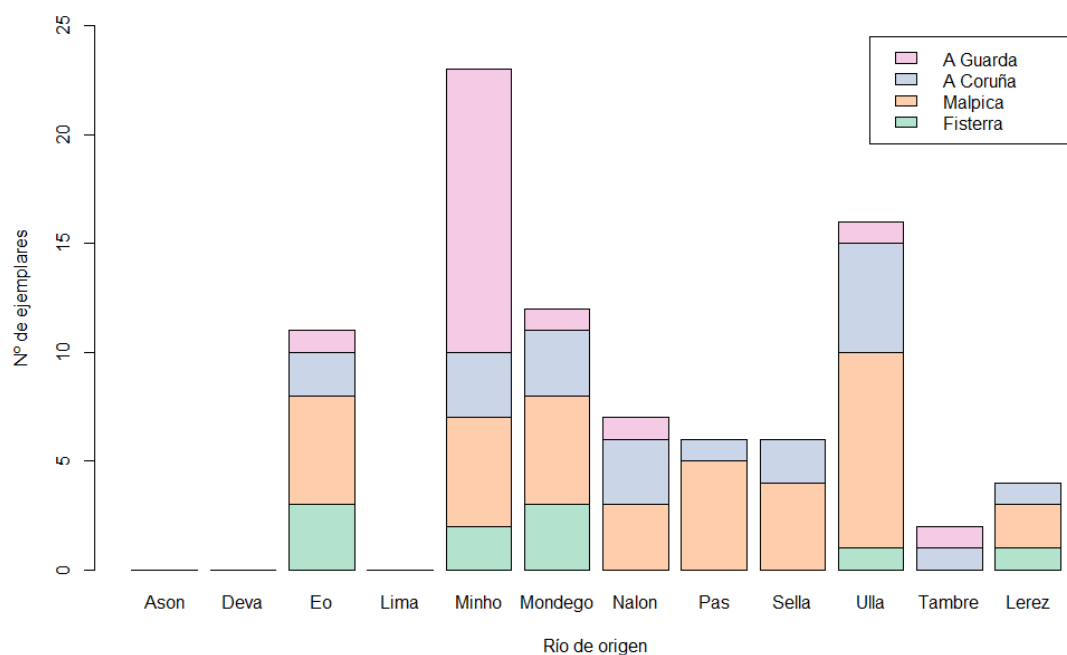


Figura 27: Clasificación en sus respectivos ríos de origen de los ejemplares adultos capturados localizaciones costeras de A Guarda, A Coruña, Malpica y Fisterra. Únicamente fueron clasificados los ejemplares cuya probabilidad mínima de asignación a un río concreto fue igual o superior al 80%.

Tabla 9: Número de ejemplares de diferentes orígenes capturados en diferentes puntos de la costa de Galicia.

	Asón	Deva	Eo	Lima	Minho	Mondego	Nalón	Pas	Sella	Ulla	Tambre	Lérez
Fisterra	0	0	3	0	2	3	0	0	0	1	0	1
Malpica	0	0	5	0	5	5	3	5	4	9	0	2
A_Coruña	0	0	2	0	3	3	3	1	2	5	1	1
A_Guarda	0	0	1	0	13	1	1	0	0	1	1	0

Los resultados obtenidos de la clasificación de los ejemplares adultos capturados en medio marino (figura 27 y tabla 9), indican que el mayor número de ejemplares provienen de los ríos Minho y Ulla seguidos de los ríos Mondego y Eo. Ejemplares de estos cuatro ríos se encontraron en todas las localizaciones costeras contempladas en el estudio. Además, en la costa de A Coruña,

aparecieron ejemplares de todos los ríos a excepción de aquellos de los que no se capturó ningún ejemplar (Asón, Deva y Lima).

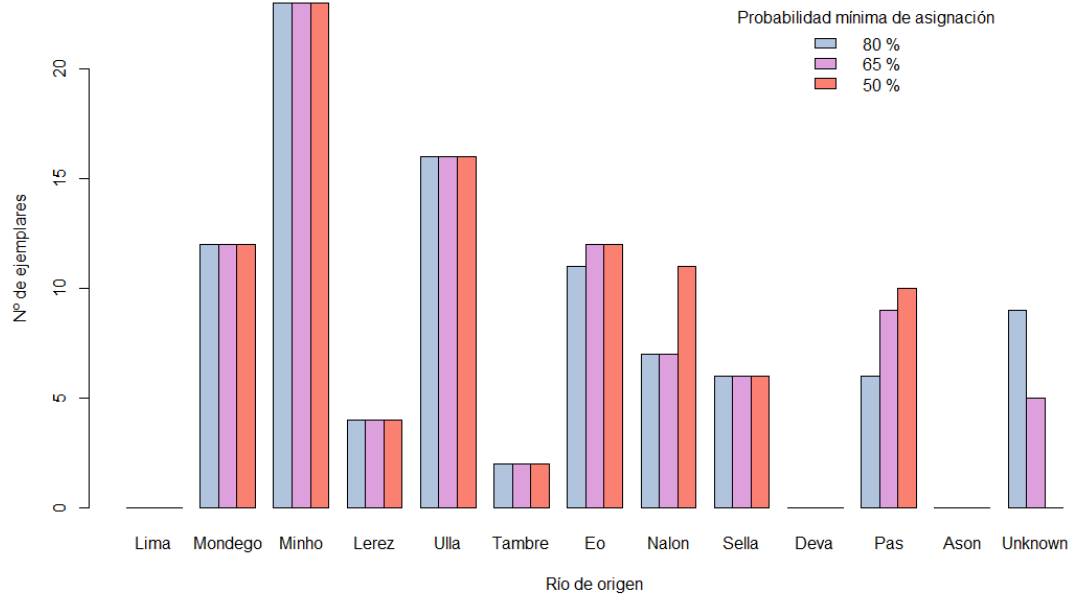


Figura 28: Número de ejemplares capturados en medio marino y asignados a su río de origen en función de los diferentes umbrales de probabilidades mínimas de asignación a un río específico. Aquellos ejemplares que no alcancen un umbral fueron clasificados como de origen desconocido (*unknown*).

Tabla 10: Número de ejemplares capturados en el medio marino asignados a su río de origen en función a los diferentes umbrales de probabilidad mínima de asignación. Unk. Indica origen desconocido (*unknown*).

	Lima	Mondego	Minho	Lérez	Ulla	Tambre	Eo	Nalón	Sella	Deva	Pas	Asón	Unk.
80%	0	12	23	4	16	2	11	7	6	0	6	0	9
65%	0	12	23	4	16	2	12	7	6	0	9	0	5
50%	0	12	23	4	16	2	12	11	6	0	10	0	0

En contraste con lo ocurrido en la clasificación de los ejemplares capturados en el medio fluvial, la mayoría de los ejemplares capturados en el medio marino presentan una probabilidad de asignación a un río en concreto superior al 80% (figura 28 y tabla 10). De hecho, cuando el umbral es rebajado al 50% de probabilidad mínima de asignación, no existe ningún ejemplar clasificado como de origen desconocido.

Capítulo 4

Discusión y conclusiones

En este capítulo final se interpretan los resultados obtenidos durante la realización del estudio y se justificará la consecución de los objetivos propuestos explicando de forma detallada las principales conclusiones alcanzadas. Además, se desarrollan los principales puntos débiles del proyecto y se intentan proponer una serie de medidas a aplicar para subsanar en la medida de lo posible las mencionadas debilidades en vista de futuros estudios aplicados a especies similares o extensiones del presente trabajo.

4.1. Rendimiento y comparación de los modelos

Por un lado, cuando ajustamos los modelos de clasificación para el estudio de la filopatria, observamos que, en primera instancia, los modelos *machine learning* basados en árboles de decisión, *boosting* y *random forest*, presentan un mejor rendimiento que el modelo bayesiano ingenuo y el modelo QDA. Al aumentar el tamaño muestral empleado para el entrenamiento de los clasificadores, se mejoró el rendimiento de todos los modelos, además, el modelo bayesiano ingenuo fue el que experimentó una mayor mejoría alcanzando una precisión media del 80%. No obstante, pese al aumento percibido en la calidad predictiva de los modelos generadores, estos no llegaron a alcanzar la precisión de los modelos basados en árboles de decisión.

Por otro lado, en el caso de los modelos ajustados para el análisis de la dispersión, todos alcanzaron un rendimiento competente con el tamaño muestral más pequeño y, si bien en el cómputo global los modelos con mejor rendimiento continuaron siendo aquellos basados en árboles de decisión, no se observó una diferencia tan acusada respecto a los modelos generadores tal y como ocurría en el caso anterior. Con mayores tamaños muestrales, se obtuvieron mejores resultados en todos los modelos, sin embargo, se hizo patente el hecho de que alcanzado cierto umbral de precisión, aumentar el tamaño muestral tan solo tuvo un leve efecto sobre la mejoría de la precisión de los modelos.

4.2. La importancia del ratio isotópico

A colación de los resultados obtenidos, es evidente que los modelos ajustados para el análisis de la dispersión en el mar ofrecieron mejores resultados que los modelos ajustados para el estudio de la filopatria. De hecho, los resultados obtenidos por parte de los modelos ajustados para el primer caso con el menor tamaño muestral, son equiparables a los resultados observados por los modelos del segundo caso con el mayor de los tamaños muestrales.

La única diferencia en el ajuste de ambos modelos fue la utilización o no del ratio isotópico $\text{Sr}^{87}/\text{Sr}^{86}$ como parámetro para la clasificación. Numerosos estudios tales como los mencionados trabajos de Walter *et al.* (2008), Martin *et al.* (2015) o Randon *et al.* (2018) o defienden el uso de esta relación isotópica como parámetro clave para la discriminación del origen de especies diádromas más allá de las estudiadas en el presente trabajo.

Otra prueba de la importancia del ratio isotópico se evidencia al observar las probabilidades de asignación a un río en concreto de los ejemplares de ambos estudios. Mientras que en el estudio de la filopatria observamos que el número de ejemplares con una probabilidad de asignación superior al 80% es bajo, en el estudio de la dispersión la mayor parte de los ejemplares ya son asignados a un río en este umbral. Además, en el primer caso podemos ver como el número de ejemplares asignados va en aumento a medida que desciende el umbral mínimo, sin embargo, en el segundo caso, como la mayoría de las alosas ya se encuentran clasificadas, a medida que desciende el umbral son menos las alosas a asignar, hasta el punto de que cuando se establece el umbral al 50% todas las alosas se encuentran asignadas a un río. Por su parte, en el estudio de la filopatria todavía restarían por clasificar siete individuos al fijar la probabilidad mínima de asignación a un río determinado en el 50%.

De esta forma y en vista de los resultados, se respalda la idea de que efectivamente el ratio isotópico entre estos isótopos del Sr es un factor que permite la discriminación entre diferentes cursos fluviales. Ahora bien, la carencia de este parámetro a la hora de clasificar los ejemplares reproductores capturados en los ríos, se pudo llegar a compensar de forma parcial en los modelos de clasificación. Así pues, y tal y como se expuso anteriormente, el aumento del tamaño muestral a la hora de entrenar los modelos permitió la obtención de resultados equiparable a aquellos obtenidos con un tamaño muestral menor pero empleado el ratio $\text{Sr}^{86}/\text{Sr}^{87}$. Sin embargo, el aumento del tamaño muestral tiene un efecto finito sobre los umbrales de precisión que se pueden llegar a alcanzar y a iguales tamaños muestrales, los modelos que dispongan de este parámetro isotópico siempre obtendrán un mejor rendimiento.

4.3. Origen de los ejemplares capturados en el medio fluvial

En el siguiente apartado se intentará dar una explicación a los resultados obtenidos en la clasificación de los ejemplares de alosa reproductores capturados en el medio fluvial. Con este fin se abordarán los diferentes ríos contemplados en el estudio y que han sido agrupados por criterios de proximidad y presencia/ausencia, *a priori*, de poblaciones de las especies objetivo.

4.3.1. Los ríos Miño y Ulla

En vista de los resultados obtenidos, observamos que el río Miño se trata del curso fluvial del cual provienen la mayoría de los ejemplares capturados en las diferentes cuencas analizadas y al

cual llegan a reproducirse un mayor número de ejemplares de entre los mismos ríos. El Miño es el río más largo de Galicia y uno de los ríos más caudalosos de España (340m³/s de caudal medio).

Este curso fluvial presenta poblaciones estables de ambas especies de alosas, e históricamente la pesca de la especie *A. alosa* en el tramo internacional del río Miño (TIRM) ha sido una importante actividad económica para el sector pesquero de la zona. Se trata pues, de un río que debido a sus características presenta la capacidad de albergar un gran número de ejemplares de ambas especies de alosa (capacidad de carga) cuya reproducción es viable en dichas aguas. Como consecuencia y atendiendo a los resultados, la población del río Miño parece actuar como una población fuente, dando lugar a ejemplares que, en lugar de volver a su río de origen, exploran otros cursos fluviales con el fin de reproducirse. De esta forma, encontramos ejemplares nacido en el río Miño en todos los ríos analizados.

No obstante, el río Miño también actúa como receptor de ejemplares de origen externo. De hecho, se capturaron más ejemplares originarios del río Ulla en el río Miño que en el propio Ulla. El río Ulla se trata del otro río, junto al Miño, del que se tiene constancia de la existencia de una población de alosas. Sin embargo, presenta una serie de características (menor caudal y longitud) que se pueden extrapolar a una capacidad de carga menor a la del Miño.

4.3.2. Los ríos Lérez y Tambre

Tal y como se mencionó anteriormente, en Galicia tan solo se tiene constancia de la existencia de dos ríos que alberguen poblaciones de alosas, los ríos Miño y Ulla. Sin embargo, en Galicia existen ríos que cuyas características fisicoquímicas permiten de forma potencial la existencia de estas especies en sus aguas. En el caso de los ríos Lérez y Tambre estas condiciones se cumplen y en vista de los resultados, parece ser que existen ejemplares nacidos en estos ríos que fueron capturados en los ríos donde sí se tenía constancia de la existencia de estas especies.

El hallazgo de estas especies en ríos donde no se tenía constancia de la existencia de poblaciones, presenta nuevas implicaciones en su gestión ecológica. Siendo *A. alosa* y *A. fallax* especies amparadas por diversas figuras de protección a nivel nacional y europeo, el descubrimiento de nuevas poblaciones permitirá a las autoridades competentes trazar nuevos planes para su gestión y protección.

4.3.3. El río Mondego

Si bien las alosas son especies con un marcado comportamiento filopátrico, existen ejemplares de estas especies han remontado ríos con el fin de reproducirse que distan cientos de kilómetros respecto a su río de origen (Martin *et al.* 2015). En el presente estudio, se han encontrado ejemplares reproductores tanto en el río Miño como en el Ulla originarios del río Mondego, río portugués cuya desembocadura dista más de 190 km de la desembocadura del río Miño.

Como ya se comentó, entre el río Miño y Ulla, el primero debido a sus características es el que presenta una mayor capacidad de carga y así se ve reflejado en los resultados, sin embargo, se han encontrado un mayor número de ejemplares originarios del río Mondego en el río Ulla que en el río Miño siendo este último además el más próximo al Mondego de los dos.

4.3.4. El comportamiento filopátrico

Las especies *A. alosa* y *A. fallax* presentan un marcado comportamiento filopátrico (Nachón 2016), no obstante, existen ejemplares que, en lugar de remontar sus ríos de origen con el fin de reproducirse, se adentran en cursos fluviales diferentes (Martin *et al.* 2015). En el presente estudio hemos podido observar que el grado de filopatria de estas especies en aguas gallegas dista de alcanzar el 100%, siendo una importante cantidad de ejemplares los que se adentran a reproducirse en ríos diferentes al de su origen.

La forma en la que un ejemplar de alosa se orienta y posibilita la llegada a su río de origen con fines reproductivos todavía es desconocido. De la misma forma, también es desconocido el motivo que lleva a ciertos ejemplares a remontar un río diferente. No obstante, se especula que las alosas puedan orientarse de forma similar a la realizada por el salmón, especie anádroma con un marcado comportamiento filopátrico también.

Tal y como se ha presentado a lo largo del trabajo, los ríos presentan una diferente composición microquímica que permite diferenciarlos entre ellos. Cuando desemboca un río en el océano, el agua de ambos se mezcla de forma paulatina ya que las características fisicoquímicas de las aguas son muy diferentes. Esto crea una pluma que se extiende desde la desembocadura del río y se va adentrando en el océano hasta desaparecer, cuando la composición del agua vuelve a ser homogénea debido a un proceso de dilución y mezcla. Los salmones reproductores son capaces de detectar este gradiente químico (la pluma) “huelen” en el agua y son capaces de reconocer su río de origen a través de ese olor característico de cada cuenca.

En el caso de las alosas, de ser cierta esta hipótesis explicaría en gran parte los resultados obtenidos en el estudio. Para empezar, al ser el río Miño el más caudaloso de Galicia y uno de los más caudalosos de España, su pluma es capaz de adentrarse en el océano en mucha mayor medida que el resto de los ríos gallegos. Como consecuencia de esto y debido a la relativa proximidad de los diferentes ríos de la costa gallega es posible que muchos ejemplares reproductores se dejen llevar por la pluma del río Miño y remonten un río distinto al suyo por equivocación.

No obstante, esto no explicaría todavía a mayor cantidad de ejemplares del río Mondego capturados en aguas del Ulla respecto al Miño, ya que el río Mondego y el río Miño son más próximos entre sí tanto en distancia de sus desembocaduras como en composición microquímica, lo que provocaría que ambos ríos “oliesen” de una forma más similar. Así pues, se debe seguir

investigando los mecanismos a través de los cuales estas especies se orientan y eligen sus ríos de reproducción.

4.4. Análisis de la dispersión en el medio marino

En conjunto, las capturas realizadas en Malpica y A Coruña suponen la mayor parte del total muestral. Estas dos localizaciones se encuentran relativamente próximas entre sí y parecen ser una zona de interés para estas especies puesto que en ellas se reúnen alosas provenientes tanto de ríos de la vertiente cántabra como de ríos portugueses. El motivo por el cual las alosas originarias de diferentes puntos peninsulares se desplazan cientos de kilómetros para confluir en aguas de Galicia no está todavía esclarecido. No obstante, se hipotetiza que estas migraciones a lo largo de la costa puedan deberse a la búsqueda de alimento, lo que supondría que ejemplares de diferentes orígenes confluyesen e interactuasen en la costa gallega con fines tróficos.

Por otra parte, el motivo por el cual en A Guarda la mayoría de las capturas provienen del río Miño, se debe a que este punto de muestreo se encuentra muy próximo a la desembocadura del río al cual se le supone la mayor capacidad de carga de estas especies de Galicia. De este modo, una importante parte de los ejemplares salidos de este río optarían por quedarse en las inmediaciones costeras, presentando así un comportamiento que contrasta con el carácter migratorio de otros congéneres.

Así pues, se puede concluir que en Galicia se lleva a cabo la interacción entre comunidades de alosas provenientes de puntos tan distantes entre sí como la costa portuguesa y la cántabra. Este hecho presenta abre la puerta a comenzar a hipotetizar sobre una posible dinámica meta poblacional que se pudiese estar estableciendo entre las poblaciones de alosas ibéricas. Además, ante los resultados obtenidos y siendo las alosas especies paraguas⁴, sería de interés promover y designar como áreas protegidas las zonas costeras en las que estas especies tienden a congregarse.

4.5. Limitaciones del trabajo y propuestas

El presente estudio no deja de adolecer de una serie de “puntos débiles” sujetos a ser mejorados en futuras investigaciones. Las principales debilidades las encontramos en la cantidad de información disponible para la realización del estudio. Por un lado, debido a que estas especies han sido poco estudiadas, los datos disponibles de las mismas distan de ser amplios. Debido a esto, para la realización del trabajo no nos fue posible contar con muestras de juveniles de todos los ríos analizados por lo que se tuvo que recurrir a la inferencia de diferentes parámetros con el fin de remediar en la medida de lo posible esta falta de información. Por otro lado, si bien

⁴ Especies seleccionadas para tomar decisiones relacionadas con la conservación, usualmente porque protegiendo estas especies, se protegen de forma indirecta muchas otras especies que componen la comunidad de su hábitat.

contamos con un rango espacial aceptable en lo que respecta a la localización de los ríos estudiados, existen cursos fluviales dentro de este rango espacial de los cuales no poseemos muestras de agua, por lo que nos fue imposible contemplarlos en el estudio.

Otra limitación del estudio la encontramos en el análisis de los otolitos de los ejemplares adultos capturados en los ríos. Dicho análisis fue llevado a cabo con anterioridad a la conceptualización del presente estudio y no se cuantificó el ratio isotópico $\text{Sr}^{87}/\text{Sr}^{86}$. Este parámetro es clave para la correcta asignación de los ejemplares capturados a su río de origen. Como consecuencia, los resultados obtenidos en el estudio de la filopatria de ambas especies presentan una mayor imprecisión respecto a los resultados del estudio de la dispersión en el mar, donde sí nos fue posible calcular los valores de este parámetro en los ejemplares capturados.

Con el fin de llevar a cabo futuras investigaciones en base al presente trabajo, se propone realizar un proyecto de recopilación de información con el fin de completar en la medida de lo posible los “huecos” presentes en la base de datos empleada para la realización de este estudio.

Finalmente, con el propósito de mejorar la precisión de los resultados obtenidos y la robustez de los modelos empleados, sería conveniente para futuros estudios profundizar en la selección de los diferentes hiperparámetros de los modelos a través de metodologías de búsqueda que permitan la obtención de valores óptimas de forma que sea asequible en términos computacionales. Además, la exploración de nuevas metodologías de clasificación junto con el ensamblado de modelos podría dar lugar a un aumento de la calidad predictiva cuando la cantidad de información disponible es limitada.

Anexo

En el siguiente apartado se recogen los resultados del proceso de validación cruzada llevado a cabo para la selección óptima de los hiperparámetros de los modelos basados en árboles de decisión *stochastic gradient boosting* y *random forest*. Asimismo, se muestran los valores óptimos obtenidos para cada uno de estos modelos teniendo en cuenta los diferentes tamaños muestrales manejados a lo largo del estudio y de si el modelo ajustado fue empleado para el estudio de la filopatria o de la dispersión durante la fase marina de las alosas.

Modelos: stochastic gradient boosting

Modelos seleccionados para el estudio de la filopatria

$n = 10$

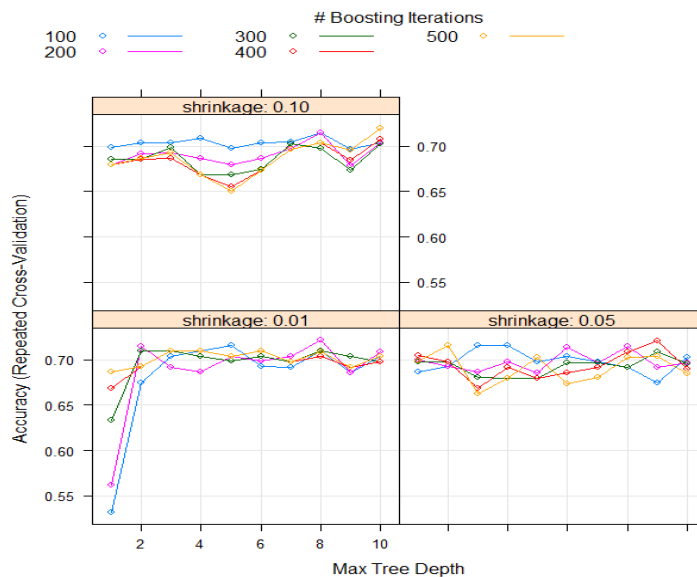


Figura 29: Selección de hiperparámetros óptimos del modelo *stochastic gradient boosting* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la filopatria.

- Modelo óptimo:
 - Número de árboles: 200
 - Número de divisiones del árbol: 7
 - Parámetro de regularización: 0.01

$n = 30$

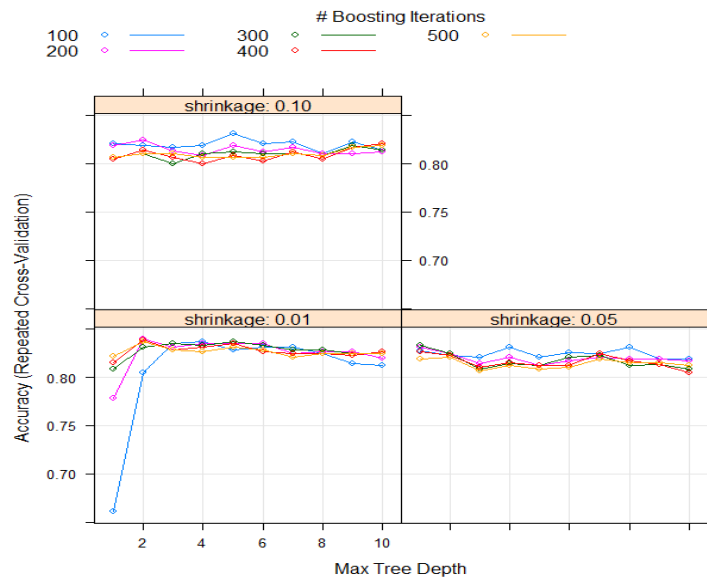


Figura 30: Selección de hiperparámetros óptimos del modelo *stochastic gradient boosting* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la filopatría.

- Modelo óptimo:
 - Número de árboles: 200
 - Número de divisiones del árbol: 2
 - Parámetro de regularización: 0.01

$n = 60$

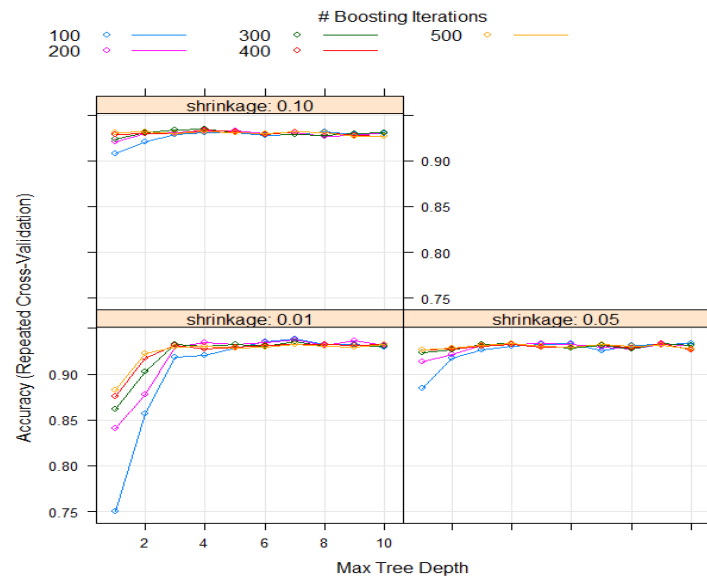


Figura 31: Selección de hiperparámetros óptimos del modelo *stochastic gradient boosting* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la filopatría.

- Modelo óptimo:
 - Número de árboles: 100
 - Número de divisiones del árbol: 7
 - Parámetro de regularización: 0.01

Modelos seleccionados para el estudio de la dispersión

$n = 10$

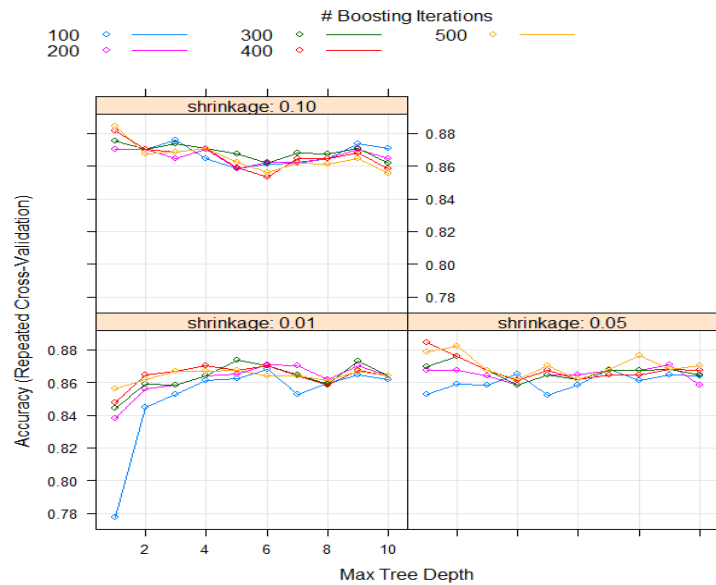


Figura 32: Selección de hiperparámetros óptimos del modelo *stochastic gradient boosting* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la dispersión en fase marina.

- Modelo óptimo:
 - Número de árboles: 400
 - Número de divisiones del árbol: 1
 - Parámetro de regularización: 0.05

$n = 30$

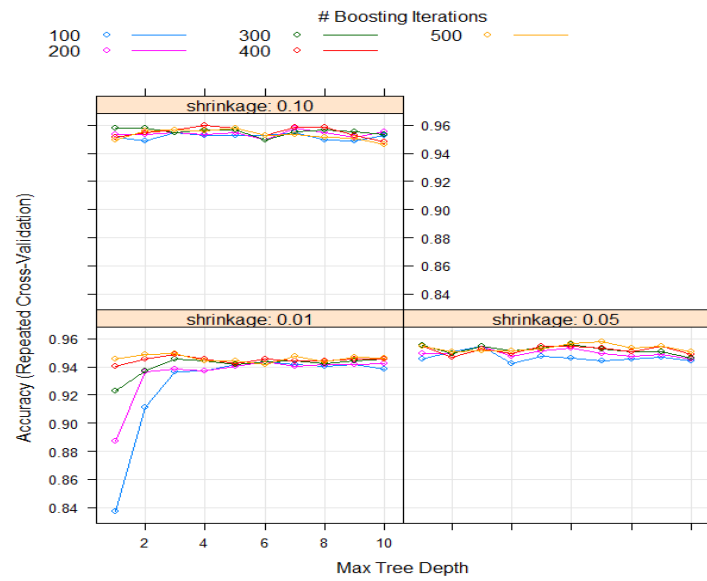


Figura 33: Selección de hiperparámetros óptimos del modelo *stochastic gradient boosting* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la dispersión en fase marina.

- Modelo óptimo:
 - Número de árboles: 400
 - Número de divisiones del árbol: 4
 - Parámetro de regularización: 0.1

$n = 60$

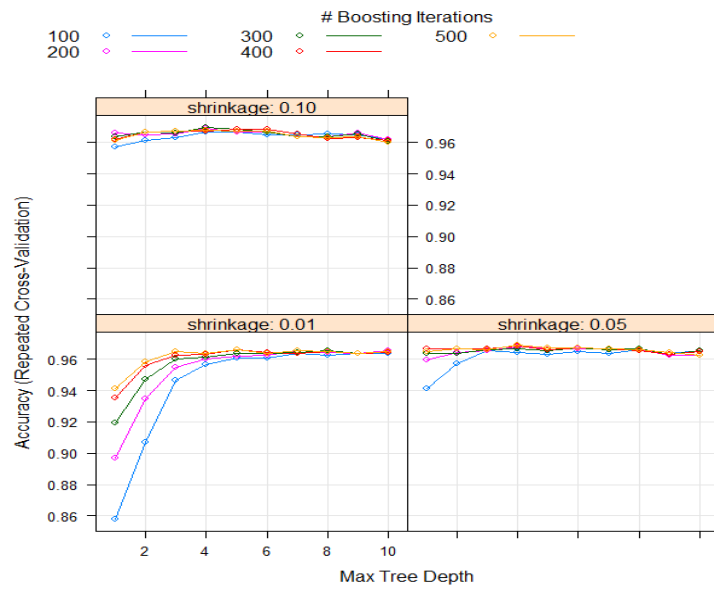


Figura 34: Selección de hiperparámetros óptimos del modelo *stochastic gradient boosting* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la dispersión en fase marina.

- Modelo óptimo:
 - Número de árboles: 300
 - Número de divisiones del árbol: 4
 - Parámetro de regularización: 0.1

Modelos: random forest

Modelos seleccionados para el estudio de la filopatria

$n = 10$

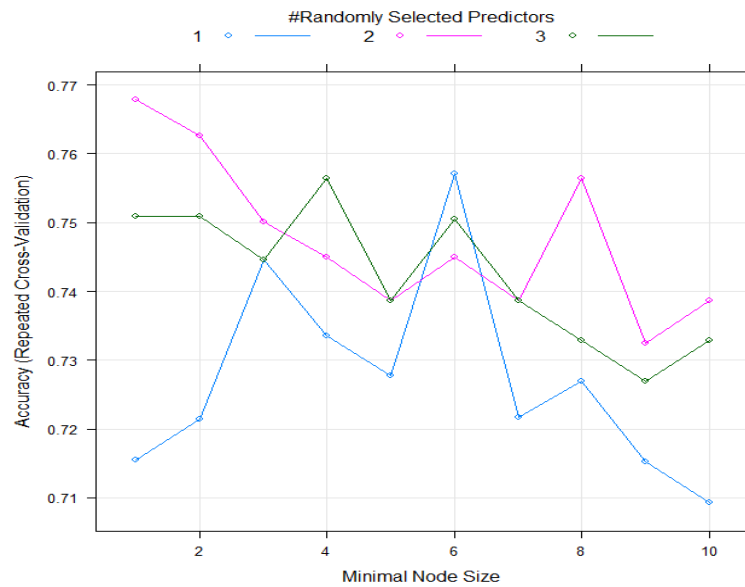


Figura 35: Selección de hiperparámetros óptimos del modelo *random forest* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la filopatria.

- Modelo óptimo:
 - Número de predictores: 2
 - Número mínimo de observaciones en el nodo final: 1
 - Número prefijado de árboles: 1000

$n = 30$

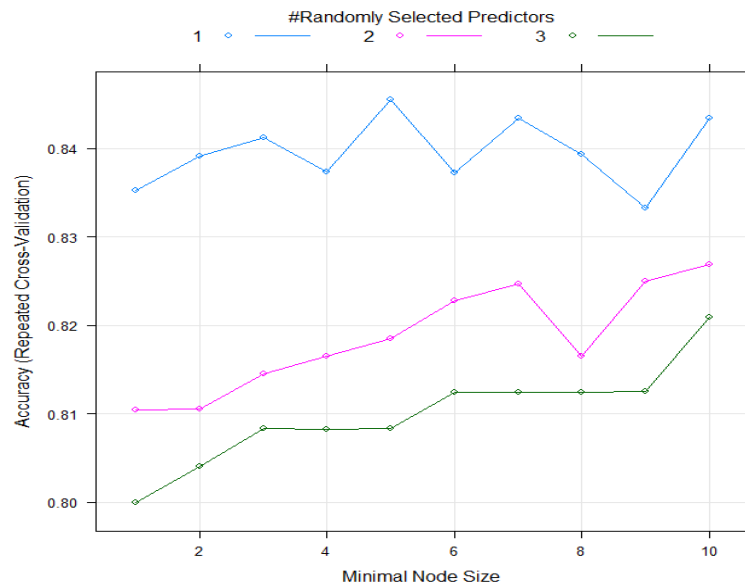


Figura 36: Selección de hiperparámetros óptimos del modelo *random forest* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la filopatria.

- Modelo óptimo:
 - Número de predictores: 1
 - Número mínimo de observaciones en el nodo final: 5
 - Número prefijado de árboles: 1000

$n = 60$

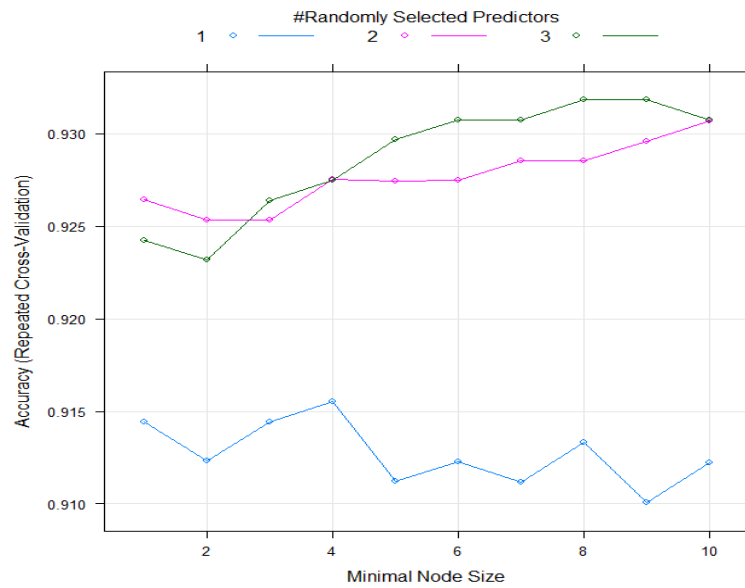


Figura 37: Selección de hiperparámetros óptimos del modelo *random forest* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la filopatria.

- Modelo óptimo:
 - Número de predictores: 3
 - Número mínimo de observaciones en el nodo final: 8
 - Número prefijado de árboles: 1000

Modelos seleccionados para el estudio de la dispersión

$n = 10$

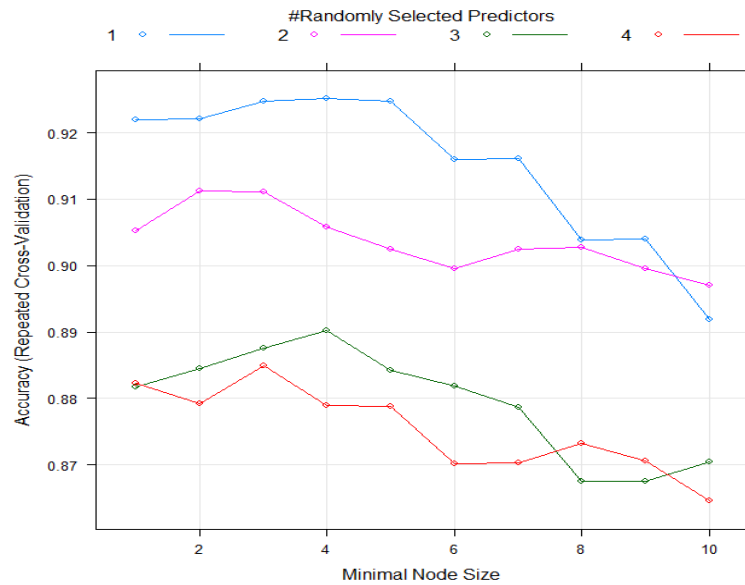


Figura 38: Selección de hiperparámetros óptimos del modelo *random forest* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la dispersión.

- Modelo óptimo:
 - Número de predictores: 1
 - Número mínimo de observaciones en el nodo final: 4
 - Número prefijado de árboles: 1000

$n = 30$

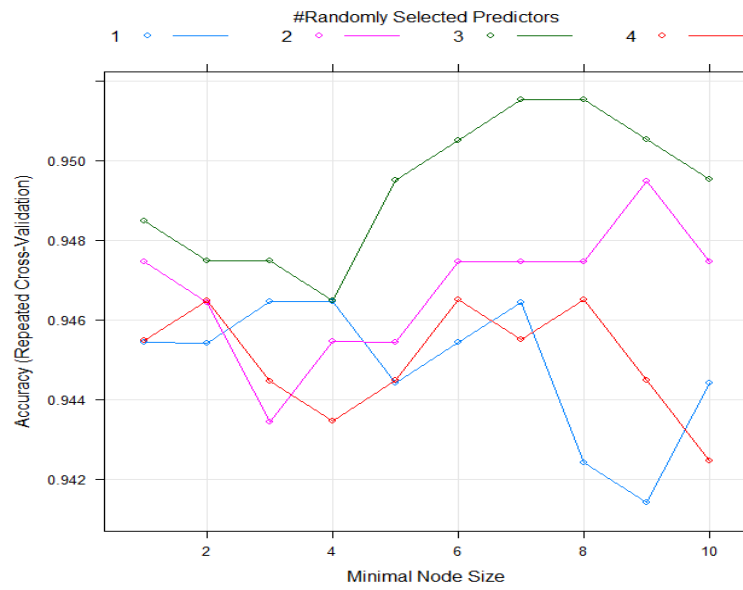


Figura 39: Selección de hiperparámetros óptimos del modelo *random forest* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la filopatria.

- Modelo óptimo:
 - Número de predictores: 3
 - Número mínimo de observaciones en el nodo final: 7
 - Número prefijado de árboles: 1000

$n = 60$

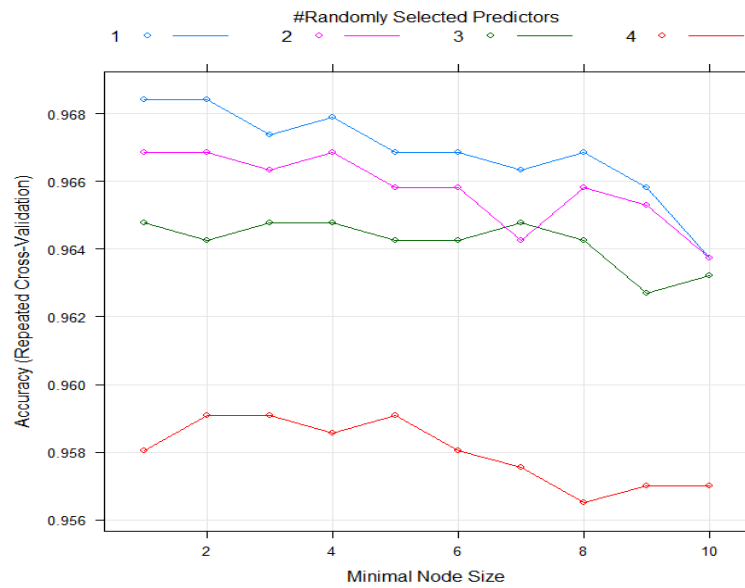


Figura 40: Selección de hiperparámetros óptimos del modelo *random forest* mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la filopatria.

- Modelo óptimo:
 - Número de predictores: 1
 - Número mínimo de observaciones en el nodo final: 1
 - Número prefijado de árboles: 1000

Referencias bibliográficas

- Amat J. (2018). Machine Learning con R y caret.
https://www.cienciadedatos.net/documentos/41_machine_learning_con_r_y_caret. Accedido 20 de junio de 2022.
- Aprahamian, M. W., Baglinière, J. L., Sabatié, R., Alexandrino, P. y Aprahamian, C. D. (2003b). *Alosa alosa and Alosa fallax spp.: literature review and bibliography*. RandD technical report W1-014/TR, Environment Agency, Bristol.
- Aprahamian, M. W., Baglinière, J. L., Sabatié, R., Alexandrino, P., Thiel, R. y Aprahamian, C. D. (2003a). Biology, status, and conservation of the anadromous Atlantic twaite shad *Alosa fallax fallax*. *Am. Fish. Soc. Symp.*, 35: 103-124.
- Baglinière, J. L., Sabatié, M. R., Rochard, E., Alexandrino, P. y Aprahamian, M. W. (2003). The allis shad *Alosa alosa*: biology, ecology, range, and status of populations. *Am. Fish. S. S.*, 35: 85-102.
- Bath, G. E., Thorrold, S. R., Jones, C. M., Campana, S. E., McLaren, J. W. y Lam, J. W. H. (2000). Strontium and barium uptake in aragonitic otoliths of marine fish. *Geochim. Cosmochim. Acta.*, 64: 1705-1714.
- Campana, S. E. y Thorrold, S. R. (2001). Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations?. *Can. J. Fish. Aquat. Sci.*, 58: 30-38.
- Cox, E. R. y Snell, E. J. (1968). A General Definition of Residuals. *J R Stat Soc Series B Stat Methodol.*, 30: 248-265.
- Farrell, J. y Campana, S. E. (1996). Regulation of calcium and strontium deposition on the otoliths of juvenile tilapia, *Oerochromis niloticus*. *Comp. Biochem. Physiol.*, 115: 103-109.
- Fernández, R., Costa, J. y Oviedo, M. (2021). Aprendizaje Estadístico (apuntes de la asignatura de Aprendizaje Estadístico del Máster en Técnicas Estadísticas).
https://rubenfcasal.github.io/aprendizaje_estadistico.
- Gahagan, B. I., Vokoun, J. C., Whitledge, G. W., Schultz, E. T. (2012). Evaluation of Otolith Microchemistry for Identifying Natal Origin of Anadromous River Herring in Connecticut. *Mar. Coast. Fish.*, 4: 358-372.
- Guinand, B., Topchy, A., Page, K.S., Burnham-Curtis, M.K., Punch, W.F. y Scribner, K.T. (2002). *J. Hered.*, 93: 260-269.

- King, J. J. y Roche, W. K. (2008). Aspects of anadromous Allis shad (*Alosa alosa* Linnaeus) and Twaite shad (*Alosa fallax* Lacépède) biology in four Irish Special Areas of Conservation (SACs): status, spawning indications and implications for conservation designation. *Hydrobiologia*, 602: 145-154.
- Martin, J. (2013). Développement de la microchimie élémentaire et isotopique ($^{87}\text{Sr}:^{86}\text{Sr}$) des otolithes de saumon Atlantique: évaluation du potentiel pour un appui à la gestion piscicole dans le bassin de l'Adour. Thèse de Doctorat. Université de Pau et des Pays de l'Adour, Pau, France.
- Martin, J., Rougemont, Q., Drouineau, H., Launey, S., Jatteau, P., Bareille, G., *et al.* (2015). Dispersal capacities of anadromous Allis shad population inferred from a coupled genetic and otolith approach. *Can. J. Fish. Aquat. Sci.* 72(7): 991-1003.
- Mirzaei, S., Borzadaran, G. R., Amini, M. y Jabbari, H. (2017). Some properties of the Canberra inequality index. *Hacettepe J. Math. Stat.*, 46: 1159-1174.
- Mota, M. (2014). Biology and Ecology of the Allis shad, *Alosa alosa* (Linnaeus, 1758), in the Minho River. Tese de Doutoramento. Instituto de Ciências Biomédicas Abel Salazar da Universidade de Porto, Porto, Portugal.
- Nachón, D.J. (2017). Dinámica poblacional y microquímica de los otolitos de las poblaciones de saboga, *Alosa fallax* (Lacépède, 1803), de los ríos Ulla y Miño. Tesis, Universidade de Santiago de Compostela, Santiago de Compostela, España.
- Nachón, D.J., Mota, M., Antunes, C., Servia, M.J., and Cobo, F. (2016). Marine and continental distribution and dynamic of the early spawning migration of twaite shad (*Alosa fallax* (Lacépède, 1803)) and allis shad (*Alosa alosa* (Linnaeus, 1758)) in the north-west of the Iberian Peninsula. *Mar. Freshw. Res.* 67(8): 1229-1240.
- Nachón, M., Breille, G., Drouineau, H., Tabouret, H., Taverny, C., Boisneau, C., Berail, S., Pécheyran, C., Claverie, F. y Daverat, F. (2020). 1980s population-specific compositions of two related anadromous shad species during the oceanic phase determined by microchemistry of archived otoliths. *Can. J. Fish. Aquat. Sci.* 77: 164-176.
- Proctor, C. H. y Thresher, R. E. (1998). Effects of specimen handling and otolith preparation on concentration of elements in fish otoliths. *Mar. Biol.*, 131: 681-694.
- Randon, M., Daverat, F., Bareille, G., Jatteau, P., Martin, J., Pécheyran, C., y Drouineau, H. (2018). Quantifying exchanges of Allis shads between river catchments by combining otolith microchemistry and abundance indices in a Bayesian model. *ICES J. Mar. Sci.* 75: 9-21.

- Sedgwick, P. (2015). A comparison of parametric and non-parametric statistical tests. *BMJ*, 350: h2053.
- Smoliński, S., Schade, F.M. y Berg, F. (2020). Assessing the performance of statistical classifiers to discriminate fish stocks using Fourier analysis of otolith shape. *Can. J. Fish. Aquat. Sci.*, 77: 674–683.
- Tabouret, H. (2009). Recherche des marqueurs d'exposition aux contaminants et de fréquentation des habitats chez l'anguille *Anguilla anguilla* de l'estuaire de l'Adour: de la réponse moléculaire à la microchimie de l'otolithe. Thèse de Doctorat, Université de Pau et des Pays de l'Adour, Pau, France.
- Taverny, C. (1991). Contribution à la connaissance de la dynamique des populations d'aloses (*Alosa alosa* L. et *Alosa fallax* Lacépède), dans le système fluvio-estuarien de la Gironde: pêche, biologie et écologie. Étude particulière de la dévalaison et de l'impact des activités humaines. Thèse de Doctorat, Université de Bordeaux 1, Bordeaux, France.
- Taverny, C. y Elie, P. (2001). Répartition spatio-temporelle de la grande alose *Alosa alosa* (Linné, 1766) et de l'alose feinte *Alosa fallax* (Lacépède, 1803) dans le Golfe de Gascogne. *Bull. Fr. Pêche Piscic.*, 362/363: 803-821.
- Walther, B. D. y Thorrold, S. R. (2006). Water, not food, contributes the majority of strontium and barium deposited in the otoliths of a marine fish. *Mar. Ecol. Prog. Ser.*, 311: 125-130.
- Walther, B. D., Thorrold, S. R. y Olney, J. E. (2008). Geochemical signatures in otoliths record natal origins of American Shad. *T. Am. Fish. Soc.*, 137: 57-69.
- Wells, B., Rieman, B. E., Clayton, J. L., Horan, D. L., y Jones C. M. (2003). Relationships between Water, Otolith, and Scale Chemistries of Westslope Cutthroat Trout from the Coeur d'Alene River, Idaho: The Potential Application of Hard-Part Chemistry to Describe Movements in Freshwater. *Trans. Am. Fish. Soc.*, 132: 409-424.
- Zhu, J., Zou, H., Rosset, S. y Hasti, T. (2009). Multi-class AdaBoost. *Stat. Interface*, 2: 349-360.

Apéndices

Apéndice de figuras

Figura 1: Ejemplares adultos de <i>A. alosa</i> arriba y <i>A. fallax</i> abajo. Fuente: Modificado de Nachón (2017).	16
Figura 2: Esquema ilustrativo del material analizado para la realización del estudio. Los datos contorneados de rojo fueron empleados como información auxiliar relevante en el proceso de estimación de diversos parámetros.	21
Figura 3: Área de estudio considerada para la elaboración del proyecto correspondiente a la región noroeste de la península ibérica. Los ríos analizados aparecen resaltados en diferentes coloraciones. Fuente: elaboración propia.	22
Figura 4: Concentraciones de bario (Ba), estroncio (Sr) y Sr87/Sr86 (Iso) correspondientes a la microquímica de los otolitos de los juveniles pertenecientes a cada uno de los ríos estudiados.	28
Figura 5: Matriz de distancias entre la composición microquímica de los diferentes ríos contemplados en el estudio. La distancia fue calculada en función a las concentraciones de estroncio (Sr), bario (Ba) y el ratio isotópico de los isótopos del Sr.	30
Figura 6: Ejemplo de árbol de decisión basado en la temática del presente estudio. Fuente: elaboración propia.	33
Figura 7: Precisión o accuracy medio de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_1 = 10$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.	42
Figura 8: Precisión o accuracy de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_1 = 10$). Cada punto es una realización del proceso de validación cruzada.	43
Figura 9: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la filopatria ($n_1 = 10$).	44
Figura 10: Precisión o accuracy medio de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_2 = 30$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.	45
Figura 11: Precisión o accuracy de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_2 = 30$). Cada punto es una realización del proceso de validación cruzada.	46
Figura 12: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la filopatria ($n_2 = 30$).	47
Figura 13: Precisión o accuracy medio de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_3 = 60$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.	48
Figura 14: Precisión o accuracy de cada uno de los modelos de clasificación contemplados en el estudio de la filopatria ($n_3 = 60$). Cada punto es una realización del proceso de validación cruzada.	49
Figura 15: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la filopatria ($n_3 = 60$).	50
Figura 16: Precisión o accuracy medio de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_1 = 10$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.	52

Figura 17: Precisión o accuracy de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_1 = 10$). Cada punto es una realización del proceso de validación cruzada.	53
Figura 18: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la dispersión ($n_1 = 10$).	54
Figura 19: Precisión o accuracy medio de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_2 = 30$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.	56
Figura 20: Precisión o accuracy de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_2 = 30$). Cada punto es una realización del proceso de validación cruzada.	57
Figura 21: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la dispersión ($n_2 = 30$).	58
Figura 22: Precisión o accuracy medio de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_3 = 60$). En el interior de cada círculo rojo se muestra la precisión media obtenida mediante un proceso de validación cruzada.	59
Figura 23: Precisión o accuracy de cada uno de los modelos de clasificación contemplados en el estudio de la dispersión ($n_3 = 60$). Cada punto es una realización del proceso de validación cruzada.	60
Figura 24: Comparación de la precisión alcanzada al emplear la muestra de entrenamiento frente a la muestra de test en los diferentes modelos de clasificación para el estudio de la dispersión ($n_3 = 60$).	61
Figura 25: Clasificación de los ejemplares adultos capturados en el río Ulla (azul) y el río Miño (rojo), en sus respectivos ríos de origen.	63
Figura 26: Número de ejemplares capturados en medio fluvial y asignados a su río de origen en función de los diferentes umbrales de probabilidades mínimas de asignación a un río específico. Aquellos ejemplares que no alcancen un umbral fueron clasificados como de origen desconocido (unknown).	64
Figura 27: Clasificación en sus respectivos ríos de origen de los ejemplares adultos capturados localizaciones costeras de A Guarda, A Coruña, Malpica y Fisterra. Únicamente fueron clasificados los ejemplares cuya probabilidad mínima de asignación a un río concreto fue igual o superior al 80%.	65
Figura 28: Número de ejemplares capturados en medio marino y asignados a su río de origen en función de los diferentes umbrales de probabilidades mínimas de asignación a un río específico. Aquellos ejemplares que no alcancen un umbral fueron clasificados como de origen desconocido (unknown).	66
Figura 29: Selección de hiperparámetros óptimos del modelo stochastic gradient boosting mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la filopatria.	73
Figura 30: Selección de hiperparámetros óptimos del modelo stochastic gradient boosting mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la filopatria.	74
Figura 31: Selección de hiperparámetros óptimos del modelo stochastic gradient boosting mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la filopatria.	75
Figura 32: Selección de hiperparámetros óptimos del modelo stochastic gradient boosting mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la dispersión en fase marina.	76

Figura 33: Selección de hiperparámetros óptimos del modelo stochastic gradient boosting mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la dispersión en fase marina.	77
Figura 34: Selección de hiperparámetros óptimos del modelo stochastic gradient boosting mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la dispersión en fase marina.	78
Figura 35: Selección de hiperparámetros óptimos del modelo random forest mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la filopatria.	79
Figura 36: Selección de hiperparámetros óptimos del modelo random forest mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la filopatria.	80
Figura 37: Selección de hiperparámetros óptimos del modelo random forest mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la filopatria.	81
Figura 38: Selección de hiperparámetros óptimos del modelo random forest mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 10 ejemplares por río para el estudio de la dispersión.	82
Figura 39: Selección de hiperparámetros óptimos del modelo random forest mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 30 ejemplares por río para el estudio de la filopatria.	83
Figura 40: Selección de hiperparámetros óptimos del modelo random forest mediante un proceso iterativo de validación cruzada utilizando un tamaño muestral de 60 ejemplares por río para el estudio de la filopatria.	84

Apéndice de tablas

Tabla 1: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la filopatria ($n_1 = 10$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.	44
Tabla 2: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la filopatria ($n_2 = 30$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.	47
Tabla 3: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la filopatria ($n_3 = 60$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.	51
Tabla 4: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la dispersión ($n_1 = 10$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.	55
Tabla 5: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la dispersión ($n_2 = 30$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.	58
Tabla 6: Diferencias significativas entre los diferentes modelos de clasificación para el estudio de la dispersión ($n_3 = 60$) al aplicar la prueba de Mann-Whitney-Wilcoxon para muestras pareadas. p-valores menores al nivel de significación $\alpha=0.05$ indican la existencia de diferencias significativas.	61
Tabla 7: Número de ejemplares de diferentes orígenes capturados en los ríos Ulla y Miño.	63
Tabla 8: Número de ejemplares capturados en el medio fluvial asignados a su río de origen en función a los diferentes umbrales de probabilidad mínima de asignación.	64
Tabla 9: Número de ejemplares de diferentes orígenes capturados en diferentes puntos de la costa de Galicia.	65
Tabla 10: Número de ejemplares capturados en el medio marino asignados a su río de origen en función a los diferentes umbrales de probabilidad mínima de asignación. Unk. Indica origen desconocido (unknown).	66