UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Unravelling the predictive factors of spatial extent variation in European coastal teleost fishes

Matilde Alves Correia

Mestrado em Bioestatística

Trabalho de Projeto orientado por:

Doutor David Abecasis

Professor Doutor Tiago Marques

2024

Acknowledgements

I would like to express my gratitude to Dr. David Abecasis, one of my supervisors, for giving me the opportunity to develop my thesis with telemetry data, providing me new insights on this area and allowing me to work in the field of environmental biology. I would also like to express my gratitude to Professor Dr. Tiago Marques, my other supervisor, for placing a vote of confidence in me and having accepted me as his student. To both, thank you for your guidance and support, for always being willing to help me and teach me new things, but above all for always being present.

I also would like to thank Miguel Gandra, without whom this project would not have been possible, or at least it would have been harder, for his work in data processing, for having created the code to estimate the KUDs and for explaining me every step of it.

I am specially grateful to all the researchers that collected and shared data so that this project could exist.

I am also grateful to all the people who took the time to listen, help and motivate me during this journey, in particular, my master's colleagues Miguel Martins, João Louro, Miriam Dora, Vítor Palma and Bruno Martins, POrTAS and 3 litros de ucal.

To my dearest friend, Daniela Leal, whom I couldn't thank enough, thank you for always being here since day one and for growing with me throughout this challenging journey that is life. I love you.

Por fim, um agradecimento especial e muito grande aos meus pais, por sempre me terem apoiado, incentivado e feito de tudo para que eu pudesse seguir os meus sonhos e fazer aquilo que gosto.

Resumo

A ecologia do movimento é um campo bem estudado que dá especial atenção aos padrões de deslocação dos seres vivos e aos fatores e mecanismos que os influenciam. Os animais precisam de se mover para satisfazer necessidades básicas fundamentais à sua sobrevivência, como procurar alimento, reproduzir-se e evitar predadores. Geralmente, os animais não se distribuem aleatoriamente no espaço [Burt, 1943], preferindo locais onde sabem que podem encontrar segurança e recursos. Esse comportamento resulta na criação de padrões de deslocação e distribuição, uma vez que os animais tendem a habitar áreas que oferecem melhores condições.

Compreender como os animais se distribuem e os fatores que influenciam essa distribuição é crucial para implementar medidas eficazes de gestão e conservação das espécies. Para além disso, essa compreensão é essencial para prever alterações populacionais decorrentes de mudanças ambientais, como alterações climáticas, catástrofes naturais ou até mesmo intervenções humanas.

Até ao momento, muitos estudos foram realizados com o intuito de desvendar os fatores por detrás do movimento das espécies. Um dos fatores que se sabe ter influência na utilização do espaço pelos indivíduos é a massa corporal, havendo evidências de que esta componente escala linearmente com a área utilizada [Lindstedt et al., 1986; Turner et al., 1969; Udyawer et al., 2023] e que indivíduos mais pesados tendem a ocupar maiores áreas. Outro fator que parece ter impacto na distribuição das espécies é o nível trófico. O nível trófico limita o movimento dos indivíduos tanto a nível da procura de alimento como de necessidades de abrigo [Nash et al., 2015]. Isto porque, por um lado, os indivíduos na base da teia alimentar geralmente têm menores necessidades energéticas e acesso a alimentos abundantes e amplamente distribuídos, como erva ou algas marinhas, o que reduz a necessidade de viagens extensas. Por outro lado, esses indivíduos podem adotar estratégias de vida sedentária ou de baixa mobilidade, pois não apenas têm recursos abundantes nas proximidades, como também evitam expor-se a predadores permanecendo em locais protegidos. O habitat é outra característica que desempenha um papel crucial no movimento dos indivíduos. O habitat pode ser caracterizado pela sua diversidade e complexidade. Habitats mais complexos, como recifes de corais, oferecem oportunidades abundantes para abrigo e alimentação, resultando numa menor necessidade de percorrer grandes extensões. Contrariamente, em habitats menos estruturados e menos complexos, como o mar aberto, espera-se que os animais cubram áreas maiores para encontrar os recursos adequados aos seus requerimentos [Simpfendorfer, 2012]. Existem ainda fatores adicionais ou características biológicas que podem impactar a utilização do espaço, como a migração, o período de desova ou a importância comercial. No entanto, estes traços ainda não se encontram bem estudados e há informações limitadas sobre a sua relação com a distribuição.

Apesar dos esforços significativos para tentar compreender o comportamento e os padrões de distribuição das espécies, a investigação em ecologia do movimento terrestre é muito mais abundante do que a aquática [Udyawer et al., 2023; Nash et al., 2015]. Isto pode ser justificado pelo facto de apenas a partir do último século se terem desenvolvido técnicas e instrumentos capazes de rastrear e recolher dados de espécies aquáticas de maneira fácil e eficaz, como é o caso da telemetria acústica. Esta técnica utiliza em simultâneo dois tipos de aparelhos: um transmissor que é fixado ao animal e que emite sinais sonoros e um recetor que capta o sinal emitido pelo transmissor quando este se encontra dentro da área de deteção do recetor. Devido a estes avanços tecnológicos e metodológicos, a biotelemetria tem vindo a crescer cada vez mais e já existem alguns estudos focados em espécies marinhas, incluindo peixes. No entanto, existe uma lacuna na investigação da distribuição dos peixes. Até à data, é possível encontrar estudos de distribuição e de padrões espaciais de diversas espécies de peixes, porém são escassos os que reúnam dados de várias espécies e levem a cabo uma investigação multiespécies por forma a encontrar padrões comuns de distribuição de um grupo mais diversificado.

Neste projeto, o objetivo foi preencher essa lacuna. Foram analisadas oito características biológicas, ecológicas e socio-económicas – comprimento, massa corporal, vulnerabilidade, longevidade, nível trófico, habitat, migração e importância comercial – e a sua influência na área de utilização, tanto no *home range* como na *core area*, de 850 indivíduos de 30 espécies de peixes teleósteos. Os dados são provenientes de vários estudos realizados ao longo da costa da Europa e foram recolhidos através de telemetria acústica.

Para tal, foram utilizados modelos lineares generalizados mistos, utilizando a família gama e função de ligação logarítmica, e considerando o indivíduo e a espécie como efeitos aleatórios. Primeiro procedeu-se à modelação do *home range* e *core area* com cada uma das variáveis explicativas separadamente. Esta análise considera-se importante do ponto de vista ecológico, pois permite prever a área de utilização dos peixes caso apenas se tenha acesso a uma das características estudadas. Neste sentido, a variabilidade do *home range* e *core area* parece ser explicada pelo nível trófico, habitat (fator de três níveis: bentopelágico, demersal e pelágico-nerítico) e migração (fator de dois níveis: migratório e não migratório). Nomeadamente, no geral a área de utilização parece aumentar com o nível trófico e ser maior para indivíduos pelágico-neríticos e que fazem migrações. Posteriormente, os modelos foram ajustados com mais do que uma variável explicativa, atingindo-se um melhor desempenho dos modelos. O melhor modelo capaz de explicar a variabilidade do *home range* incluiu o habitat e a importância comercial (fator de 3 níveis: elevada, média e baixa) e o melhor modelo capaz de explicar a variabilidade da *core area* inclui o comprimento e o habitat.

O período de desova foi também testado, mas, neste caso, entre indivíduos da mesma espécie, por forma a observar a existência de relações significativas entre o período de desova e a àrea de utilização de cada espécie. Das 24 espécies analisadas, 16 mostraram ter relações significativas entre o período de desova e a área de utilização (tanto o *home range*, como a *core area*). Destas 16 espécies, a maioria (11 espécies) mostrou evidências de que o *home range* e a *core area* são maiores quando os indivíduos se encontram em período de desova, enquanto uma minoria (5 espécies) mostrou evidências para se afirmar o contrário, que o *home range* e a *core area* são maiores quando os indivíduos se encontram fora do período de desova. O facto de nem todas as espécies seguirem o mesmo padrão de utilização de espaço

no que toca ao período de desova é importante de se ter em conta em estudos de impacto ambiental para garantir o bem estar e manutenção de cada uma das espécies.

Este projeto é o primeiro passo na vertente de investigação multiespécies de peixes teleósteos com o propósito de compreender os padrões de movimentação. Compreender como um grupo se distribui e quais são as suas exigências e os fatores de preferência permite-nos não só obter conhecimentos valiosos para a gestão e aplicação de medidas eficientes, mas também extrapolar informação para espécies que possam ser mais difíceis de observar e para as quais não se tenha informação para desenvolver um estudo.

Até agora, a investigação científica tem privilegiado o estudo de espécies individuais, deixando em segundo plano os estudos multiespécies. Esta abordagem singular tem as suas vantagens, permitindo uma análise detalhada e profunda das características e comportamentos de uma única espécie. No entanto, esse foco pode limitar a compreensão das dinâmicas ecológicas mais amplas que ocorrem em ambientes naturais, onde múltiplas espécies interagem e coexistem.

A realização de mais estudos multiespécies é essencial para validar e expandir as tendências e evidências observadas em projetos pioneiros, como o presente estudo. Um aumento no número e na diversidade de estudos permitirá uma melhor generalização dos resultados, ajudando a confirmar se os padrões observados são consistentes em diferentes contextos e regiões. Além disso, uma maior quantidade de dados permitirá a utilização de métodos estatísticos mais robustos e a construção de modelos preditivos mais precisos, que podem ser aplicados na gestão de recursos naturais e na conservação da biodiversidade.

Para tal, como já acontece em várias áreas científicas, a colaboração entre investigadores de diferentes instituições e áreas torna-se essencial, de forma a que haja uma maior partilha de dados e metodologias, por forma a não só reduzir os custos da investigação como também acelerar o seu progresso. Este esforço coletivo contribui para a ciência básica, proporcionando novas ideias e informações, mas também tem aplicações práticas na conservação e gestão sustentável dos ecossistemas aquáticos.

Palavras Chave: Telemetria Acústica, Estudo Multiespécies, Peixes Teleósteos, Distribuição de Utilização, Modelos Lineares Generalizados Mistos.

Abstract

Movement ecology is a well-studied field that has provided significant insights into the factors influencing fish distribution. However, most studies focus on single species, with limited attention given to multi-species approaches that investigate distribution patterns applicable to a broader range of fish. This project analysed the influence of eight traits - length, body mass, vulnerability, longevity, trophic level, habitat, migration, and commercial importance - on the utilisation area of 874 individuals from 30 different teleost fish species. These data were collected through acoustic telemetry from various studies conducted along the European coast. Using generalized linear mixed-effects models, we found evidence suggesting that the variability in utilisation area is well-explained by length, trophic level, habitat, and migration alone, while controlling for random effects such as individual and species. Overall, the utilisation area appears to increase with trophic level and is larger for pelagic and migratory individuals. However, models that included multiple traits showed improved explanatory power, with the best model for home range variability incorporating habitat and commercial importance. This model revealed that home range size was larger for pelagic and commercially important species. Furthermore, the best model for explaining core area variability included length and habitat, demonstrating that core area increases with body length and is larger for pelagic species. We also investigated the relation between spawning season and utilisation area within species. Of the 24 species analysed, 16 exhibited significant associations between spawning season and changes in home range and core area. Among these, 11 species expanded their utilisation area during the spawning season, while 5 reduced their area of use. Spawning season emerged as a key factor influencing utilisation area, with spawning individuals generally occupying larger areas. This project represents an initial step in multi-species research and has revealed important key movement patterns. In the future, further studies are needed to confirm the patterns observed, and to allow for more detailed and specific analyses of the influence of spawning season.

Keywords: Telemetry, Multi-Species Study, Teleost Fish, Utilisation Distribution, Generalised Linear Mixed-Effects Models.

Index

List of Figures

List of Tables

List of Abbreviations

- **AIC** - Akaike Information Criterion
- **COA** - Centre Of Activity
- **ETN** - European Tracking Network
- **GAM** - Generalised Additive Model
- **GAMM** - Generalised Additive Mixed-Effects Model
- **GLM** - Generalised Linear Model
- **GLMM** - Generalised Linear Mixed-Effects Model
- **ID** - Identification
- **KDE** - Kernel Density Estimation
- **KUD** - Kernel Utilization Distribution
- **LRT** - Likelihood Ratio Test
- **LM** - Linear Model
- **LMM** - Linear Mixed-Effects Model
- **MCP** - Minimum Convex Polygon
- **MLE** - Maximum Likelihood Estimation
- **MPA** - Marine Protected Area
- **Q-Q plot** - Quantile-Quantile plot
- **TMB** - Template Model Builder
- **UD** - Utilisation Distribution
- **VIF** - Variance Inflation Factor

Chapter 1

Introduction

This chapter presents the introduction. Here, the motivation behind undertaking this project is presented by summarising the current state of the art and emphasising the gaps or issues in current knowledge that warrant investigation. The objectives of the study are also defined, delineating its specific aims and goals and discussing what broader implications the findings may have. Additionally, the structure of the project document is outlined.

1.1 Motivation

Animal movement, defined as a change in an individual's spatial position over time, is fundamental to life [Nathan et al., 2008]. Animals need to move to fulfil their needs and carry out essential activities for survival, such as searching for food and shelter, reproducing and parenting, and escaping predators. Wild animals usually do not roam randomly [Burt, 1943]. They distribute themselves in areas where they know they can find security and resources. This behaviour leads to space use patterns, as animals tend to inhabit areas that offer better conditions.

Movement ecology is a scientific field dedicated to studying the movement of individuals and the mechanisms and factors underlying these movements. Understanding the spatial use and distribution patterns of species is crucial, as it helps predict the health of populations under changing climate conditions and enables the incorporation of effective measures for the management and conservation of threatened and endangered species. One such measure is the designation of protected areas, as home range size is one of the factors considered in the designation of protected areas. It is therefore crucial to understand how the home range size changes according to certain traits and the events underlying the typical movements of species, making them prefer certain places over others and triggering the use of larger or smaller areas to implement and delineate more effective boundaries. Nevertheless, caution is needed when designing reserves and protected areas, as they must meet the spatial needs of all target species to ensure the maintenance of populations.

Many studies have been carried out to unveil the drivers of animal space use patterns. One of the

factors found to influence the space used by individuals is body mass. There is evidence that body mass scales linearly with home range in terrestrial mammals [Lindstedt et al., 1986; Turner et al., 1969]. Generally, larger animals tend to move over larger spatial extents than smaller animals to satisfy their greater energy requirements [Nash et al., 2015]. Lower metabolic requirements can be achieved within smaller areas. According to Hendriks [2007], geographic ranges of species increase with body mass, with fish generally occupying smaller areas than mammals and birds. Additionally, carnivorous individuals tend to occupy larger areas than herbivorous ones Lindstedt et al. [1986], which leads us to another factor that may underlie the distribution patterns of animals: trophic level. The trophic level represents the position an individual occupies in a food web. Typically, predators are found at the top of the food web, while herbivores and primary producers are found at the bottom. The trophic level limits the movement of individuals both in terms of food search and shelter needs [Nash et al., 2015]. On the one hand, individuals at the bottom of the food web usually have lower energy requirements and access to abundant and widely distributed food, such as grass or seaweed (depending on whether they are marine or terrestrial organisms), which reduces the need for extensive travel. On the other hand, these individuals may adopt sedentary or low-mobility life strategies, as they not only have ample resources nearby, but also avoid exposing themselves to predators by staying in sheltered locations. Afonso et al. [2022] found that trophic level seemed to influence the maximum diving depth of elasmobranchs, which tended to increase with the trophic level. Another factor that plays a crucial role in the movement of individuals is habitat. Habitat is characterised by its diversity and complexity. More complex habitats, such as coral reefs, provide abundant opportunities for shelter and feeding, resulting in smaller occupied areas. Conversely, in less structured, less complex, habitats, like the open ocean, animals are expected to cover larger areas to find suitable resources [Simpfendorfer, 2012]. According to Topping et al. [2005], there is evidence that habitat shape and type have a significant influence on the home range size of *Semicossyphus pulcher*. There are additional factors or biological characteristics that seem to be related to space use, such as migration, spawning season, or commercial importance. However, these traits are not well studied, and there is limited information about their relationship with home range. Concerning migration, it is logically expected to observe differences in space use between individuals engaging in large-scale migrations and those that are more sedentary and attached to a specific location [Kropil et al., 2015]. Furthermore, some species undertake migrations during the spawning season to areas favourable for the birth and growth of offspring, potentially resulting in an increased occupancy area [Afonso et al., 2008a]. The relationship between movement patterns and genetic structure is also significant and something to consider in the light of movement ecology studies. Theoretically, the higher the capability of dispersion, the lower the inter-population genetic structure of a species [Gandra et al., 2021]. In other words, if individuals of a species have a greater ability to move and spread over large areas, there will be more mixing between different populations of that species, resulting in less genetic differentiation between populations because genes are more widely shared across the range of the species. Gandra et al. [2021] discovered a statistically significant link between commercial importance and genetic differentiation. Their findings indicated that non-commercial species showed greater genetic differentiation compared to species of minor commercial,

commercial, or high commercial importance. These findings suggest that highly commercial species tend to disperse more than non-commercial species, however this has never been properly assessed.

Despite considerable efforts to understand species behaviour and movement patterns, research in movement ecology has largely concentrated on terrestrial taxa [Udyawer et al., 2023; Nash et al., 2015]. The factors mentioned above have been well studied in terrestrial mammals [Lindstedt et al., 1986; Kropil et al., 2015; Kie et al., 2002], but they are far less clear for marine species. This leaves a notable gap in our understanding of aquatic species, emphasising the need for more in-depth studies on the movement patterns of marine and freshwater organisms to better comprehend their spatial dynamics and conservation requirements. Fortunately, recent advances in technology have allowed the development of new techniques to study these groups more easily and efficiently. Biotelemetry is an example of this, as it is a tracking tool that allows the detection, location and tracking of marine individuals [Crossin et al., 2017], making it possible to carry out fish movement studies. In fact, many marine researchers have embraced this tracking tool and created collaborative networks to share their data, leading to a growing sense of unity and collaboration within the field. This collective effort not only fosters a stronger research community, but also significantly increases the quality and depth of studies in movement ecology, allowing for much larger datasets, not only for more reliable results, but also for more subjects to be studied in different regions of the world. As a result, our understanding of the behaviour and distribution patterns of marine species is becoming increasingly comprehensive and sophisticated. With these extensive datasets, researchers can address a wide range of questions and conservation challenges, particularly those aimed at comparing and investigate species behaviour across different global regions, between various reserves, and within and outside protected areas. Additionally, they allow for the identification of factors driving these differences, providing deeper insights into the dynamics of marine ecosystems.

1.2 Objectives

In this project, we aim to study and evaluate some biological, ecological and socio-economic factors that explain the distribution patterns of teleost fish. Teleost fish are one of the most significant groups within the fish community, comprising approximately 95% of all existing fish species and hence naturally serving as the primary target for commercial fishing. Studying these animals is essential to ensure their sustainability, promoting the conservation of populations and the health of aquatic ecosystems. Many studies have focused on understanding the biology, behaviour, and distribution of certain teleost species. However, there is a lack of research aimed at determining whether the movement patterns observed in specific species can be generalised to a broader and more diverse group. This can be crucial. If we can identify common patterns among various teleost species, we can infer that species for which we lack detailed information may exhibit similar behaviours and movement patterns. This could help their management and conservation where such potential relevant knowledge is missing. To address this gap, this project serves as a pioneer study, bringing together data from 30 teleost species collected across 15 different studies conducted in Europe, with the intent to assess the variability of the distribution area

used by teleost fish based on biological, ecological and socio-economic traits such as length, body mass, longevity, vulnerability, trophic level, habitat and migration strategies, as well as commercial importance. We also aim to understand if there is any relation between utilisation area and spawning season both across and within species, since different species might have distinct movement strategies during spawning season. The findings will contribute to understanding whether there are common movement patterns among teleost species and how biological traits influence the activity space they use. This will significantly advance our knowledge of teleost fish ecology and movement patterns, providing valuable insights for management and conservation practices. It will enable the development of more effective strategies for managing teleost fish populations, especially within Europe, where the data comes from. Additionally, it lays a solid foundation for future studies on movement ecology, highlighting the importance of multi-species research approaches.

1.3 Project Report Structure

This report is divided into two main chapters, Chapter 2 and Chapter 3.

Chapter 2 provides a theoretical framework, presenting and describing the foundational theoretical concepts of the study. This chapter aims to offer a comprehensive understanding of the origin of telemetry data and its collection methods, as well as the statistical analysis implemented to answer the questions. It is divided into four subchapters, covering the topics of (1) telemetry, (2) utilisation distribution, (3) models and (4) model comparison methods.

Chapter 3 focuses on our study itself and its development, and it is organised into six subchapters. Subchapter 1 introduces the data and gives a brief description of the variables used. Subchapter 2 details the statistical methods employed in data analysis. Subchapter 3 presents the results of data analysis. Subchapter 4 discusses the results obtained. Subchapter 5 presents what I have learnt from this project. And finally, Subchapter 6 summarises the main conclusions drawn from the project.

Chapter 2

Theoretical Framework

This chapter presents an overview of the relevant concepts and methodologies employed in this study, from the methods used in data collection to the ones used in statistical analysis. The aim is to provide foundational knowledge to the study, enabling readers with diverse backgrounds to gain an understanding of its nature and development.

2.1 Telemetry

Telemetry is a form of wireless communication that measures and transmits data from a remote source to a receiving station. Currently, numerous fields benefit from this technology, from drones to a variety of applications using remote sensors. One of them is biotelemetry.

Biotelemetry is a term used in ecology to refer to a range of electronic tracking tools that utilise telemetry to detect and monitor animals [Crossin et al., 2017]. This technique requires two types of devices: the transmitter, also known as tag, and the receiver. Depending on the transmitter and the nature of the emitted signal, telemetry can be classified in three distinct types: acoustic, radio and satellite telemetry [Abecasis et al., 2018]. All these techniques are similar in the way that to function they need a transmitter to be attached to the animal and a receiver to capture the signal. Radio telemetry transmits signals at high frequencies between 20 and 300 MHz and uses wire antennas to capture the signal. However, these radio signals do not propagate in salt water [Hussey et al., 2015]. Satellite telemetry transmits signals at a UHF frequency of 401.650 MHz and the signal is captured by satellites of the ARGOS system [Reine, 2005]. Acoustic telemetry transmits signals at low frequencies between 20 and 300 kHz, and, because of that, this is the most suitable type of telemetry for use in both marine environments, since sounds, and hence the signals, are not easily absorbed by water, as happens with radio and satellite telemetry.

Here, the primary focus will be on acoustic telemetry. Acoustic telemetry is, nowadays, the most popular technique among biologists to study and monitor movement, behaviour and distribution patterns of marine animals. There has been a significant advancement in this sector over the last 30 years [Abecasis et al., 2018]. The number of studies increased, and this led to improvements in devices and

data collection, and a set of numerous new discoveries have emerged in the light of science. There are a lot of different acoustic transmitters and the choice of which to use must be properly considered, taking into account the animals to be tagged, battery longevity and detection range. Weight of transmitters can range between 0.3 g and 40 g. This wide range is important because it enables the study of many different species, particularly small species and juveniles. For ethical reasons, it is common practice, among scientists, to use transmitters that weigh no more than 2% of the individual's total body weight [Winter, 1983]. Transmitters can be implanted internally or externally in an animal's body. They can be surgically implanted in the peritoneal cavity, inserted in the stomach through the mouth or be externally attached with adhesives, glues or other instruments. Transmitters can also be differentiated based on the nature of the signal they emit, where they can be either "pulsed" or "coded" [Reine, 2005]. Pulsed or standard transmitters emit simple pulses at a selected rate or in response to certain events. A simple pulse is best described as a single and brief fluctuation in the amplitude, frequency or other property of the signal, like a "beep" [Allen, nd]. Due to their simplicity, they are efficient in presence/absence studies that do not require extensive information collection, i.e. where individual identification is not required. On the other hand, coded transmitters are more complex and emit sequences of pulses using the same frequency but random emission times between each pulse to create a unique code that distinguishes and identifies all transmitters. These transmitters are particularly useful in studies that require simultaneous monitoring of several individuals or where additional information, individual specific, needs to be collected.

Data typically collected by coded acoustic telemetry consist of the ID of the tagged individual, the ID of the receiver and the time and date stamp. The ID of the receiver can be associated to a location based on metadata. Tags can also be equipped with sensors to transmit environmental data, such as depth or water temperature, or individual behavioural or physiological data, such as acceleration, heart rate, stomach pH, depth and more [Crossin et al., 2017; Abecasis, 2008].

For telemetry to work, a receiver is also needed. The instrument used as a receiver in acoustic telemetry is an hydrophone, which is capable of capturing the signals from the tags underwater. The signal reception can be done actively, in real time, or passively, with receiver stations installed. Therefore, two types of acoustic telemetry exist: active and passive. In active acoustic telemetry, the hydrophone goes on a vessel, continuously following and locating the tags. In passive acoustic telemetry there are receiver stations, which are fixed and installed in strategic places, usually in the bottom of the ocean and without too many noisy elements around. Passive acoustic telemetry offers several advantages over active acoustic telemetry, primarily because it does not require continuous human monitoring, thus allowing for extended monitoring periods. Additionally, it enables the simultaneous monitoring of multiple individuals. Until recently, there was a limitation in the ability to precisely locate individuals. However, more recently, software such as Innovasea's Fathom Position [INNOVASEA, 2021] and the YAPS R package [Baktoft et al., 2017] have been developed to precisely locate individuals based on the time of signal arrival at multiple receivers, sometimes achieving an accuracy of less than 5 metres [Whoriskey et al., 2022]. Despite these advancements, there are still some disadvantages, such as the challenge of tracking individuals if they move outside the detection range, and the need for multiple detections at different

sensors within a small time window, which requires dense arrays to be efficient.

Detection of animals may depend on many factors, but arguably the most important and one that has been recently a matter of discussion in the literature is the detection range of the receivers. It is fundamental to take detection range into account when doing studies about movement and distribution of species, because lack of such knowledge can cause errors in data analysis and the interpretation of results, leading to wrong conclusions [Kessel et al., 2014]. It is therefore important to understand how the detection range is defined and how it can be applied in subsequent analyses. Detection range can be seen as the effective distance at which a receiver is able to detect a transmitter [Kessel et al., 2014]. This distance is influenced by the location of the receiver and the surrounding environment [Goossens et al., 2022]. Open water locations have larger detection ranges because there are fewer obstacles to interfere with the signal transmission. In contrast, locations with many surrounding obstacles, like kelp or rocks, are more likely to have a reduced detection range because, as the signal travels from the transmitter to the receiver, there is a loss of energy due to refraction and attenuation in signal propagation [Kessel et al., 2014]. Properties of the water body, such as salinity, turbidity and temperature, will also influence the propagation of the signal, with the water being more or less conductive depending on their combination at a given site [Goossens et al., 2022]. Additionally, anthropomorphic and environmental noise can also reduce the detection range, by overlapping with the sound of the transmitter [Reubens et al., 2019; Huveneers et al., 2016; Abecasis, 2008]. On the other hand, the maximum distance can also be influenced by the efficiency and range of both the transmitter and the receiver [Hellström et al., 2022]. Powerful transmitters have the capacity to travel longer distances and receivers with better quality have the capacity of capture signals further away [VEMCO, 2014]. Although environmental conditions vary over space and time and, therefore, there is no standard definition of detection range, it is important to include it in statistical analyses of studies using presence/absence data to account for the probability of detecting individuals.

Beyond the detection range, there is another aspect that can occur and must be taken into account when analysing the data, namely false positive detections. This happens when a tagged animal's transmission collides with ambient noise or with another animal's transmission, resulting in an unknown code ID or in one that already exists in the system [Crossin et al., 2017]. Such events are common and can cause false detections that could lead to inaccurate conclusions if not recognized. One could see two detections of the same fish far apart as evidence for extremely fast displacement, when in reality it might simply be a false positive in one of the locations. Over the years, acoustic telemetry has become a vital tool for studying aquatic animals, and researchers have been working to improve this technique and make it more accurate. Network organisational structures have been created, such as the Ocean Tracking Network (OTN), the Integrated Tracking of Aquatic Animals across the Gulf of Mexico (iTAG), the Integrated Marine Observing System (IMOS), the Aquatic Tracking Array Platform (ATAP) and the European Tracking Network (ETN). These networks serve as repositories where researchers can deposit their aquatic biotelemetry data [Abecasis et al., 2018], making it available for others worldwide. This fosters extensive collaboration among aquatic tracking researchers, enabling them to share infrastructure

and expand monitoring areas for species with vast migration patterns. Furthermore, these networks facilitate interactions within the biotelemetry community, allowing researchers to connect and familiarise themselves with each other's work.

2.2 Utilisation Distribution

With technology advancements over the last 50 years, the study and understanding of the distribution patterns of species became an easier task, and a new scientific concept emerged: the Utilisation Distribution. Utilisation Distribution (UD) is a concept known in ecology to describe the space used by individuals in their habitat over a given period of time [Winkle, 1975]. In statistical terms, it represents the probability of finding an animal in a given area and over a given period of time. This metric is usually generated using animal location data, which is then processed to obtain the total area observed and the probability of a specific area. Among the areas most studied and of most interest to biologists and ecologists are those referred to as the home range and the core area [Cholaquidis et al., 2023]. Home range is generally defined as the area occupied by an individual in its normal activities of food gathering, mating and parental care, being that occasional departures from this area should not be considered [Burt, 1943]. In this way, and statistically speaking, the home range has been associated with the area occupied by the animals 95% of the time. Regarding the core area, it refers to the area where the animal spends 50% of its time [Laver and Kelly, 2008], which usually translates into a smaller, more central area of activity.

There are various methods for estimating the UD. Among them, one of the best known and most widely used are Kernel methods [Laver and Kelly, 2008], usually called Kernel Utilisation Distribution (KUD) or Kernel Density Estimation (KDE). These methods use kernel functions that are non-negative, real-valued and integrable functions, which satisfy the following two requirements [NUMXL, 2016]:

$$\int_{-\infty}^{+\infty} k(u)\,du = 1, \tag{2.1}$$

$$k(u) = k(-u) \tag{2.2}$$

where $k(u)$ represents the kernel function. The equation (2.1) indicates the normality requirement, ensuring that the total area under the curve of the kernel function is equal to 1. The equation (2.2) indicates that $k(u)$ is an even function, meaning that the kernel is symmetrical about the y axis, i.e. the function has the same value for u and $-u$.

KUDs work as follows: Firstly, the animal locations are distributed on a plane. A kernel function is applied to each point, which acts as a smoothing function. Finally, all the kernel functions are added together. Below is the formula for the basic KUD function:

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right) \tag{2.3}$$

where $\hat{f}(x)$ is the estimated density at point x, representing the probability density function, N is the total number of observations, h is the bandwidth or smoothing parameter, $k(.)$ is the kernel function, which assigns weights based on the value of $\frac{x-x_i}{h}$, being x the point at which the probability density is being estimated and x_i a data point from the sample.

It should be noted that with the development of this technique, new generalisations and adaptations have been made to the formula [Worton, 1975]. The final result is a probability density function, where the peaks of the function correspond to the places where there is the greatest probability of finding an animal. Usually, the 50^{th} and 95^{th} percentiles are used as proxies for core area and home range, respectively [Laver and Kelly, 2008]. The UD can also be visualised in the form of contour maps, heat maps or three-dimensional surface maps, allowing for a better understanding and interpretation of the space used by individuals.

An important parameter to take into account when estimating KUD is the kernel bandwidth [Kraft et al., 2023]. The bandwidth is a smoother parameter that controls the neighbourhood size within which observed locations contribute to the density estimate at a point [Silverman, 1986]. A high bandwidth widens the distribution over each point, making distant points more influential, increasing the overall area of the area used, and making resulting surfaces smoother. In addition, they are more likely to take into account the uncertainty of the estimates and smooth out possible sampling errors, eliminating small-scale details and retaining only the most notable and important features. On the other hand, a small bandwidth provides greater detail at small scales, but tends to be more sensitive to sampling errors and could lead to severe overfitting, in the sence that noise in the data would drive the estimation procedure and patterns found not being generalisable to other animals or areas.

The bandwidth can be constant (fixed kernel) or it can vary (adaptive kernel) over space [Millspaugh et al., 2006]. The adaptive kernel is advantageous in that it allows different locations to have different smoothing, and the margins of the distributions that correspond to the most uncertain locations because there are few observations have a higher smoothing. However, based on studies of UD simulations carried out by Seaman [1996] and Silverman [1986], it was observed that the adaptive kernel showed greater bias than the fixed kernel approach, based on the relative mean squared error. Thus, the fixed kernel method is usually preferable for home range estimation [Millspaugh et al., 2006]. For both approaches, an automatic bandwidth selection method that minimises the errors associated with UD estimates is desirable [Millspaugh et al., 2006].

2.3 Models

Models are simple and generalised representations of reality. As George Box famous quote says, ”All models are wrong, but some are useful”. In particular, models are widely used to understand, explain, and predict events based on explanatory variables and their relationships. Countless mathematical models have been developed over the years, applicable across all scientific fields. The selection and application of statistical models for data analysis demands a critical and careful approach, since inappropriate choices

can lead to incorrect interpretations of the data, inaccurate predictions, and ultimately, decision-making based on wrong information.

In this project we concentrate on regression models. Most models come with a set of assumptions that, if not met, may result in biased estimates and invalid inferences [Gelman, 2006]. Models can vary in complexity and it is important to acknowledge that the complexity of the model should match the nature of the data being analysed. Overly simplistic models may fail by not capturing important nuances, while overly complex models risk interpreting noise as significant patterns [Hastie et al., 2009].

Therefore, an exploratory data analysis should be conducted before selecting a model. This preliminary analysis can provide valuable insights into the data structure, the relationships between variables, and their distributions, guiding the choice of the most appropriate model.

Furthermore, it is essential to test the chosen model's assumptions, adjust its complexity as necessary, and validate its predictions through techniques like cross-validation and splitting data into training and testing sets. These practices help ensure that the model not only fits well to the available data but also retains the ability to generalise to new data.

Some of the most popular regression models will be presented now, starting with the simplest one, the linear model (LM), before progressing to encompass some of its generalisations, namely the generalised linear model (GLM) and the generalised additive model (GAM), and further the corresponding mixed models.

2.3.1 Linear Regression Model

Linear regression models are amongst the simplest statistical models available, yet are some of the most powerful available to help researchers answer ecological questions. They are used to model the relationship between the independent variable(s) and a continuous dependent variable [Schmidt and Finan, 2018]. Linear regression can be classified as simple when there are only measurements of two variables, i.e. one dependent variable that is explained by one independent variable, and as multiple when there are measurements of more than two variables and the dependent variable is explained by the remaining variables. A linear regression is described by the following equation:

$$Y_i = \beta_0 + \sum_{p=1}^{k} \beta_p x_{pi} + \epsilon_i, \quad (i = 1, 2, \ldots, n) \tag{2.4}$$

$$\epsilon_i \quad \text{i.i.d.} \quad N(0, \sigma^2)$$

where Y_i is the dependent variable for observation i, $(\beta_0, ..., \beta_p)$ are the model parameters to be estimated, x_{pi} is the value of the independent variable p for observation i, and ϵ_i is the error associated with observation i. ϵ_i represents how much the observed value differs from the expected value according to the model, often referred as the residuals ($\epsilon_i = y_i - \hat{y}_i$); in general the estimate and the observed value will be different [Casson and Farmer, 2014].

This type of model is usually considered alongside a set of assumptions:

(1) **Linearity:** There must be a linear relationship between the dependent, y_i, and independent, x_{pi}, variables. This linearity can be expected a priori, from domain knowledge, and visually inspected using scatterplots, which should reveal a straight-line relationship rather than a curvilinear one.

(2) **Independence:** The residuals must be independent. There should be no autocorrelation between the errors, which means that the value of one error should not be able to predict the value of another error. The independence of the residuals can be analyzed by plotting the residuals the order in which the observations were obtained or by plotting them against X. The assumption is met if the points are randomly distributed along the horizontal axis, with no sign of clear patterns of dependence. The presence of positive or negative clusters suggests there is correlation. Usually the best is to enforce this assumption by selecting a random sample from the population under study. Correlation tends to appear due to some unmodelled heterogeneity in the data, e.g. when an unrecorded covariate Z influences the response Y but is unavailable.

(3) **Homoscedasticity:** The variance of the residuals must be constant across all levels of the independent variables and equal to σ^2. The homoscedasticity can be checked by plotting the residuals against the observed and/or fitted values. The assumption is met if the dispersion of the points remains constant along the horizontal axis.

(4) **Normality:** The residuals must be normally distributed with mean 0 and variance σ^2. The normality of the residuals can be assessed by using a normal probability plot (Q-Q plot) of the residuals, which plots the empirical quantiles of the residuals against the theoretical quantiles of the standardised normal distribution. The assumption is met if the points on such a plot fall close to the diagonal reference line. There are also a variety of statistical tests for normality, such as the Kolmogorov-Smirnov test, the Shapiro-Wilk test, the Jarque-Bera test, and the Anderson-Darling test [Das and Imon, 2016]. However, such tests are very sensitive, and since real data rarely follow a perfect normal distribution, it is almost impossible for the tests not to violate the normality assumption for larger sample sizes, for any usual significance level. So, it is recommended to look at the Q-Q plots to draw conclusions about the normality assumption. Additionally, this assumption is usually less relevant for large sample sizes, when the Central Limit Theorem (CLT) tends to apply. The CLT implies that, given a sufficiently large sample size, the sampling distribution of a mean will converge to a normal distribution, regardless of the shape of the population distribution. So even if this assumption is violated for large samples when observing the Q-Q plots, models may still produce valid inferential results [Schmidt and Finan, 2018]. It is therefore preferable to focus on the violations of other assumptions, rather than violations of normality assumption, when the dataset is large.

There is another characteristic to consider when the model has more than one independent variable,

and that is multicollinearity. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, meaning they contain similar information about the variance in the dependent variable. This might be a problem since correlated variables explain similar variability of the model and can lead to inflated standard errors, resulting in wider confidence intervals and reduced statistical significance for predictors that may actually be important. Multicollinearity should therefore be avoided. For that, the independent variables, x_{pi}, must be linearly independent of each other. If they are not independent, it means that they are somehow correlated, and changes in one variable will be associated with changes in another. Multicollinearity can be checked using correlation matrices or the variance inflation factor (VIF $= 1/(1 - R^2)$). For correlation matrices, R^2 values close to 1 indicate high correlation between variables and values close to 0 indicate almost complete independence between variables. The VIF measures how much the variance of a regression coefficient increase due to multicollinearity. If the VIF is equal to 1, the variables are linearly independent, and if the VIF is greater than 10, the variables are strongly linearly dependent, indicating multicollinearity.

Sometimes, some of the assumptions are not met, but there are options to circumvent this situation to allow the linear models to be applied. One of these techniques is data transformation. Data transformation is a process in which changes are made to one or more variables (Osborne, 2002), either dependent or independent variables. Some of the most commonly used transformations are logarithmic, inverse, square root [Manikandan, 2010] and Box-Cox, but there are many others. To fulfil the assumption of linearity, transformations are usually applied firstly to the independent variables [Zelmer, nd]. However, if this fails, transformations on the dependent variable may be attempted. This must be the first step before moving on to analyse other assumptions. To fulfil the normality and homoscedasticity assumptions, transformations are usually applied to the dependent variable, which allows the stability of the variance of the residuals, approximating them to a normal distribution. It is often the case that a transformation designed to overcome problems that arise when one of the assumptions is not met, simultaneously solves problems related to other assumptions [Poole and O'Farrell, 1971]. The other technique employed when the assumptions are not met correspond to extensions to the classical linear regression model. Nowadays, there are several extensions of linear models, such as polynomial regression models or splines, which can be used when there is a non-linear relationship between the dependent and independent variables, or generalised linear regression models, which can be used when the data do not follow a normal distribution. It is much more sensible to use a regression model that is sensible for the data at hand than to try to shoehorn a dataset into unrealistic assumptions via data transformation.

2.3.2 Generalised Linear Model

Generalised Linear Models (GLMs) are extensions of linear models, which is a special case of GLMs. They are a solution when the assumption of normality is not satisfied and the errors do not follow a normal distribution. This type of models still requires the independence assumption. A sort of linearity is also required, but not that between the independent and dependent variables on the response scale. GLMs

are characterised by having a random component and a systematic (or structural) component [Nelder and Wedderburn, 1972]. The random component is responsible for explaining the random variability that cannot be explained by the linear combination of the independent variables. This variability is therefore assumed to be explained by another distribution, necessarily from the exponential family, covering a range of useful distributions that include the Normal, Binomial, Bernoulli, Poisson, Exponential, Gamma, Negative Binomial and Multinomial. The general form for the density function of the exponential family is:

$$f(y_i, \theta_i, \phi) = \exp\left\{r(\phi)\left[y_i\theta_i - g(\theta_i)\right] + h(y_i, \phi)\right\}, \quad (i = 1, 2, \ldots, n) \tag{2.5}$$

where θ_i is the natural location parameter and ϕ is the scale parameter. This component describes the probability distribution of the dependent variable and, because of that, the decision of which distribution to use must be made based on the nature of the relationship between the independent and the dependent variables.

The systematic component of GLMs is responsible for linking the independent variable(s) to the dependent variable through a linear predictor. Then it consists of two main elements: the linear predictor and the link function [Myers and Montgomery, 1997]. The linear predictor is a linear combination of the explanatory variables, each weighted by its respective coefficient, represented as:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}, \quad (i = 1, 2, \ldots, n) \tag{2.6}$$

where η_i is the linear predictor associated with observation i, β_0 is the intercept, $(\beta_1, ..., \beta_k)$ are the coefficients of the k independent variables and $(x_{1i}, ..., x_{ki})$ are the observations related to each independent variable. Note that above in the linear regression case the same linear predictor was considered.

The linear predictor is related to the expected value of the dependent variable through a link function. The link function is denoted by $g(.)$ and is related with the linear predictor by the expression:

$$g(\mu_i) = \eta_i, \quad (i = 1, 2, \ldots, n) \tag{2.7}$$

where $g(.)$ is the link function, μ_i is the expected value of the dependent variable and η_i is the linear predictor. There are a variety of link functions that can be used, but they must be chosen according to the distribution selected and the nature of the dependent variable being analysed. The fundamental property of the link function is that it constrains the predicted response to be within a given admissible range of values. Some of the most used link functions are the logit, the log, the identity, the inverse, the probit and the square root. To illustrate how the link function works when applied to the dependent variable, here is the example of the logit:

$$\text{logit}(\mu_i) = \ln\left(\frac{\mu_i}{1 - \mu_i}\right) = \eta_i, \quad (i = 1, 2, \ldots, n) \tag{2.8}$$

Then, the model produced will be:

$$\mu_i = \frac{1}{1 - e^{-\eta_i}} = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})}}, \quad (i = 1, 2, \ldots, n) \tag{2.9}$$

where $logit(.)$ is the link function, μ_i is the expected value of the dependent variable for observation i, and η_i is the linear predictor. For the log function, it will be:

$$\ln(\mu_i) = \eta_i, \quad (i = 1, 2, \ldots, n) \tag{2.10}$$

And the model produced will be:

$$\mu_i = e^{\eta_i} = e^{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}}, \quad (i = 1, 2, \ldots, n) \tag{2.11}$$

where $ln(.)$ is the link function, μ_i is the expected value of the dependent variable for observation i, and η_i is the linear predictor.

For each distribution, there is a natural link function derived by setting the natural location parameter equal to the linear predictor:

$$\theta_i = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}, \quad (i = 1, 2, \ldots, n) \tag{2.12}$$

This is known as the canonical link function. It simplifies the likelihood equations, making it easier to estimate model parameters. For the Binomial distribution, the canonical function is the logit, for the Poisson distribution it is the log, for the Normal distribution it is the identity, for the Gamma distribution it is the inverse, and so on. While the canonical function is often the preferred choice, it is important to consider other link functions as well, which might be easier to interpret (the best example is the Gamma, which canonical inverse link function leads very often to errors in interpretation and even inadmissible estimates). This can help to ensure the best fit for the data being analysed.

2.3.3 Generalised Additive Model

Generalised Additive Models (GAMs) are extensions of generalised linear models [Grego, 2012], which are also special cases of GAMs. They are used when the relationship between the independent and dependent variables is complex and non-linear [Pedersen et al., 2019]. These non-linear relations are modelled using a smooth function. The smooth functions are non-parametric functions - in that the parameters involved are not interpretable, but they do have parameters - applied separately to each independent variable, allowing for the determination of non-linear relationships in a flexible way. In this way, the linear predictor present in GLMs is replaced by an additive predictor in GAMs, that contains the smooth functions, denoted by $\eta(X)$, which takes the following form:

$$\eta(X) = a + \sum_{p=1}^{k} f_p(X_p) \tag{2.13}$$

where $\eta(X)$ is the additive predictor, a is the intercept and $\sum_{p=1}^{k} f_p(X_p)$ is the sum of the smooth functions of each independent variable. The smooth functions can be estimated through smoothing scatterplots techniques. There are several scatterplot smoothers, one of them being the penalised regression splines. Smoothers require the choice of a span to operate. The span is responsible for the width to be considered around a point when smoothing. The larger the span, the smoother the curve [Yee and Mitchell, 1991]. The best span is the one that provides the best estimates in terms of the sum of squared prediction errors, on average. Given that each variable is separately smoothed, different variables can have different smoothers and spans. In addition, each independent variable has an additive effect that can be evaluated examining the function. If $f_p(X_p) = \beta_p x_p$ for all p, the GAM is a GLM. Just as in the GLM, the additive predictor is related to the expected value of the dependent variable through a link function. The link function is denoted by $g(.)$ and is related with the linear predictor by the expression:

$$g(\mu) = \eta(X) \tag{2.14}$$

where $g(.)$ is the link function and $\eta(X)$ is the additive predictor.

2.3.4 Mixed-effects Models

Mixed-effects models are used when data is collected in a hierarchical or grouped structure, inducing some dependence, i.e. correlation, between observations. These models incorporate two types of effects: the fixed effects and the random effects. The fixed effects are usually the primary focus of the research question. The random effects account for the variability within the data, induced by the correlation across observations, that cannot be explained by the fixed effects (nor the residual variation). Random effects are usually associated with specific groups or clusters in the data. The most traditional use of random effects in general is to account for multiple measurement made on the same individual, and are called individual random effects. But other possible grouping of observations might be present. For example, if the collected data belongs to different species or was collected in different sites, and these are not all sites/species of interest but can be seen as a random sample of all possible sites/species, the species or the sites may be seen as random effects, assigning different effects to each species or sites. The general formula [Werf, nd] for the mixed-effects models is the following:

$$Y_{ij} = X_{ij}\beta + Z_{ij}u_j + \epsilon_{ij} \tag{2.15}$$

where Y_{ij} represents the response variable for observation i in group j, X_{ij} represents the fixed effect predictors, β is the vector of the fixed effect coefficients, Z_{ij} represents the random effect predictors, u_j is the vector of random effect for group j, and ϵ_{ij} is the error term associated with observation i in group j.

Random effects can be combined with the different models we described above, leading to linear mixed-effects models (LMM), the generalised linear mixed-effects models (GLMM) and the generalised additive mixed-effects models (GAMM). Equation (2.15) is the formula corresponding to LMM. For the

GLMM, the right side of the formula remains the same, but in the left side the link function is added, getting:

$$g(\mathbb{E}[Y_{ij}]) = X_{ij}\beta + Z_{ij}u_j + \epsilon_{ij} \tag{2.16}$$

where $g(.)$ is the link function. For the GAMM, the link function is added to the left side of the formula and the smooth functions are added to the right side, getting:

$$g(\mathbb{E}[Y_{ij}]) = X_{ij}\beta + f(X_{ij}) + Z_{ij}u_j + \epsilon_{ij} \tag{2.17}$$

where $g(.)$ is the link function and $f(X_{ij})$ are the smooth functions.

2.4 Model Comparison Methods

In a given statistical analysis, it is common to consider different plausible models that might differ in regression type, parameters, or other aspects. To determine which model best fits the data and is most appropriate for inferences or predictions, several comparison techniques can be employed. Some key methods include the log-likelihood, the likelihood ratio test (LRT), and the Akaike Information Criterion (AIC). Other valuable methods are the Bayesian Information Criterion (BIC), deviance, cross-validation, and adjusted R^2. Although there are many methods available to evaluate model performance, our focus here will be on the first three mentioned above.

2.4.1 Log-Likelihood

The log-likelihood is a fundamental concept in statistics that measures the log-likelihood of observing a set of data given a statistical model. It can be used to respond to different problems. The log-likelihood function can be used as a model comparison tool. Given a statistical model with parameters θ and a set of observed data $x = (x_1, x_2, \ldots, x_n)$, the likelihood function $L(\theta \mid x)$ measures how likely it is to observe the data x given the parameter values θ [Patefield, 1977]. The likelihood function is then defined as:

$$L(\theta \mid x) = P(X = x \mid \theta) \tag{2.18}$$

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{2.19}$$

where X represents the random variable that generates the data x in (2.18) and $f(x_i \mid \theta)$ represents the probability density function or probability mass function of x_i given the parameter θ in (2.19).

Then, the log-likelihood function $\ell(\theta \mid x)$ is simply the natural logarithm of the likelihood function:

$$\ell(\theta \mid x) = \ln L(\theta \mid x) \tag{2.20}$$

$$\ell(\theta \mid x) = \sum_{i=1}^{n} \ln f(x_i \mid \theta) \tag{2.21}$$

Using the log-transformed likelihood is advantageous because it converts the product of probabilities (in the likelihood function) into a sum of logarithms, making calculations easier and avoiding issues with numerical precision. To compare the models, one must look to the values of the log-likelihood. A higher value indicates that the observed data is more likely under the proposed model and, consequently, indicates the best model.

The log-likelihood function is also a fundamental aspect of maximum likelihood estimation (MLE), where the goal is to find the parameter values that maximize the likelihood of the observed data, which is equivalent to maximize the log-likelihood function. To do that one must take the derivative of $\ell(\theta \mid x)$ with respect to θ, setting it to zero, and solving for θ, like this:

$$\frac{\partial \ell(\theta \mid x)}{\partial \theta} = 0 \tag{2.22}$$

This process allows statisticians to find the most probable parameter values given the observed data, making the log-likelihood function an essential tool in statistical modeling and inference. The log-likelihood is crucial not only as a method for comparing models but also for various other processes, such as constructing confidence intervals and fitting models. It is also the basis to perform likelihood ratio tests, another method for comparing models we describe next.

2.4.2 Likelihood Ratio Test

The Likelihood Ratio Test (LRT) is used to compare nested models, where one model is a special case of the other, with fewer parameters (by setting some parameters to 0). This test evaluates whether the more complex model provides a significantly better fit than the simpler model [Chen et al., 2020]. Mathematically, if the simpler model has k parameters and the more complex model has $k+m$ parameters, the hypothesis can be written as:

$$\begin{aligned} H_0 &: \quad \beta_{k+1} = \beta_{k+2} = \cdots = \beta_{k+m} = 0 \\ H_1 &: \quad \exists\, \beta_{k+q} \neq 0, \quad q \in \{1, 2, \ldots, m\} \end{aligned}$$

This means that we are testing whether the additional parameters of the more complex model are significantly different from zero. If they are, the additional parameters significantly improve the fit. If all additional parameters are zero, then the simpler model is preferred.

The test statistic, G, is defined as:

$$G = -2 \ln \left(\frac{L_0}{L_1} \right) \underset{H_0}{\sim} \chi^2_q \tag{2.23}$$

$$G = -2(\ln L_0 - \ln L_1) \underset{H_0}{\sim} \chi^2_q \tag{2.24}$$

where $\ln L_0$ is the log-likelihood of the restrict (simpler) model and $\ln L_1$ is the log-likelihood of the full model. Under H_0, the statistic G follows a chi-squared distribution with q degrees of freedom, being q the difference in the number of parameters between the two nested models [Woolf, 1957].

The null hypothesis H_0 is rejected at a significance level α if $G > \chi^2_{q;1-\alpha}$.

2.4.3 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a measure used to assess the parsimony of a statistical model, taking into account the balance between the model's ability to fit the data (via the likelihood) and its complexity (via the number of parameters) [Bevans, 2020]. The AIC is often used in model selection problems, where there is a need to choose between alternative models to describe a set of data. It provides a way of comparing non-nested models that may have different numbers of parameters. The aim is to select the model that best fits the data, but which is also simple enough to avoid overfitting, i.e. avoiding the model to explain very well the data but with poor generalisation to new data. The formula of AIC is:

$$\text{AIC} = 2k - 2\ln(L) \tag{2.25}$$

where k is the number of parameters of the model and L is the maximum value of the likelihood function of the model which represents the best fit of the model. So, considering AIC, the most parsimonious, and hence considered better model, is the one with lowest AIC.

AIC works by penalising models with more parameters because they tend to overfit the data and reduce their predictive power for new data. In this way, there is a balance between the fit and the parsimony of the model. The AIC is a versatile tool that can be used to compare different types of models. It is capable of accurately comparing models with different number of parameters (simple linear model vs. multiple linear model), as well as different types of models (linear model vs. generalised linear model).

Chapter 3

Study Research

This chapter presents the study research. Firstly, a brief description of the data and the variables is provided. All modifications made to the data from its initial collection to the point prior to modelling are explained, affording an understanding of the original, unprocessed dataset and the methodology employed up to the point of creating the final and clean dataset. Then, the statistical methods are presented with a comprehensive account of the data analysis development, as well as the results. Finally, a discussion is held and conclusions are drawn.

3.1 The Data

3.1.1 From Telemetry Data to KUDs (Response Variables)

Passive acoustic telemetry data was downloaded from the European Tracking Network (ETN) online platform. The data originated from 15 studies conducted throughout coastal waters of Europe (Appendix 1), and included detections from different groups of marine species, including teleost fishes.

The raw dataset of detections comprised the transmitter ID, which corresponds to the individual ID, the date and time of detection, the coordinates of the receiving device, and other information that was not relevant for this study and was therefore discarded.

To investigate the influence of biological, ecological and socio-economic characteristics on the range occupied by individuals, it was necessary to convert the detection data, i.e. convert presence/absence data over specific coordinates, into a measure of utilisation area, and corresponding summary statistics.

Firstly, raw detections were processed to estimate short-term centres of activity (COA) for each individual in 60-minute intervals [Simpfendorfer et al., 2002]. This process enabled us to estimate more reliable positions for individuals, as it used weighted means based on the frequency of signal receptions at each receiver over a specific time period. This was a crucial step because it not only increased the accuracy of the fishes' positions as it reduced the dimensionality of the dataset by consolidating multiple detections into a single coordinate per hour.

After estimating the COAs, we proceeded to calculate the KUDs, using a bandwidth of 200m. The choice of bandwidth is always somewhat arbitrary, but it was chosen after testing different values and having knowledge of the study areas and species. Also, when h is mentioned in the literature for this type of study, it is usually between 100 and 300. As this was a comparative study and not a study to accurately estimate the KUDs, it was the consistency of the bandwidth that was important, rather than its value. To ensure a balanced comparison of the individuals' utilisation areas, KUDs were computed on a weekly basis. This approach was necessary because monitoring periods varied widely between individuals. KUDs calculated over each individual entire monitoring period would mean that longer-monitored individuals would tend to show larger utilisation areas, which could lead to confounding in comparisons. Calculating the KUDs on a weekly basis ensured a fair and comparable measure of utilisation across all individuals. This interval is sensible as it allows us to solve the problem of different monitoring periods and also because the utilisation area (home range and core area) may change with the seasonal seasons.

We focused on estimating the KUD95 and KUD50, which represent proxies for the home range and core area, respectively. This involved calculating the 95^{th} and 50^{th} percentiles of each individual's KUD per week. The weekly KUD95 and KUD50 values per individual were then used as response variables in our statistical analysis.

3.1.2 Biological, Ecological and Socio-Economic Traits (Explanatory Variables)

Based on literature and domain knowledge, all of the potential predictors considered were initially selected because *a priori* we believed that they could have an influence, or help predict, the response variables. We selected biological and ecological traits that had already been studied in some intraspecific studies, that could provide interesting results and insights on their relationship with the area used by teleosts, and for which we were able to obtain information, either by collecting it at the time of tagging or through reliable literature or online sources. Therefore, eight traits were selected as explanatory variables to investigate their relationship with the variability of home range and core area. These included 5 numerical variables and 3 categorical variables.

The selected numerical variables were:

- Standardised length: length of the individual divided by the maximum length of the species;
- Standardised body mass: body mass of the individual divided by the maximum body mass considered for the species;
- Longevity: species longevity;
- Vulnerability: species vulnerability;
- Trophic level: species trophic level.

Length and body mass were standardised to allow fair comparisons between species with different sizes and masses. Different species have naturally varying sizes and masses, which can influence observed characteristics (such as activity range or habitat use) and standardising these variables to a common size or weight facilitates direct comparisons between species of different sizes.

The selected categorical variables were:

- Habitat: a three level factor, categorised as demersal, benthopelagic and pelagic-neritic. Demersal individuals inhabit the lowest section of the water column, often living and feeding near or on the seabed. Benthopelagic individuals occupy both the bottom and midwater layers. Pelagic-neritic individuals reside in the upper and middle parts of the water column;
- Migration: a binary variable categorised as non-migratory and oceanodromous, with oceanodromous species undergoing migrations within the ocean;
- Commercial importance: categorised as a three level factor, with levels high, medium, and minor.

Individual length was measured directly when animals were tagged. Body mass was estimated using length and two other parameters, and , according to the formula:

$$M = \alpha L^{\beta} \tag{3.1}$$

where M represents individual body mass, L represents individual length, α describes body shape and condition, and β indicates isometric growth in body proportions [Froese et al., 2014]. Both of these parameters and the all other traits were obtained from FishBase [Froese and Pauly, 2024], a website that provides information on numerous fish species. If no information was available on the site, a literature search was conducted.

Our initial hypotheses were that spatial extent would: **increase** with length, body mass, longevity and trophic level; **decrease** with vulnerability; and **be greater** for pelagic, oceanodromous and highly commercial important individuals.

Table 3.1: Species traits. This table presents the biological, ecological and socioeconomic traits extracted for each species.

Specie	Mean LengthStd (max=1)	Mean BodyMassStd (max=1)	Longevity (years)	Vulnerability (max=100)	Trophic Level (max=5)	Habitat	Migration	Commercial Importance	Spawning Season
Dactylopterus volitans	0.786	0.553	20	36.5	3.65	benthopelagic	non-migratory	minor	SA
Dentex dentex	0.557	0.291	28	66.3	4.53	benthopelagic	non-migratory	high	SS
Dicentrarchus labrax	0.482	0.113	30	69.1	3.47	benthopelagic	oceanodromous	high	W
Diplodus cervinus	0.891	0.929	18.6	54.6	2.99	demersal	non-migratory	medium	W
Diplodus sargus	0.597	0.193	10	63.4	3.38	demersal	non-migratory	high	SS
Diplodus vulgaris	0.552	0.196	12	47.8	3.52	demersal	non-migratory	medium	W
Epinephelus marginatus	0.493	0.115	60	71.5	4.1	demersal	non-migratory	high	SS
Gadus morhua	0.223	0.009	25	65.4	4.09	benthopelagic	oceanodromous	high	SS
Labrus bergylta	0.523	0.189	29	66.6	3.97	demersal	non-migratory	medium	SS
Lichia amia	0.445	0.165	16.2	90	4.5	pelagic-neritic	oceanodromous	medium	SS
Lithognathus mormyrus	0.595	0.199	12	40.4	3.42	demersal	non-migratory	minor	SS
Pagellus erythrinus	0.499	0.207	15	39.8	3.48	benthopelagic	non-migratory	medium	SA
Pagrus pagrus	0.373	0.087	18	66.4	3.86	benthopelagic	non-migratory	high	SS
Pomatomus saltatrix	0.338	0.056	9	57.6	4.53	pelagic-neritic	oceanodromous	high	SS
Pseudocaranx dentex	0.406	0.079	49	74.3	3.92	benthopelagic	oceanodromous	medium	SS
Sciaena umbra	0.508	0.295	21	63.6	3.75	benthopelagic	oceanodromous	medium	SS
Scorpaena porcus	0.578	0.225	18	50.5	3.92	demersal	non-migratory	minor	SS
Scorpaena scrofa	0.405	0.119	25	67.5	4.25	demersal	non-migratory	medium	SS
Seriola dumerili	0.344	0.044	15	54	4.66	pelagic-neritic	oceanodromous	medium	SS
Seriola rivoliana	0.384	0.101	13.5	75.8	4.45	pelagic-neritic	oceanodromous	medium	SS
Serranus atricauda	0.748	0.549	16	60.7	4.27	demersal	non-migratory	medium	SS
Serranus cabrilla	0.315	0.041	8	36	3.35	demersal	non-migratory	minor	SS
Serranus scriba	0.405	0.119	16	38.1	3.82	demersal	non-migratory	minor	SS
Solea senegalensis	0.492	0.157	26.4	49.2	3.25	demersal	non-migratory	high	SS
Sparisoma cretense	0.803	0.427	8	35.9	2.86	demersal	non-migratory	high	SS
Sparus aurata	0.519	0.042	11	40.2	3.7	demersal	non-migratory	high	SA
Sphyraena viridensis	0.557	0.239	14	68.9	4.31	pelagic-neritic	oceanodromous	medium	SS
Spondyliosoma cantharus	0.433	0.600	22.7	37.2	3.34	benthopelagic	non-migratory	medium	SS
Umbrina cirrosa	0.726	0.169	26.5	40.4	3.41	demersal	non-migratory	minor	SS
Xyrichtys novacula	0.466	0.938	8	36.3	3.51	demersal	non-migratory	minor	SS

3.1.3 Study Design Parameters (Explanatory Variables)

Detection data was obtained from 15 different studies carried out all over Europe, meaning that each one was performed under different settings regarding the data collection. Some of the properties that varied across studies were the monitored area (km^2), the number of receivers installed and the maximum distance between receivers (km). The minimum convex polygon (MCP in km^2, represents the smallest polygon that can be created with all the receivers contained in it and all the angles less than 180º) also varied across studies. This metric was estimated using the coordinates of the receivers and taking into account the coastline to remove areas on land. Through these numbers, it was also possible to calculate the receiver density, dividing the number of receivers by the MCP. Receiver density gives us an idea of sampling intensity and is an important metric in telemetry studies. In the end, five numerical study design parameters were included in the dataset as explanatory variables as they may be relevant to explain the behaviour of the KUDs.

Table 3.2: Study Design Parameters. This table presents the various species and corresponding studies, including the array ID and its design parameters.

File	Species	ArrayID	NReceivers	MonitoredArea_km2	MaxDistReceivers_km	MCP_km2	ReceiverDensity
Dactylopterus_volitans	*Dactylopterus volitans*	1	98	15.62	55.55	363.5428	0.27
Dentex_dentex1	*Dentex dentex*	15	26	2.72	3.21	1.4701	17.69
Dentex_dentex2	*Dentex dentex*	1	98	15.62	55.55	363.5428	0.27
Dicentrarchus_labrax1	*Dicentrarchus labrax*	2	86	12.64	87.39	2990.2885	0.03
Dicentrarchus_labrax2	*Dicentrarchus labrax*	1	98	15.62	55.55	363.5428	0.27
Diplodus_cervinus	*Diplodus cervinus*	1	98	15.62	55.55	363.5428	0.27
Diplodus_sargus1	*Diplodus sargus*	3	23	3.56	7.11	3.3126	6.94
Diplodus_sargus2	*Diplodus sargus*	4	15	1.5	1.46	0.6133	24.46
Diplodus_sargus3	*Diplodus sargus*	4	15	1.5	1.46	0.6133	24.46
Diplodus_sargus4	*Diplodus sargus*	15	26	2.72	3.21	1.4701	17.69
Diplodus_sargus5	*Diplodus sargus*	5	24	3.7	42.98	114.3315	0.21
Diplodus_sargus6	*Diplodus sargus*	1	98	15.62	55.55	363.5428	0.27
Diplodus_vulgaris1	*Diplodus vulgaris*	6	11	2.16	4.5	5.8522	1.88
Diplodus_vulgaris2	*Diplodus vulgaris*	1	98	15.62	55.55	363.5428	0.27
Epinephelus_marginatus1	*Epinephelus marginatus*	7	45	8.38	74.03	355.9882	0.13
Epinephelus_marginatus2	*Epinephelus marginatus*	8	27	0.89	1.62	0.3332	81.03
Epinephelus_marginatus3	*Epinephelus marginatus*	1	14	1.84	2.43	1.3607	10.29
Epinephelus_marginatus4	*Epinephelus marginatus*	1	98	15.62	55.55	363.5428	0.27
Gadus_morhua1	*Gadus morhua*	9	25	3.4	2.68	2.0148	12.41
Gadus_morhua2	*Gadus morhua*	10	55	7.64	9.25	5.4664	10.06
Gadus_morhua3	*Gadus morhua*	11	11	0.83	0.78	0.1982	55.5
Labrus_bergylta	*Labrus bergylta*	12	12	1.2	1.11	0.4318	27.79
Lichia_amia	*Lichia amia*	1	98	15.62	55.55	363.5428	0.27
Lithognathus_mormyrus	*Lithognathus mormyrus*	1	98	15.62	55.55	363.5428	0.27
Pagellus_erythrinus	*Pagellus erythrinus*	1	98	15.62	55.55	363.5428	0.27
Pagrus_pagrus1	*Pagrus pagrus*	7	45	8.38	74.03	355.9882	0.13
Pagrus_pagrus2	*Pagrus pagrus*	1	98	15.62	55.55	363.5428	0.27
Pomatomus_saltatrix	*Pomatomus saltatrix*	1	98	15.62	55.55	363.5428	0.27
Pseudocaranx_dentex	*Pseudocaranx dentex*	7	45	8.38	74.03	355.9882	0.13
Sciaena_umbra1	*Sciaena umbra*	13	19	0.69	0.53	0.1447	131.31
Sciaena_umbra2	*Sciaena umbra*	1	98	15.62	55.55	363.5428	0.27
Scorpaena_porcus	*Scorpaena porcus*	13	19	0.69	0.53	0.1447	131.31
Scorpaena_scrofa1	*Scorpaena scrofa*	13	19	0.69	0.53	0.1447	131.31
Scorpaena_scrofa2	*Scorpaena scrofa*	1	98	15.62	55.55	363.5428	0.27
Seriola_dumerili	*Seriola dumerili*	1	98	15.62	55.55	363.5428	0.27
Seriola_rivoliana	*Seriola rivoliana*	14	4	0.79	08.08	12.2923	0.33
Serranus_atricauda	*Serranus atricauda*	7	45	8.38	74.03	355.9882	0.13
Serranus_cabrilla	*Serranus cabrilla*	6	21	2.77	2	2	10.5
Serranus_scriba	*Serranus scriba*	6	25	4.7	2.74	3.8682	6.46
Solea_senegalensis	*Solea senegalensis*	3	23	3.56	7.11	3.3126	6.94
Sparisoma_cretense	*Sparisoma cretense*	7	45	8.38	74.03	355.9882	0.13
Sparus_aurata1	*Sparus aurata*	13	19	0.69	0.53	0.1447	131.31
Sparus_aurata2	*Sparus aurata*	1	98	15.62	55.55	363.5428	0.27
Sphyraena_viridensis1	*Sphyraena viridensis*	14	4	0.79	08.08	12.2923	0.33
Sphyraena_viridensis2	*Sphyraena viridensis*	1	98	15.62	55.55	363.5428	0.27
Spondyliosoma_cantharus	*Spondyliosoma cantharus*	1	98	15.62	55.55	363.5428	0.27
Umbrina_cirrosa	*Umbrina cirrosa*	1	98	15.62	55.55	363.5428	0.27
Xyrichtys_novacula	*Xyrichtys novacula*	6	21	2.41	1.64	1.2431	16.89

3.1.4 Data Cleaning

For this study, only teleost fish were considered, with all other species excluded from the unfiltered dataset (six species of cartilaginous fish and one of crustacean). There were five individuals who were missing length information, so to prevent potential modelling issues, the detections for those individuals were removed. These individuals were removed because their absence would not significantly impact the results, given the large dataset. Individuals with less than five different COAs were also removed to avoid errors in the `kernelUD()` function from the `adehabitatHD` package [Calenge, 2006] in R, i.e. when estimating the KUDs, since the function only works with a minimum of 5 different locations. Naturally, with fewer than five detections it can be difficult to accurately estimate the individual space use pattern.

The final dataset contained 25612 weekly KUDs from 874 individuals of 30 different species. Some of these species were monitored in more than one study and therefore in different geographical locations. A new variable termed File was created to distinguish the species monitored in more than one site, treating them, effectively, as different species. This was done because the same species can differ in biological and behavioural characteristics if they are far apart geographically and are under different environmental pressures. Appendix 2 provides a better understanding of the origin and dimensionality of the data.

Table 3.3: Characterisation of all variables.

	Variable	**Meaning**	**Type**	**Codification**
Response variables	KUD95	Home range	Numerical	
	KUD50	Core area	Numerical	
Traits	LengthStd	Standardised length	Numerical	
	BodyMassStd	Standardised body mass	Numerical	
	Vulnerability	Vulnerability	Numerical	
	Longevity	Longevity	Numerical	
	Troph	Trophic level	Numerical	
	Habitat	Habitat	Categorical	Benthopelagic, Demersal, Pelagic-neritic
	Migration	Migration	Categorical	Non-migratory, Oceanodromous
	ComImport	Commercial importance	Categorical	High, Medium, Minor
Study design parameters	MonitArea_km2	Monitored area (km^2)	Numerical	
	MCP_km2	Minimum Convex Polygon (km^2)	Numerical	
	NReceivers	Receiver number	Numerical	
	MaxDistReceivers	Maximum distance between receivers	Numerical	
	ReceiverDensity	Receiver density	Numerical	
Random effects variables	Transmitter	Identification of the individual	Categorical	Note: 25612 different codifications
	File	Species by study	Categorical	Note: 48 different codifications
	Species	Species	Categorical	Note: 30 different codifications

3.2 Statistical Methods

All analyses were performed in R [R Core Team, 2024] via RStudio [Posit team, 2023] and a list of the required packages is available at the end of this document. For all tests we used a significance level of $\alpha = 0.05$.

Exploratory Data Analysis

First, an exploratory data analysis was carried out for both the response variables (KUD95 and KUD50) and the explanatory variables (biological, ecological and socio-economic traits and study design parameters). This is an important step before moving on to deeper and more specific analysis, because it gives us a better perception of the nature of the data, its dimensionality and how it behaves.

Boxplots of the response variables were constructed to assess the presence of discrepant observations. To determine whether outliers should be included in the dataset, an analysis was conducted. Potential outliers, as judged by the Interquartile Range (IQR) method (limit +/- 1.5*IQR), were initially removed, and the behaviour of the remaining data was carefully observed. If, after removing these observations, additional observations stood out as potential outliers, the decision would be made to retain all the data. Conversely, if the removal of outliers resulted in a distribution without any further outliers, the decision would be to discard them.

As this project is based on regression models, in this exploratory analysis, some of the assumptions of linear models were checked. Given they are the simplest models, and a good starting point for analysis, evaluating whether they can be applied or whether another type of model, more general and potentially non-linear, is needed.

Regarding the explanatory variables, correlation between all numerical traits and all study design parameters was inspected to avoid multicollinearity in the models. We used Pearson's coefficient correlation method. To test the linearity assumption for the numerical explanatory variables, two methods were employed. Firstly, we inspected it visually, plotting the KUDs against each of the explanatory variables, to ascertain whether there was evidence of a linear pattern. To confirm our suspicions, we employed Pearson's coefficient correlation method.

In an attempt to solve the problem of non-linearity of the data, transformations of the explanatory variables and of the response variables were explored.

Model Type Testing

To confirm the inadequacy of the linear model to explain the data, simple and multiple linear models were implemented and a residual analysis was performed. We plotted the empirical quantiles of the residuals against the theoretical quantiles of the standardised normal distribution to check the normality assumption and plotted the residuals against the fitted values to check the homoscedasticity assumption.

We then went on to consider other types of models, such as GLM, GAM and also mixed models. Mixed models were applied since there was a dependence between data collected on the same individual and on the same species. For all types of models, the responses were modelled as a function of each

explanatory variable individually and all together. The Gamma family and logarithmic link function were used. For the mixed models, three variables were considered as potential random effects (Species, File and Transmitter) and were experimented one at a time. The GLM and GAM models were implemented using the `glm()` and `gam()` functions of the `stats` [R Core Team, 2023] and `mgcv` [Wood, 2011] packages, respectively. The GLMM and GAMM models were implemented using the `glmmTMB()` and `gamm()` functions of the `glmmTMB` [Brooks et al., 2017] and `mgcv` packages, respectively.

To compare the different models and see which fitted best, we considered AIC. We also used the AIC of the models to observe what was the most relevant variable responsible for the random effect. A number of different approaches were also used to assess the importance of these variables when considered as fixed effects, the results of which we would expect to be consistent with the use of AIC (see Appendix 6).

Finally, to check which of these variables should be included in the models as random effects (if only one, two or all of them), we constructed full models with all the combinations of the 3 variables and analysed the AIC.

Evaluation of the significant biological, ecological and socio-economic traits and study design parameters

Initially, we analysed the significance of each trait separately in determining space use, using the GLMM model with Species and Transmitter as random effects.

Then, to evaluate if we could achieve a better fit by adding more variables, we performed a backward elimination. We first started by including all traits variables in the GLMM model: length, body mass, longevity, vulnerability, trophic level, habitat, migration, and commercial importance. We also included the monitored area and receiver density as study design parameters. In the random effects we included transmitter and species. After fitting the full model, we checked if there was any variable whose coefficient, β_j, was not statistically different from zero (i.e. $p - value > \alpha, \alpha = 0.05$), by checking the p-values for the Wald test statistics. If there was any variable in this condition, then it was a candidate to exit the model. If more than one variable was in this condition, then the choice was made to remove the one who had the largest p-value. This process was repeated until no more variables were candidates to be excluded. In the end, we compared the full model with the final model and the final model with the previous model using the likelihood ratio test of the `anova()` function to determine if they were statistically different. If no evidence of statistical differences between the two models were found, we retained the simpler model for inference, to comply with the principle of parsimony. We also performed this test to assess whether the final model, obtained with the significant variables, was statistically different from the simpler models that included each trait alone, evaluating whether the final model provided a better fit to the data by more significantly explaining the existing variability in KUDs.

To validate and assess the fit of the final models, a residual analysis was performed, using the `DHARMa` package [Hartig, 2022]. First, we plotted the residuals fitted vs. simulated to inspect the overdispersion. A Q-Q plot was also made to check the normality and the homoscedasticity was checked by plotting the residuals vs. predicted values. Finally, we assess the predictive capacity of the models by simulating data

with the DHARMa package and plotting them alongside with the real observations.

To analyse the influence of spawning season in KUDs within species, we fitted GLMM models for each species with Transmitter as a random effect to observe whether the utilisation area of individuals varied with spawning season and, if so, whether it tended to increase or decrease with spawning season.

Spawning season analysis within species

First, the spawning season of each species was assessed using FishBase. For species for which spawning season information was missing from the site, we retrieved the data from the literature. Using this information, we created two new variables. The first was a categorical variable called SpawnSeason, which indicated whether the species had their spawning season in Spring/Summer (SS), Autumn (A), or Winter (W). The second was a binary variable called Spawn. For each observation, this variable indicated whether the individual was in the spawning season. Since each observation corresponded to a week (as KUDs were estimated on a weekly basis), the new variable indicated if the individual was in (yes) or out (no) of spawning season during that week. Next, to analyse the influence of the spawning season in KUDs, we fitted GLMMs to each Species with Transmitter as a random effect to observe if there was any significant relationship between the spawning season and the utilisation area and, if so, whether the area of the individuals tended to increase or decrease with the spawning season.

3.3 Results

First the results of the exploratory analysis are shown, followed by the results of the model type testing and then the results of the evaluation of the significant traits and study design parameters.

Exploratory Analysis

The dataset comprised 25612 observations, including two response variables (KUD95 and KUD50) and eight potential explanatory variables. Table 3.4 provides a summary of the two response variables. To have a better understanding on the type of data being analysed, a table containing some of the species from different studies/sites, as well as their KUD95 and KUD50 values is shown in Appendix 3).

Table 3.4: Summary table of KUD95 and KUD50.

KUD95		**KUD50**	
Min.	:0.760	Min.	:0.1620
1st Qu.	:0.762	1st Qu.	:0.1640
Median	:0.818	Median	:0.1730
Mean	:1.124	Mean	:0.2267
3rd Qu.	:1.188	3rd Qu.	:0.2390
Max.	:14.723	Max.	:3.2780

From this table we could observe that the maximum value of both responses was significantly distant from the 3rd quartile, which could indicate the presence of outliers or observations that exhibit greater variability than expected. To investigate this further, boxplots were created (Figure 3.1).

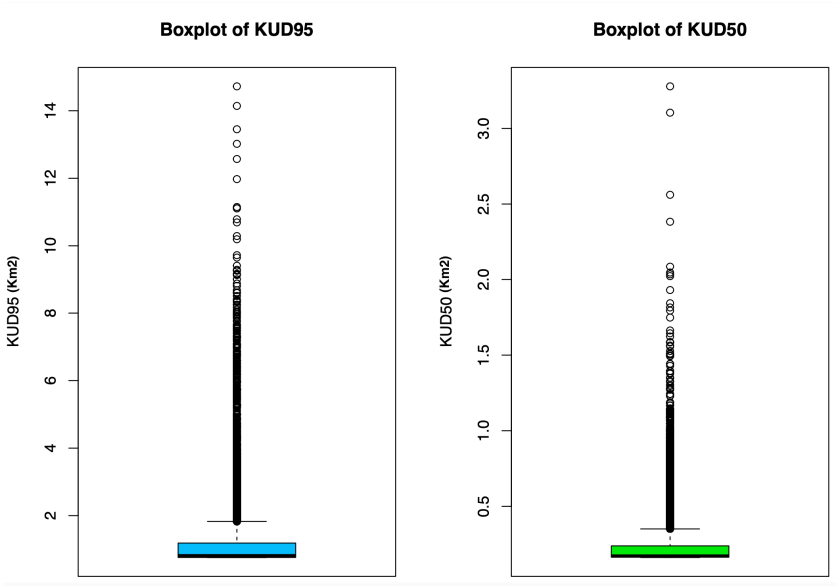


Figure 3.1: Boxplots to examine the presence of outliers. Boxplot of KUD95 on the left (blue) and the boxplot of KUD50 on the right (green).

The boxplots demonstrated that there was high concentration of observations at low values of the KUDs, with a considerable number of points outside the boxes. After we removed the outliers, new ones appeared, as can be seen in Figure 3.2. We decided not to discard any of the outliers and maintain all observations in the dataset for the analysis.

While for our results we only report the analysis including the outliers, analysis excluding them had negligible impact on the outputs and did not change the final results.

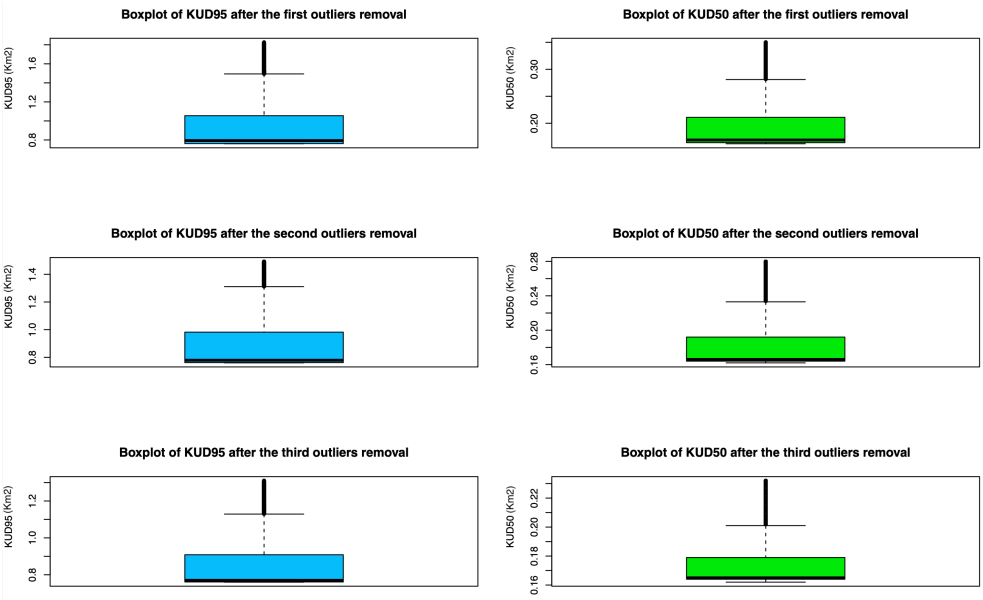


Figure 3.2: Boxplots showing the outliers after three consecutive removals. Outliers removal process of KUD95 on the left (blue) and the outliers removal process of KUD50 on the right (green). The first row shows the boxplots after the first removal, the second row the boxplots after the second removal, and the third row the boxplots after the third removal.

The correlations between the explanatory variables were assessed and showed that the only traits exhibiting a strong linear correlation ($> 70\%$) between the traits were length and body mass, with a value of approximately 0.75. All the values of the correlation analysis can be checked in Table 3.5.

Table 3.5: Pearson's correlations between biological, ecological and socio-economic traits. The numbers outlined in red correspond to the variables with a high linear correlation.

	Pearson's Correlation				
	Length Std	**Body Mass Std**	**Longevity**	**Vulnerability**	**Troph**
Length Std	1	0.7489714	-0.1539612	-0.2422920	-0.3042097
Body Mass Std		1	-0.1734132	-0.3117613	-0.1870678
Longevity			1	0.3942677	0.2072966
Vulnerability				1	0.5659924
Troph					1

Concerning the correlations between the experimental design parameters (Table 3.6), we found strong linear correlations ($> 70\%$) between monitored area and number of receivers, and between monitored area and maximum distance between receivers, with approximately 0.98 and 0.74, respectively.

Table 3.6: Pearson's correlations between experimental design parameters. The numbers outlined in red correspond to the variables with a high linear correlation.

	Pearson's Correlation				
	Receiver Density	**Monitored Area**	**MCP**	**Number of Receivers**	**Max. Dist. Receivers**
Receiver Density	1	-0.3835498	-0.2094113	-0.2748357	-0.4524428
Monitored Area		1	0.4583335	0.9845062	0.7488135
MCP			1	0.4697432	0.5886984
Number of Receivers				1	0.6589268
Max. Dist. Receivers					1

Regarding the inspection of the linearity assumption, the results indicated that there was no discernible linear relationship between the response and the numerical explanatory variables. The plots demonstrated evidence of non-linear patterns for all the variables (Appendices 4 and 5) and the correlation test showed very small values (Table 3.7), being that the correlation values for KUD95 ranged from |0.03| to |0.33| and the correlation values for KUD50 ranged from |0.02| to |0.33|.

Table 3.7: Pearson's correlations between each response variables (KUD95 and KUD50) and the explanatory variables.

	Pearson's Correlation	
	KUD95	**KUD50**
Length Std	-0.0215178	-0.0124221
Body Mass Std	-0.0553196	-0.0451847
Longevity	-0.0806592	-0.0925148
Vulnerability	-0.0492247	-0.0836845
Troph	0.0608373	0.0350152
Receiver Density	-0.1214700	-0.1059148
Monitored Area	0.3207865	0.3134040
MCP	0.0804618	0.0640651
Number of Receivers	0.3311147	0.3287046
Max. Dist. Receivers	0.1316282	0.1082455

According to the boxplots created to inspect the plausibility of a linear model for the categorical explanatory variables (Figure 3.3) there seemed to be no evidence of differences between the groups or categories for each variable. This implies that the influence of the factor covariates is likely to be mild.

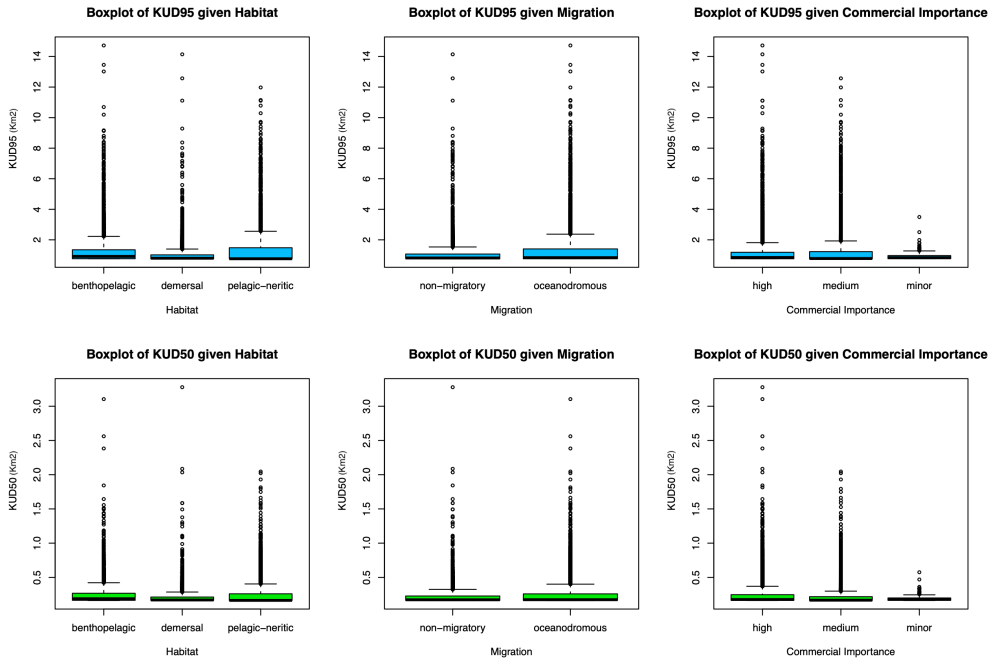


Figure 3.3: Boxplots displaying the distribution of the response variable for each category of the explanatory variables.

The transformations performed on the explanatory variables carried out to ascertain the linearity assumption showed no significant improvements in the linearity aspect, with the highest correlation value between KUD95 and the explanatory variables being |0.36|, observed with the exponential transformation, and the highest value between KUD50 and the explanatory variables being |0.35|, also with the exponential transformation. We also checked if by any chance the normality and homoscedasticity assumptions were met after the transformations but we had no positive results. All this information is summarised in Table 3.8.

Table 3.8: Summary table of the transformations performed on the explanatory variables. The table shows the Pearson's correlation range between the transformed explanatory variables and the KUDs, and also if the linearity, normality and homoscedasticity assumptions are met. The use of "No" means that, after adjusting the linear model for each transformed explanatory variable and inspecting the residuals, none of them satisfied the assumption.

Explanatory Variables Transformations

Transformation	Correlation Range		Linearity		Normality		Homoscedasticity	
	KUD95	**KUD50**	**KUD95**	**KUD50**	**KUD95**	**KUD50**	**KUD95**	**KUD50**
Logarithmic	(|0.02| , |0.25|)	(|0.007| , |0.26|)	No	No	No	No	No	No
Square root	(|0.02| , |0.30|)	(|0.009| , |0.30|)	No	No	No	No	No	No
Exponencial	(|0.02| , |0.36|)	(|0.01| , |0.35|)	No	No	No	No	No	No
Inverse	(|0.004| , |0.19|)	(|0.003| , |0.20|)	No	No	No	No	No	No

Transformations on the response variables were also tried, but the underlying assumptions of the linear model were never verified (see Table 3.9.), implying that some extension to the linear model would be required.

Table 3.9: Summary table of the transformations performed on the response variables. The table shows the Pearson's correlation range between the explanatory variables and the transformed KUDs, and also if the linearity, normality and homoscedasticity assumptions are met. The use of "No" means that, after adjusting the linear model for each transformed explanatory variable and inspecting the residuals, none of them satisfied the assumption.

Response Variables Transformations

Transformation	Correlation Range		Linearity		Normality		Homoscedasticity	
	KUD95	**KUD50**	**KUD95**	**KUD50**	**KUD95**	**KUD50**	**KUD95**	**KUD50**
Logarithmic	(|0.03| , |0.36|)	(|0.01| , |0.35|)	No	No	No	No	No	No
Square root	(|0.03| , |0.36|)	(|0.01| , |0.35|)	No	No	No	No	No	No
Exponencial	(|0.005| , |0.23|)	(|0.003| , |0.22|)	No	No	No	No	No	No
Inverse	(|0.003| , |0.34|)	(|0.008| , |0.34|)	No	No	No	No	No	No

Model Type Testing

As the data did not fulfil the linear model assumptions (Linearity, Normality, Homoscedasticity and Independence), we attempted to fit other types of models, which did not require some of these assumptions to hold. We fitted the responses against each one of the traits. The AIC of those models were analysed (Table 3.10) and we realised that the best fitted model, with the lowest AIC, was the generalised linear mixed model using Transmitter as the random effect, for both KUD95 and KUD50. Therefore we decided to continue the analysis with the GLMM.

Table 3.10: AIC of the different types of models fitted (LM, GLM, GAM, GLMM and GAMM). The first table corresponds to models with respect to KUD95 and the second table to models with respect to KUD50. Models named with T, F and S, correspond to mixed-effects models in which the random effects are attributed to transmitter (T), to file (F) or to species (S). The numbers outlined in red correspond to the models with the lowest AIC, i.e. the best models.

AIC KUD95

	Length Std	**Body Mass Std**	**Longevity**	**Vulnerability**	**Troph**	**Habitat**	**Migration**	**Commercial Importance**
lm	60408.26	60341.62	60252.94	60357.98	60325.15	59545.84	59783.05	60400.97
glm	35354.67	35194.80	34951.62	35242.26	35173.93	33309.53	33916.45	35333.69
gam	42791.40	42506.22	42313.90	42706.31	42397.54	39499.58	40261.07	42658.62
glmm_T	10331.08	10330.63	10325.84	10332.37	10328.52	10242.59	10304.26	10318.17
glmm_F	20408.20	20457.95	20491.38	20491.43	20488.42	20467.43	20480.58	20492.95
glmm_S	25565.58	25623.81	25626.29	25624.76	25622.11	25608.63	25616.83	25627.65
gamm_T	16441.62	16438.86	16383.01	16440.07	16315.53	16323.07	16400.94	16435.14
gamm_F	30060.72	30157.54	30657.31	30654.78	30654.85	30636.69	30647.16	30655.11
gamm_S	38088.52	38618.86	38999.43	38995.83	38996.26	38977.48	38982.97	38995.75

AIC KUD50

	Length Std	**Body Mass Std**	**Longevity**	**Vulnerability**	**Troph**	**Habitat**	**Migration**	**Commercial Importance**
lm	-31828.99	-31877.38	-32045.20	-32005.03	-31856.46	-32396.89	-32191.53	-31846.21
glm	-52604.94	-52704.33	-53090.12	-52955.22	-52657.77	-53780.41	-53333.40	-52643.88
gam	-47252.02	-47446.11	-47897.16	-47660.03	-47370.47	-48992.27	-48532.86	-47290.79
glmm_T	-75923.01	-75920.93	-75930.00	-75921.04	-75923.01	-75984.63	-75937.60	-75942.37
glmm_F	-65932.14	-65875.79	-65829.55	-65829.82	-65832.43	-65850.73	-65839.26	-65827.93
glmm_S	-61214.33	-61148.93	-61149.69	-61151.52	-61153.70	-61164.81	-61157.96	-61148.70
gamm_T	10269.63	10267.54	10223.12	10266.08	10132.52	10179.97	10241.27	10261.3
gamm_F	22567.53	22765.65	23189.64	23189.07	23187.82	23167.91	23180.12	23189.53
gamm_S	29667.58	30072.04	30353.56	30351.3	30349.77	30333.2	30339.13	30350.69

Regarding the importance of each one of the variables associated with the random effects (Species, File and Transmitter), the AIC showed that the Transmitter was more important in explaining the variability of the data, followed by File and finally Species.

As for the analysis carried out to check which of these random effect variables should be included in the models (if only one, two, or all of them), we observed the AICs of the fitted full models with all

possible combinations of the random effects and the results showed that the best fitted model was the GLMM with Transmitter and File as random effects (Table 3.11), since these models had the lowest AIC. However, after performing `anova()` to compare `glmm_TF` with `glmm_TS`, we found no evidence that these two models differed, so we decided to continue our study with Transmitter and Species as random effects, as Species is the simpler variable between File and Species.

Table 3.11: AIC of the GLMMs combining all random effects variables. Models named with T, F and S, correspond to mixed-effects models in which the random effects are attributed to transmitter (T), to file (F) and/or to species (S).

	AIC	
	KUD95	**KUD50**
glmm_T	9994.907	-76204.64
glmm_F	20357.009	-65971.34
glmm_S	22496.812	-63412.70
glmm_TF	9849.845	-76365.69
glmm_TS	9946.192	-76245.35
glmm_FS	20359.009	-65969.34
glmm_TFS	9851.845	-76363.69

Evaluate the significant biological, ecological and socio-economic traits and study design parameters

In our separate evaluation of each trait to determine its influence on home range and core area, we found that trophic level, habitat, and migration significantly explained the variability of KUD95 and KUD50 when considered individually. Our results showed that home range and core area increased with trophic level. For a 1 unit increase in trophic level, the home range increased by an average of 28% and the core area by 22%. In terms of habitat, space use was largest for pelagic individuals, followed by benthopelagic, and then demersal individuals. Specifically, pelagic individuals had, on average, 58% more home range than benthopelagic individuals, while demersal individuals had 9% less home range compared to benthopelagics. For core area, pelagic individuals had 43% more than benthopelagics, and demersal individuals had 9% less. Regarding migration, oceanodromous individuals had larger space use compared to non-migratory ones. The home range was, on average, 37% larger for oceanodromous individuals compared to non-migratory ones, and the core area was, on average, 28% larger for oceanodromous individuals compared to non-migratory ones.

Regarding the backward elimination, carried out to investigate which were the significant variables that explained the variability of KUDs, a table containing the whole process, from the first fitted model with all the selected variables to the last one with only the ones considered to correspond to the best model, is shown in Appendix 7 for KUD95 and KUD50.

The results of the likelihood ratio test comparing the full model with the final model showed no

evidence that the two models were significantly different, for both KUD95 and KUD50, so we retained the smaller model to comply with the principle of parsimony.

For both response variables, KUD95 and KUD50, the backward elimination excluded seven out of ten explanatory variables. For KUD95, variables were removed in the following order: receiver density, longevity, body mass, trophic level, migration, vulnerability and length. For KUD50, variables were removed in the following order: body mass, receiver density, longevity, migration, trophic level, vulnerability and commercial importance. In the end, KUD95 remained with habitat, commercial importance and monitored area, and KUD50 remained with length, habitat and monitored area (see Table 3.12, as these were the variables whose coefficient was statistically different from zero (at the 5% significance level). We found that home range (KUD95) was larger for pelagic individuals and individuals of high commercial importance and increased with monitored area, and core area (KUD50) increased with length and monitored area, and was larger for pelagic individuals. More details on these results and their interpretation can be found in Final Models Equations.

Table 3.12: Parameters of the best models for KUD95 and KUD50. The table on the left corresponds to the final KUD95 model and the table on the right to the final KUD50 model.

Best model KUD95

	Estimate	Std. Error	p-value
(Intercept)	-0.074	0.071	0.2972
Habitatdemersal	-0.079	0.076	0.3005
Habitatpelagic-neritic	0.499	0.104	1.56e-06 ***
ComImportmedium	-0.144	0.073	0.0479 *
ComImportminor	-0.111	0.097	0.2549
MonitArea_km2	0.029	0.003	< 2e-16 ***

Best model KUD50

	Estimate	Std. Error	p-value
(Intercept)	-1.753	0.074	< 2e-16 ***
LengthStd	0.220	0.100	0.027846 *
Habitatdemersal	-0.125	0.076	0.100496
Habitatpelagic-neritic	0.365	0.104	0.000465 ***
MonitArea_km2	0.022	0.003	< 2e-16 ***

The likelihood ratio test comparing the final model with the ones including each trait alone showed significant differences, indicating improvements in the model fit after the inclusion of additional variables.

The results of the residual analysis performed on the final models showed overdispersion, as we can see in Figure 3.4, which was also confirmed by the dispersion test, which gave a p-value < 2.2e-16 for KUD95 fitted model and a p-value = 0.008 for KUD50 fitted model. The residual analysis also revealed that the simulated residuals did not follow a uniform distribution (Figure 3.5) (in `DHARMa` package, the residuals are designed to follow a uniform distribution under the assumption that the model is correctly specified) and were heteroscedastic (Figure 3.6). These results suggested that the models were not good at explaining the variability of the data. However, after we plot the real values and compare them with the simulated ones, it was reassuring to see that the patterns were very similar (See Figure 3.7 and 3.8 to check the comparisons for KUD95 and Appendix 8 and 9 to check the comparisons for KUD50).

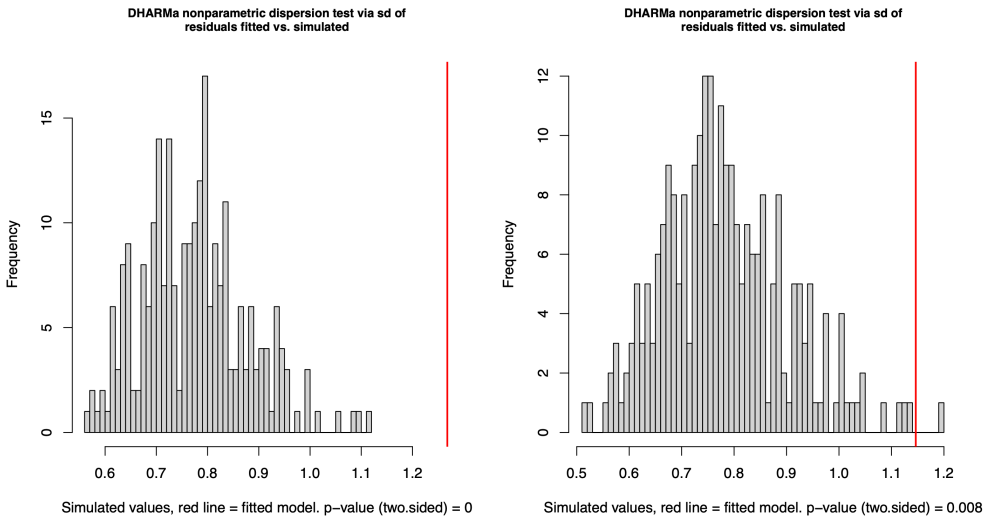


Figure 3.4: Overdispersion plots of the final models. The plot from the left refers to the KUD95 model and the plot from the right refers to the KUD50 model.

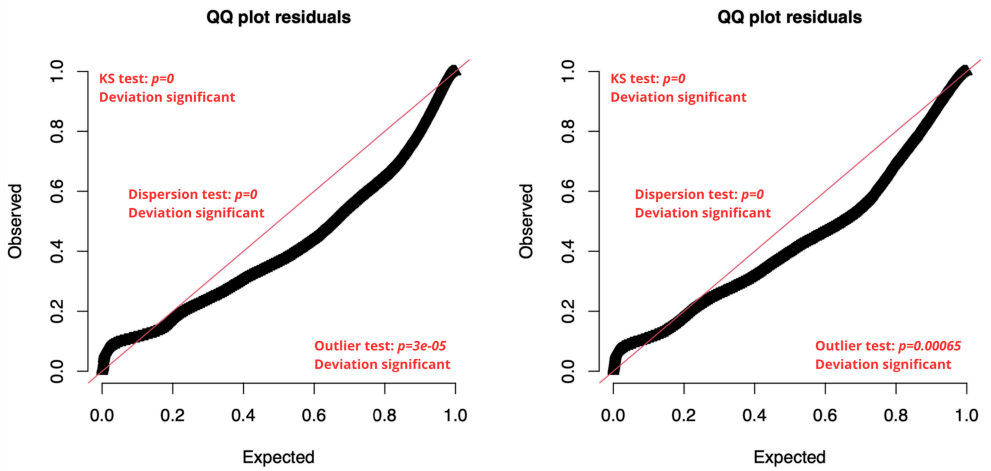


Figure 3.5: Q-Q plots residuals of the final models. The plot from the left refers to the KUD95 model and the plot from the right refers to the KUD50 model.

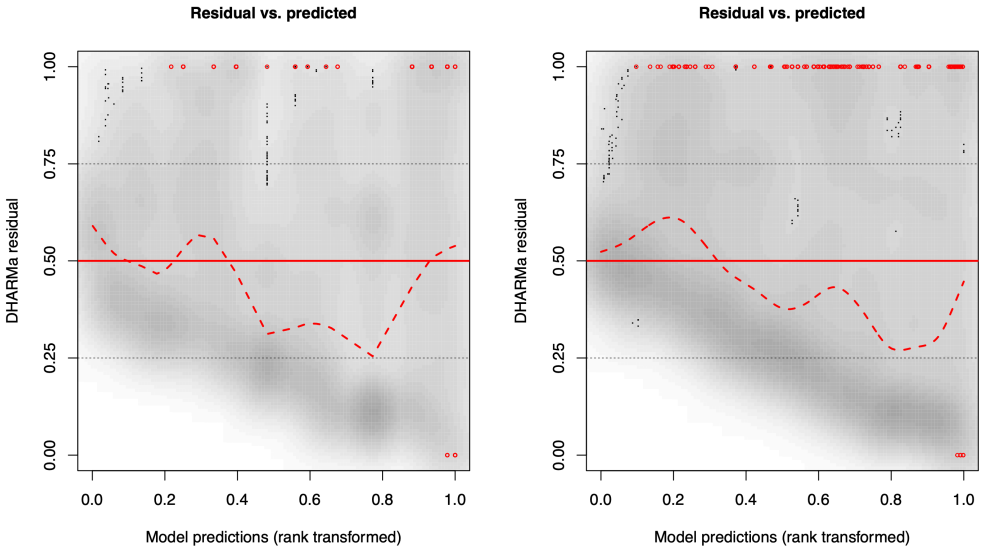


Figure 3.6: Residuals vs. predicted plots of the final models. The plot from the left refers to the KUD95 model and the plot from the right refers to the KUD50 model.

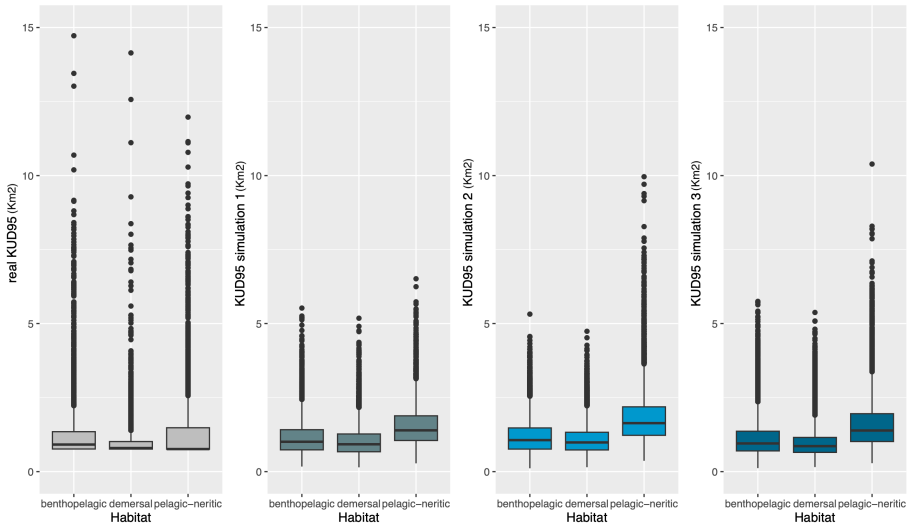


Figure 3.7: Comparison plots between observed and simulated habitat values for KUD95.

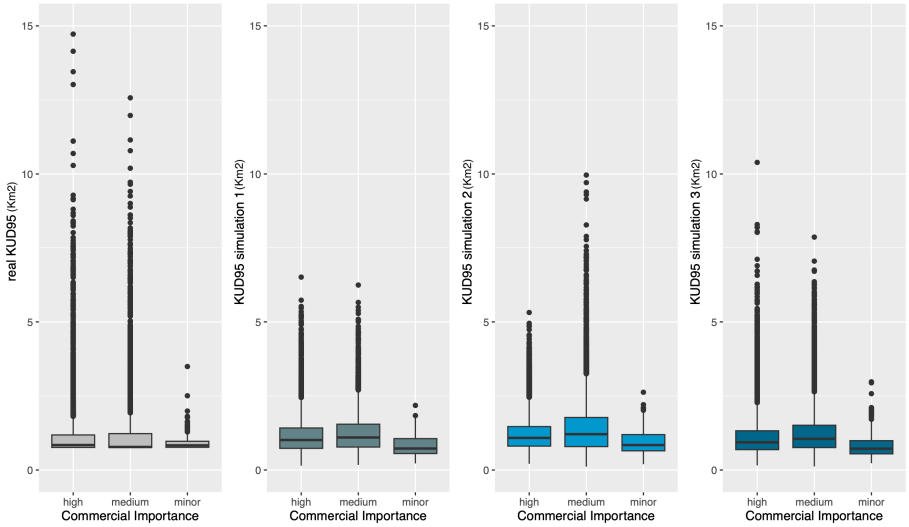


Figure 3.8: Comparison plots between observed and simulated commercial importance values for KUD95.

Spawning season analysis within species

For the analysis within species, out of 30 existing Species, 6 could not be used (*Dactylopterus volitans, Lithognathus mormyrus, Pagellus erythrinus, Serranus cabrilla, Umbrina cirrosa, and Xyrichtys novacula*), due to the lack of information to fairly compare the two categories, since they only present one of them (wether in or out the spawning season). Consequently, there are no results regarding these species, and no conclusions could be drawn about the influence of the spawning season on the area used by these species. These species are indicated with NA in Table 3.13.

Regarding the adjusted GLMMs, significant differences in the area occupied by individuals during and outside the spawning season were observed in 16 species, for both home range and core area, even though these differences occurred across different species. This suggests that for some species, the spawning season is important in explaining the space occupied 95% of the time but not 50% of the time, and vice versa. Of the 16 species that showed significant differences, 11 had larger occupied areas during the spawning season, while 5 had larger areas outside the spawning season. This pattern was consistent for both home range and core area, with increases in core area during the spawning season corresponding to increases in home range, and decreases in core area corresponding to decreases in home range. For more details about the models see Table 3.13.

Table 3.13: Coefficient of the variable spawning season (reference level is not in the spawning season) and corresponding standard errors and statistical significance for each of the species specific models. The table on the left corresponds to KUD95 models (home range) and the table on the right corresponds to KUD50 models (core area).

KUD95

	Spawnyes parameter			
Species	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Dactylopterus volitans	NA	NA	NA	NA
Dentex dentex	0.10072	0.01603	6.283	3.33e-10 ***
Dicentrarchus labrax	0.23486	0.03086	7.61	2.75e-14 ***
Diplodus cervinus	-0.25631	0.09128	-2.808	0.00498 **
Diplodus sargus	-0.008657	0.007103	-1.219	0.223
Diplodus vulgaris	-0.26811	0.09239	-2.902	0.00371 **
Epinephelus marginatus	0.04345	0.00498	8.725	<2e-16 ***
Gadus morhua	0.04808	0.01104	4.357	1.32e-05 ***
Labrus bergylta	0.022007	0.003162	6.959	3.42e-12 ***
Lichia amia	0.4826	0.1464	3.296	0.00098 ***
Lithognathus mormyrus	NA	NA	NA	NA
Pagellus erythrinus	NA	NA	NA	NA
Pagrus pagrus	0.04928	0.01805	2.730	0.00634 **
Pomatomus saltatrix	-0.05879	0.13605	-0.432	0.666
Pseudocaranx dentex	0.06570	0.02074	3.168	0.00154 **
Sciaena umbra	-0.06285	0.02144	-2.931	0.00338 **
Scorpaena porcus	-0.04611	0.02097	-2.199	0.0279 *
Scorpaena scrofa	-0.02598	0.01977	-1.31	0.189
Seriola dumerili	0.42055	0.04952	8.493	<2e-16 ***
Seriola rivoliana	0.04774	0.01209	3.949	7.86e-05 ***
Serranus atricauda	-0.002961	0.001762	-1.68	0.0929 .
Serranus cabrilla	NA	NA	NA	NA
Serranus scriba	-0.002848	0.063364	-0.045	0.964
Solea senegalensis	-0.03383	0.03830	-0.883	0.377
Sparisoma cretense	-0.002199	0.010253	-0.214	0.83017
Sparus aurata	0.02017	0.02158	0.935	0.35
Sphyraena viridensis	0.16774	0.01869	8.974	<2e-16 ***
Spondyliosoma cantharus	-0.05671	0.02129	-2.664	0.00773 **
Umbrina cirrosa	NA	NA	NA	NA
Xyrichtys novacula	NA	NA	NA	NA

KUD50

	Spawnyes parameter			
Species	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Dactylopterus volitans	NA	NA	NA	NA
Dentex dentex	0.05748	0.01527	3.765	0.000166 ***
Dicentrarchus labrax	0.19395	0.02920	6.64	3.1e-11 ***
Diplodus cervinus	-0.13142	0.06793	-1.935	0.053 .
Diplodus sargus	0.009822	0.006760	1.45	0.146
Diplodus vulgaris	-0.13813	0.06039	-2.287	0.0222 *
Epinephelus marginatus	0.028028	0.004024	6.96	3.3e-12 ***
Gadus morhua	0.04594	0.01107	4.15	3.31e-05 ***
Labrus bergylta	0.024741	0.003492	07.08	1.39e-12 ***
Lichia amia	0.2794	0.1782	1.567	0.117030
Lithognathus mormyrus	NA	NA	NA	NA
Pagellus erythrinus	NA	NA	NA	NA
Pagrus pagrus	0.03840	0.01673	2.294	0.0218 *
Pomatomus saltatrix	-0.06845	0.13018	-0.526	0.599
Pseudocaranx dentex	0.07720	0.01681	4.592	4.4e-06 ***
Sciaena umbra	-0.04746	0.01826	-2.6	0.00933 **
Scorpaena porcus	-0.05518	0.02466	-2.24	0.0253 *
Scorpaena scrofa	-0.02598	0.01977	-1.31	0.189
Seriola dumerili	0.42232	0.04948	8.535	<2e-16 ***
Seriola rivoliana	0.028314	0.009618	2.94	0.00324 **
Serranus atricauda	-0.004663	0.001812	-2.6	0.0101 *
Serranus cabrilla	NA	NA	NA	NA
Serranus scriba	0.06062	0.07423	0.817	0.414
Solea senegalensis	-0.03657	0.03680	-0.994	0.32
Sparisoma cretense	0.004549	0.009085	0.501	0.617
Sparus aurata	0.04739	0.02160	2.194	0.0282 *
Sphyraena viridensis	0.09604	0.01525	6.297	3.04e-10 ***
Spondyliosoma cantharus	-0.05414	0.02015	-2.687	0.00721 **
Umbrina cirrosa	NA	NA	NA	NA
Xyrichtys novacula	NA	NA	NA	NA

3.4 Discussion

In this study, we worked with a dataset of moderate size that presented some challenges due to its high variability. The dataset included a large number of discrepant observations. This variability could potentially be addressed by transforming the variables or removing outliers (values furthest from normality). However, these approaches were not successful. Transformations applied to both the response and explanatory variables resulted in negligible changes, with the data remaining highly dispersed and the distribution pattern remaining non-linear. With regard to outliers, it was anticipated that their exclusion would enhance the alignment of the data with the revised parameters, resulting in a distribution that more closely approximates normality. However, after each removal, the distribution readjusted with new outliers emerging. This cycle repeated multiple times. The persistence of high variability, lack of linearity, and presence of outliers, despite transformations and outlier removal, suggested that a distribution other than the Gaussian might be required to model the data. Therefore, the outliers might not indicate errors but rather values that were distant from the recurring values, possibly containing relevant biological information. Consequently, these outliers should not be, and were not, removed from the dataset. In this study, the KUD values farthest from the average simply imply the presence of individuals that use considerably larger areas on a weekly basis compared to most. This is pervasive in biological and ecological data.

The results of the correlation analysis between the explanatory variables showed that the only variables with high correlation between them were length and body mass. Therefore, to avoid multicollinearity and unreliable parameter estimates, one of these variables should typically be removed in the model study. However, not all individuals follow this pattern, since there were individuals with the same length but different body mass. In studies of animals' movement, both length and body mass must be considered, as individuals with the same length can have different body masses and therefore exhibit different movement patterns. Consequently, despite the collinearity between these two variables, we decided to include them both in the models to account for all relevant biological traits.

Regarding the study design parameters, we chose to include only the monitored area and the density of receivers while excluding the MCP, the number of receivers, and the maximum distance between receivers. The MCP and the number of receivers were excluded because the density of receivers, calculated using these two metrics, generally provides a more accurate measure for studies of the area used by individuals due to its greater detail of spatial coverage [Millspaugh and Marzluff, 2001]. Additionally, the number of receivers showed a very strong correlation with the monitored area. The maximum distance between receivers was also excluded due to its high correlation with the area monitored.

In terms of modelling, since there did not appear to be linear relationships between the response and explanatory variables, various models were tested, including generalised linear and additive models and corresponding mixed models. Generalised models were implemented because the data did not meet certain assumptions of linear models. Additionally, it was hypothesized that the data belonging to the same individual (Transmitter), species (Species), or species per study (File) might show some degree of

dependence. Therefore, mixed models were also implemented to account for this potential dependence. In GLM, GAM, GLMM and GAMM we decided to use the family Gamma, since the response variables were positive, continuous and asymmetric. We chose to use the logarithmic link function, instead of the inverse, even though this was the canonical function for the Gamma distribution, for ease of interpretation and which avoids inadmissible (e.g. negative) predictions.

We initially expected the GAMM to be the best fitting regression model due to its ability to represent non-linear relations. However, the results proved us wrong, and the GLMM emerged as the best fit. This was likely because the relationship between the explanatory variables and the response was approximately linear on the scale of the link function . In contrast, the GAMM may have overfitted the data due to its flexibility in capturing non-linear patterns that were spurious.

No statistical differences were found between models using File and Species as random effects. This suggests that treating the same species from different geographical locations as distinct species yielded similar results to treating them as the same species. Therefore, any biological differences that may exist that lead to different needs and movement patterns within the same species across locations are not statistically significant enough to warrant treating them as different species. In this study we only focused on individuals living in Europe, but it would be interesting to investigate whether individuals of a particular species in Europe differ from individuals of the same species on another continent in terms of movement patterns. In addition, and more generally, research could be carried out to try to understand whether the movement patterns of teleost fishes found in this study depend on geographical area, comparing the movement patterns of fishes from Europe with the movement patterns of fishes from other continental coasts.

Regarding the significant variables in explaining the variability of KUDs alone, both trophic level, habitat and migration supported our initial hypotheses. As expected, spatial extent generally increases with trophic level, with individuals at higher levels of the trophic chain moving more to have access to more food and meet their high energy needs. The analysis revealed that the space used was largest for pelagic individuals, followed by benthopelagic and demersal ones. Ecologically, this can be attributed to the higher abundance and concentration of organic matter on the seabed compared to the pelagic zone. The accumulation of detritus from dead organisms in the sediment allows demersal species to avoid extensive movements and remain more sedentary. Additionally, the space use seems to be larger for individuals how perform migrations comparing to sedentary individuals.

In the independent analysis of factors significantly related with the variability of KUDs, both core area and home range were explained by the same set of variables. This consistency suggests that the factors influencing the spatial range of individuals may be stable across different scales of habitat use. Specifically, the traits that determine an individual's core area (KUD50), or the central activity area, are the same traits that influence their overall home range (KUD95), including peripheral areas used less frequently. From an ecological point of view, this finding has important implications for understanding animal movement and habitat use. If the same variables are significant at both the core area and home range scales, it indicates that the fundamental ecological and biological processes driving space use are

consistent across different spatial extents. It would be interesting to look more closely at how different predictor variables relate to different levels of space use. For example, quantile regression analysis could be used to see how different characteristics affect different quantiles, rather than just the 50^{th} (core area) and 95^{th} (home range) quantiles.

However, when modeling KUDs with multiple predictor variables, we found that the significant variables differed between home range and core area. This can be attributed to the interactions between variables when analyzed together, which can alter their individual effects. As a result, some variables may have gained or lost significance due to these interactions. For home range, habitat and commercial importance proved to be relevant to account for variability and, even though habitat was already a significant variable, this model had a better performance and was statistically different from the one with only commercial importance. Although commercial importance alone did not significantly explain the variability in home range size, its combination with habitat emerged as a significant factor. This indicates a complementary relationship between the two variables, where habitat and commercial importance together provide a more robust explanation of home range variability. Here, home range was largest for pelagic individuals of high commercial importance compared to individuals of medium commercial importance. However, there seemed to be no differences between species of high and low commercial importance. This contradicts our initial hypothesis based on Gandra et al. [2021], which suggested that individuals with higher commercial importance tended to have lower genetic differentiation, and therefore, greater dispersion. Because of these results we cannot say that space extent increases with commercial importance. Further investigation is needed to explore these complex interactions and better understand the drivers of spatial use in relation to commercial importance. This is not necessarily unexpected. Commercial valued species can have sets of common biological features, but certainly there will be commercial species that are all over, while some will be quite restricted in space. So it is likely that the existing effects might be much more nuanced than a simple binary variable would allow to represent. For core area, length and habitat proved to be relevant to account for variability and, again, even though habitat was already a significant variable, this model had a better performance and was statistically different from the one with only habitat. Here, core area was largest for pelagic individuals with bigger lengths. Many studies claim that spatial extent is larger for individuals with higher body mass [Nash et al., 2015; Hendriks, 2007; Lindstedt et al., 1986; Turner et al., 1969], but we found out evidences that suggest that length is a better factor to account for variability of home range than body mass. However, given that body mass was estimated using a formula and that these values may not have been entirely accurate and may have confounded the results. Perhaps the results would have been different if body mass had been measured directly from the individuals.

Regarding the final models with more than one predictor, it is true that the residual analysis did not show a good fit of the data, revealing overdispersion, heteroscedasticity and a non-uniform residuals distribution. However, given the high dispersion in the observations, it is expected that no model will capture all variations in the data. Nonetheless, the comparison between simulations and actual values demonstrated that the model effectively captured the general patterns of individual distribution. It was

particularly adept at identifying broad trends, although it struggled to account for the variability of individuals that deviated from normality and had larger distribution areas than most. For this reason, we believe that the GLMM can be accepted as a valid model for explaining the overall variability in the data and drawing general conclusions about the variables influencing space extent patterns.

For the spawning season analysis, 6 Species were excluded because they were assigned to only one of the categories of the ”Spawn” variable, meaning they were exclusively either within or outside the spawning season, not leaving a suitable comparison term for these categories in those Species.

For some species, the utilisation area seemed to increase significantly when individuals were in the spawning season, while for others it seemed to decrease. This may be related to the different reproductive strategies of each species. Some species may expand their utilisation area during the spawning season to enhance distribution of eggs and larvae and increase survival probability [Robichaud and Rose, 2004], find suitable spawning sites [van Leeuwen et al., 2023] or explore areas with better conditions for the survival of their young. Other species, however, may reduce their area of use to concentrate on specific areas that offer optimal conditions for spawning [Ellis et al., 2012], such as protection from predators or favourable environmental conditions. There is also evidences that some species use both strategies, such as *Sparisoma cretense*, as seen in Afonso et al. [2008b]. Therefore, it makes sense that there are hardly any significant differences in the values of the area occupied during and out the spawning period.

For future studies on this subject there are some recommendations to improve the quality of the analysis. Our analysis categorised species into only three broad groups based on spawning season, which may have limited our ability to detect more specific patterns. A more detailed analysis, considering the exact spawning months for each species, could provide more accurate insights and reveal additional patterns. However, obtaining detailed data on spawning periods is challenging, as these periods can vary widely depending on geographic and environmental factors. The lack of precise information on reproductive and spawning periods of species contributes to the difficulty in conducting a more granular analysis. Therefore, it is recommended that future research seek to obtain more detailed data on spawning seasons to enable a more refined and accurate analysis. A complete understanding of spawning patterns and their influence on area use patterns may require more extensive data collection and more complex modelling, taking into account spatial and temporal variations in the spawning periods of different species.

Overall, there are important concerns to address in acoustic telemetry studies, as this technique has limitations that can affect the accuracy of estimated areas used by individuals. Overestimation or underestimation of usage areas often occurs, depending on the circumstances [Kessel et al., 2014]. For instance, if an individual temporarily moves outside the detection range of the array [Abecasis, 2008], its area of use will be underestimated. This likely resulted in unrealistic and underestimated KUDs in our dataset, particularly in cases where the monitored areas were very small, such as ArrayID 13, which covered only 0.69 km² (see Table 3.2). To minimise this issue, it’s essential to carefully design array sizes to ensure they are large enough to detect the individual consistently and avoid undetected exits. On the other hand, overestimation can occur, among others, if an individual remains within a relatively small location, but the analysis software interprets the entire area covered by a wide detection network as the area used, even

though the actual usage is much more confined. That is why positioning systems like COAs are used, to refine the spatial resolution of movement data and improve space use positions.

3.5 What I have learnt in the process

During the course of this research, I gained significant knowledge in the field of acoustic telemetry, particularly its applications in tracking fish movement and understanding habitat use. This process introduced me to the wider scientific community involved in telemetry-based research, as well as the different methodologies used to study spatial utilisation in fish populations. I also learnt how raw telemetry data are transformed into spatial metrics, in particular the estimation of utilisation areas using techniques such as kernel density estimation, and how area scale can be derived using quantiles.

Furthermore, I deepened my understanding of statistical modelling, including linear models (LM), generalised linear models (GLM), generalised additive models (GAM), and mixed-effects models.

This experience sharpened my critical thinking about statistical analysis, allowing me to discern which methods are appropriate and meaningful given the nature of the data. Throughout this project, I deliberately experimented with different modelling approaches because my primary goal was to learn by doing. I wanted to practice different methods even when simpler alternatives would have sufficed. For example, when I tried to remove outliers to see if the data could meet the assumptions of a linear regression model, I realised that I could have just used a generalised model. In biological data, it is common to encounter values that deviate from normality, and these deviations do not necessarily indicate measurement error or something abnormal. There is just intrinsically higher variability and in particular heavier tails in ecological processes. However, as it is common practice in statistics to evaluate and deal with outliers, I continued with the outlier analysis to ensure a comprehensive evaluation of the dataset. By the end of the project, I had developed enough awareness and expertise to confidently determine which methods were useful and which were not.

Additionally, I enhanced my technical skills and knowledge in tools such as Overleaf, LaTeX, BibTeX, R, and GitHub, which have been essential in improving the efficiency and quality of my workflow throughout the research process.

Finally, I realised the importance of regular meetings with a team or research group. These collaborative discussions are invaluable for exchanging ideas, accelerating scientific progress, and setting new goals, often leading to more efficient and insightful research outcomes.

3.6 Conclusion

This study provides essential insights into what influences the movement patterns of teleost fish across multiple species. By analysing traits such as length, body mass, vulnerability, longevity, trophic level, habitat, migration, and commercial importance, we identified key factors that help predict utilisation areas.

Trophic level, habitat, and migration emerged as significant when analysed alone. We confirmed that individuals in the higher trophic levels generally occupy larger areas, as well as pelagic and migratory individuals. The retention of the same variables in both KUD95 and KUD50 models using only one predictor underscores the consistency of the factors influencing animal spatial behaviour. This consistency aids in developing effective conservation and management strategies, as it highlights the key traits to consider when assessing habitat needs and space use patterns. Moreover, it simplifies the predictive modelling process, allowing for more straightforward and reliable assessments of animal movement across different spatial scales.

The most effective model for home range included however habitat and commercial importance together. The inclusion of commercial importance in the most effective model may suggest that human activities significantly shape the size of the areas used by fish, but it is also possible that this finding is only coincidental, or that fishes targeted commercially have specific habitat usage patterns. Further studies are necessary to thoroughly investigate the relationship, that is unlikely to be simple, between commercial importance and space use. The most effective model for core area included length and habitat together. These findings highlight the need to consider various biological traits together in ecological studies, to unveil the optimal conditions for fish utilisation areas.

While no significant differences were found in movement patterns among the same species across different geographical locations within Europe, future research should explore whether these patterns vary on a broader scale, comparing teleost fishes across different continents to better understand the influence of geography on movement behaviour.

Regarding the analysis of the spawning season, most species showed an increase in area use during this period. However, some species showed a decrease in area use, while for others the spawning season was not a significant factor in explaining spatial use. This variation is expected as different species have different reproductive strategies, each requiring specific conditions for spawning and rearing offspring. Some species occupy large areas, others smaller ones, while some species exhibit both behaviours during the spawning season.

This research represents a critical step in understanding the distribution patterns of multiple marine species and their interactions within their habitats. By shedding light on how different species utilise space, these findings provide valuable insights that can directly inform the management and conservation of marine ecosystems. However, there is still much to uncover. Future studies should aim to refine these results by delving deeper into additional factors such as temporal dynamics, which capture changes over time, and the influence of human activities, which often disrupt natural behaviours and habitat use.

These advancements are particularly relevant for the designation and management of Marine Protected Areas. By integrating refined spatial and behavioural data, MPAs can be better positioned and scaled to ensure the effective protection of critical habitats, such as spawning grounds or migratory routes. This evidence-based approach would maximise the ecological benefits of MPAs, supporting the long-term sustainability of marine populations and fostering resilience against ongoing challenges like climate change and habitat degradation.

In conclusion, while this study lays an important foundation, the journey towards comprehensive understanding and optimal conservation of marine species demands continued research, collaborative efforts, and the integration of innovative methodologies. The existence of collaborative networks promotes more rapid development of research and study of different topics, which would not be possible without the vast datasets generated by the contribution of different researchers around the world. Together, these efforts will ensure that conservation initiatives are grounded in robust scientific evidence, enabling them to meet the needs of marine biodiversity in an increasingly complex and changing world.

References

Abecasis, D. (2008). Aplicação de marcação convencional e telemetria no estudo dos movimentos de quatro espécies de esparídeos na ria formosa. Master's thesis, Universidade do Algarve. 6, 7, 45

Abecasis, D., Steckenreuter, A., Reubens, J., Aarestrup, K., Alós, J., Badalamenti, F., Bajona, L., Boylan, P., Deneudt, K., Greenberg, L., Brevé, N., Hernández, F., Humphries, N., Meyer, C., Sims, D., Thorstad, E. B., Walker, A. M., Whoriskey, F., and Afonso, P. (2018). A review of acoustic telemetry in europe and the need for a regional aquatic telemetry network. *Animal Biotelemetry*, 6(12). 5, 7

Afonso, A. S., Macena, B. C. L., Mourato, B., Bezerra, N. P. A., Mendonça, S., Queiroz, J., and Hazin, F. (2022). Trophic-mediated pelagic habitat structuring and partitioning by sympatric elasmobranchs. *Frontiers in Marine Science*, 9. 2

Afonso, P., Fontes, J., Holland, K. N., and Santos, R. S. (2008a). Social status determines behaviour and habitat usage in a temperate parrotfish: implications for marine reserve design. *Mar. Ecol. Prog. Ser.*, 359:215–227. 2

Afonso, P., Morato, T., and Santos, R. (2008b). Spatial patterns in reproductive traits of the temperate parrotfish *Sparisoma cretense*. *Fisheries Research*, 90(1-3). 45

Allen, J. (n.d.). Use of coded transmitter schemes to overcome radio frequency spectrum constraints in terrestrial wildlife tracking. Accessed: 2024-03-16. 6

Baktoft, H., Gjelland, K., Økland, F., and Thygesen, U. (2017). Positioning of aquatic animals based on time-of-arrival and random walk models using yaps (yet another positioning solver). *Sci Rep*, 7(14294). 6

Bevans, R. (2020). Akaike information criterion | when how to use it (example). Accessed: 2024-16-04. 18

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400. 27

Burt, W. H. (1943). Territoriality and home range concepts as applied to mammals. *Journal of Mammalogy*, 24(3):346–352. III, 1, 8

Calenge, C. (2006). The package adehabitat for the r software: tool for the analysis of space and habitat use by animals. *Ecological Modelling*, 197:1035. 25

Casson, R. J. and Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clinical Experimental Ophthalmology*, 42(6):590–596. 10

Chen, Y., Moustaki, I., and Zhang, H. (2020). A note on likelihood ratio tests for models with latent variables. *Psychometrika*, 85(4):996–1012. 17

Cholaquidis, A., Fraiman, R., and Hernandez-Banadik, M. (2023). Home range estimation under a restricted sampling scheme. *Journal of Nonparametric Statistics*, pages 1–20. 8

Crossin, G. T., Heupel, M. R., Holbrook, C. M., Hussey, N. E., Lowerre-Barbieri, S. K., and Nguyen, V. M. ... Cooke, S. J. (2017). Acoustic telemetry and fisheries management. *Ecological Applications*, 27(4):1031–1049. 3, 5, 6, 7

Das, K. R. and Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, 5(1):5–12. 11

Ellis, J. R., Milligan, S., Readdy, L., N., T., and M.J., B. (2012). Spawning and nursery grounds of selected fish species in uk waters. Technical report, Cefas Lowestoft. 45

Froese, R. and Pauly, D. (2024). Fishbase. world wide web electronic publication. Accessed: 2024-02-27. 21

Froese, R., Thorson, J. T., and Reyes, R. B. (2014). A bayesian approach for estimating length-weight relationships in fishes. *Journal of Applied Ichthyology*, 30(1):78–85. 21

Gandra, M., Assis, J., Martins, M., and Abecasis, D. (2021). Reduced global genetic differentiation of exploited marine fish species. *Molecular Biology and Evolution*, 38(4):1402–1412. 2, 44

Gelman, A., . H. J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. 10

Goossens, J., Buyse, J., Bruneel, S., Verhelst, P., Goethals, P., Torreele, E., Moens, T., and Reubens, J. (2022). Taking the time for range testing: An approach to account for temporal resolution in acoustic telemetry detection range assessments. *Animal Biotelemetry*, 10(17). 7

Grego, J. M. (2012). *Generalized Additive Models*. Wiley, 1 edition. 14

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.6. 27

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New York, NY. 10

Hellström, G., Lennox, R. J., Bertram, M. G., and Brodin, T. (2022). Acoustic telemetry. *Current Biology*, 32(16). 7

Hendriks, A. J. (2007). The power of size: A meta-analysis reveals consistency of allometric regressions. *Ecological Modelling*, 205(1-2):196–208. 2, 44

Hussey, N. E., Kessel, S. T., Aarestrup, K., Cooke, S. J., Cowley, P. D., Fisk, A. T., Harcourt, R. G., Holland, K. N., Iverson, S. J., Kocik, J. F., Mills Flemming, J. E., and Whoriskey, F. G. (2015). Aquatic animal telemetry: A panoramic window into the underwater world. *Science*, 348. 5

Huveneers, C., , Simpfendorfer, C. A., Kim, S., Semmens, J. M., Hobday, A. J., Pederson, H., Stieglitz, T., Vallee, R., Webber, D., Heupel, M. R., V., P., and Harcourt, R. G. (2016). The influence of environmental parameters on the performance and detection range of acoustic receivers. *Methods in Ecology and Evolution*, 7(7):825–835. 7

INNOVASEA (2021). Usgs conducts fine-scale positioning study in kentucky and analyzes results using fathom position. Accessed: 2024-06-28. 6

Kessel, S. T., Cooke, S. J., Heupel, M. R., Hussey, N. E., Simpfendorfer, C. A., Vagle, S., and Fisk, A. T. (2014). A review of detection range testing in aquatic passive acoustic telemetry studies. *Reviews in Fish Biology and Fisheries*, 24:199–218. 7, 45

Kie, J. G., Bowyer, R., Nicholson, M., Boroski, B., and Loft, E. (2002). Landscape heterogeneity at different scales: effects on spatial distribution of mule deer. *Ecology*, 83(2):530–544. 3

Kraft, S., Gandra, M., Lennox, R. J., Mourier, J., Winkler, A. C., and Abecasis, D. (2023). Residency and space use estimation methods based on passive acoustic telemetry data. *Movement Ecology*, 11(12). 9

Kropil, R., Smolko, P., and Garaj, P. (2015). Home range and migration patterns of male red deer *Cervus elaphus* in western carpathians. *Eur J Wildl Res*, 61:63–72. 2, 3

Laver, P. N. and Kelly, M. J. (2008). A critical review of home range studies. *The Journal of Wildlife Management*, 72(1):290–298. 8, 9

Lindstedt, L. L., Miller, B. J., and Buskirk, S. W. (1986). Home range, time, and body size in mammals. *Ecology*, 67(2):413–418. III, 2, 3, 44

Manikandan, S. (2010). Data transformation. *Journal of Pharmacology and Pharmacotherapeutics*, 1(2). 12

Millspaugh, J. J. and Marzluff, J. M. (2001). *Radio Tracking and Animal Populations*. Academic Press. 42

Millspaugh, J. J., Nielson, R. M., McDONALD, L., Marzluff, J. M., Gitzen, R. A., Rittenhouse, C. D., Hubbard, M. W., and Sheriff, S. L. (2006). Analysis of resource selection using utilization distributions. *Journal of Wildlife Management*, 70(2):384–395. 9

Myers, R. H. and Montgomery, D. C. (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, 29(3):274–291. 13

Nash, K. L., Welsh, J. Q., Graham, N. A. J., and Bellwood, D. R. (2015). Home-range allometry in coral reef fishes: comparison to other vertebrates, methodological issues and management implications. *Oecologia*, 177:73–83. III, IV, 2, 3, 44

Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., and Smouse, P. (2008). A movement ecology paradigm for unifying organismal movement research. *Proc Natl Acad Sci USA*, 105(49):19052–19059. 1

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384. 13

NUMXL (2016). Kernel density estimation (kde) plot. Accessed: 2024-02-12. 8

Patefield, W. M. (1977). On the maximized likelihood function. *The Indian Journal of Statistics*, 39(1):92–96. 16

Pedersen, E. J., Miller, D. L., Simpson, G. L., and Ross, N. (2019). Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, 7:e6876. 14

Poole, M. A. and O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, (52):145–158. 12

Posit team (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. 26

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 27

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 26

Reine, K. J. (2005). Tagging and tracking technologies for freshwater and marine fishes. Technical report, ENGINEER RESEARCH AND DEVELOPMENT CENTER VICKSBURG MS. 5, 6

Reubens, J., Verhelst, P., van der Knaap, I., Deneudt, K., Moens, T., and Hernandez, F. (2019). Environmental factors influence the detection probability in acoustic telemetry in a marine environment: results from a new setup. *Hydrobiologia*, 845:81–94. 7

Robichaud, D. and Rose, G. (2004). Migratory behaviour and range in atlantic cod: Inference from a century of tagging. *Fish and Fisheries*, 5(3). 45

Schmidt, A. F. and Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98:146–151. 10, 11

Seaman, E., P. R. A. (1996). An evaluation of the accuracy of kernel density estimators for home range analysis. *Ecology*, 77(7):2075–2085. 9

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall. 9

Simpfendorfer, C., Heupel, M., and Hueter, R. (2002). Estimation of short-term centers of activity from an array of omnidirectional hydrophones and its use in studying animal movements. *Canadian Journal of Fisheries and Aquatic Sciences*, 59(1):23–32. 19

Simpfendorfer, C. A., H. M. R. (2012). *Assessing Habitat Use and Movement*. CRC Press, 2 edition. III, 2

Topping, D., Lowe, C., and Caselle, J. (2005). Home range and habitat utilization of adult california sheephead, *Semicossyphus pulcher* (labridae), in a temperate no-take marine reserve. *Marine Biology*, 147:301–311. 2

Turner, F. B., Jennrich, R. I., and Weintraub, J. D. (1969). Home ranges and body size of lizards. *Ecology*, 50(6):1076–1081. III, 2, 44

Udyawer, V., Huveneers, C., Jaine, F., Babcock, R., Brodie, S., and Buscot, M.J. ... Heupel, M. (2023). Scaling of activity space in marine organisms across latitudinal gradients. *The American Naturalist*, 201(4):586–602. III, IV, 3

van Leeuwen, C., de Leeuw, J., and van Keeken, O. (2023). Multispecies fish tracking across newly created shallow and deep habitats in a forward-restored lake. *Movement Ecology*, 11(43). 45

VEMCO (2014). Range testing introduction. Video. Accessed: 2024-05-11. 7

Werf, J. (n.d.). Introduction to mixed models. Accessed: 2024-04-14. 15

Whoriskey, K., Baktoft, H., Field, C., L. R. J., Babyn, J., Lawler, E., and Mills Flemming, J. (2022). Predicting aquatic animal movements and behavioural states from acoustic telemetry arrays. *Methods in Ecology and Evolution*, 13(5):987–1000. 6

Winkle, W. V. (1975). Comparison of several probabilistic home-range models. *The Journal of Wildlife Management*, 39(1):118–123. 8

Winter, J. (1983). *Underwater Biotelemetry*. American Fisheries Society, Bethesda, MD, 1 edition. 6

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36. 27

Woolf, B. (1957). The log likelihood ratio tes (the g-test); methods and tables for tests of heterogeneity in contingency tables. *Annals of human genetics*, 21(4):397–409. 18

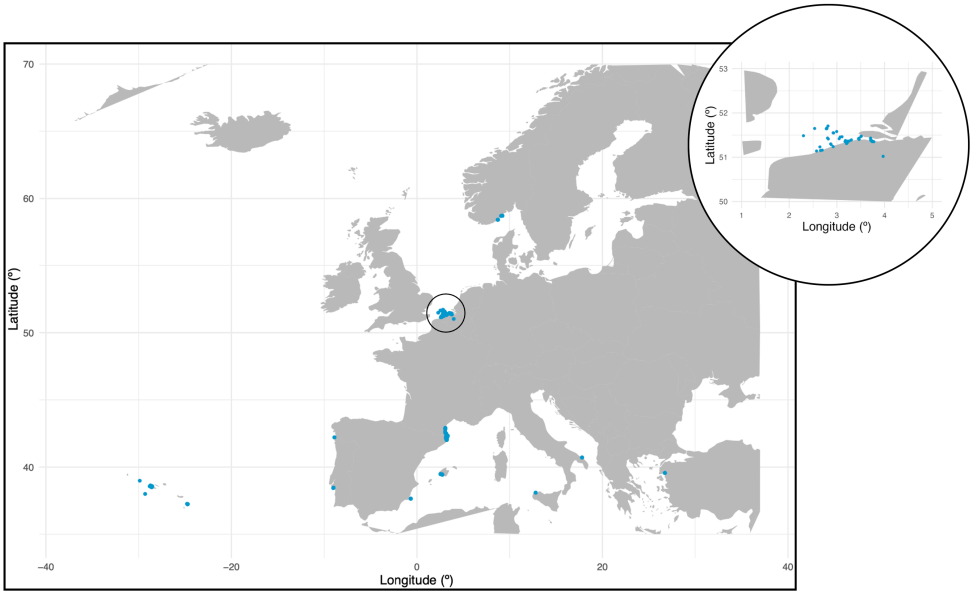
Worton, B. J. (1975). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70(1):164–168. 9

Yee, T. W. and Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science*, 2(5):587–602. 15

Zelmer, D. (n.d.). Data transformation. Accessed: 2024-01-23. 12

Appendices

Study Area



Appendix 1: Map of the study area. Each blue dot corresponds to a receiving station where transmitter signals were detected.

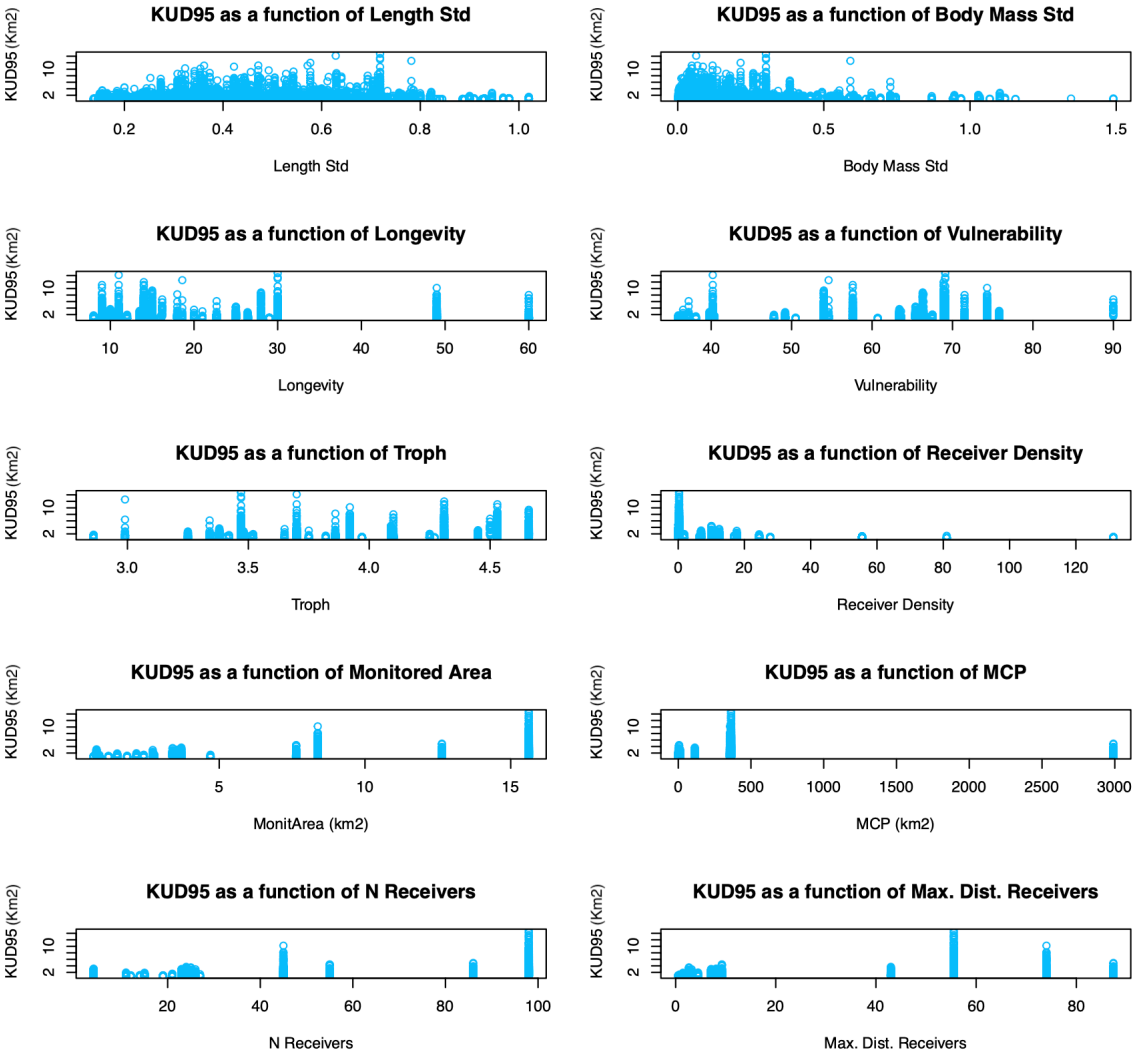
Appendix 2: Summary table of species from which KUDs were estimated. The table contains five columns, including the file variable, the name of the species, the reference of the study that collected the data, the location where the study took place, the number of individuals of each species, and the number of estimated KUDs.

File	Species	Reference	Location	N Individuals	No. of estimated KUDs
Dactylopterus_volitans	*Dactylopterus volitans*	Bernat, unpublished	NE Med Spain	1	16
Dentex_dentex1	*Dentex dentex*	Aspillaga et al. (2017)	SE Spain	19	778
Dentex_dentex2	*Dentex dentex*	Bernat, unpublished	NE Med Spain	16	599
Dicentrarchus_labrax1	*Dicentrarchus labrax*	Jolien, unpublished	Belgium	93	854
Dicentrarchus_labrax2	*Dicentrarchus labrax*	Bernat, unpublished	NE Med Spain	28	637
Diplodus_cervinus	*Diplodus cervinus*	Bernat, unpublished	NE Med Spain	4	94
Diplodus_sargus1	*Diplodus sargus*	Abecasis et al. (2015)	Portugal	16	351
Diplodus_sargus2	*Diplodus sargus*	Di Lorenzo et al. (2016)	Italy	20	660
Diplodus_sargus3	*Diplodus sargus*	Giacalone et al. (2018)	Italy	4	80
Diplodus_sargus4	*Diplodus sargus*	Aspillaga et al. (2016)	SE Spain	41	1474
Diplodus_sargus5	*Diplodus sargus*	Koeck et al. (2013)	SW France	73	1102
Diplodus_sargus6	*Diplodus sargus*	Bernat, unpublished	NE Med Spain	6	41
Diplodus_vulgaris1	*Diplodus vulgaris*	Alós et al. (2012b)	Balearic Islands	9	46
Diplodus_vulgaris2	*Diplodus vulgaris*	Bernat, unpublished	NE Med Spain	2	4
Epinephelus_marginatus1	*Epinephelus marginatus*	Afonso et al. (2016)	Azores	11	2055
Epinephelus_marginatus2	*Epinephelus marginatus*	Jose Pereñiguez, unpublished	S Spain	16	437
Epinephelus_marginatus3	*Epinephelus marginatus*	Koeck et al. (2014)	SW France	4	227
Epinephelus_marginatus4	*Epinephelus marginatus*	Bernat, unpublished	NE Med Spain	17	293
Gadus_morhua1	*Gadus morhua*	Olsen and Molan (2011)	Norway	60	1635
Gadus_morhua2	*Gadus morhua*	Martin-Kim, unpublished	Norway	56	1136
Gadus_morhua3	*Gadus morhua*	Jan Reubens et al. (2013)	Belgium	29	399
Labrus_bergylta	*Labrus bergylta*	Villegas-Rios et al. (2013)	NW Spain	25	793
Lichia_amia	*Lichia amia*	Bernat, unpublished	NE Med Spain	1	28
Lithognathus_mormyrus	*Lithognathus mormyrus*	Bernat, unpublished	NE Med Spain	2	8
Pagellus_erythrinus	*Pagellus erythrinus*	Bernat, unpublished	NE Med Spain	6	59
Pagrus_pagrus1	*Pagrus pagrus*	Afonso et al. (2009b)	Azores	20	618
Pagrus_pagrus2	*Pagrus pagrus*	Bernat, unpublished	NE Med Spain	5	36
Pomatomus_saltatrix	*Pomatomus saltatrix*	Bernat, unpublished	NE Med Spain	8	159
Pseudocaranx_dentex	*Pseudocaranx dentex*	Afonso et al. (2009)	Azores	31	1527
Sciaena_umbra1	*Sciaena umbra*	Özgül	Turkey	15	129
Sciaena_umbra2	*Sciaena umbra*	Bernat, unpublished	NE Med Spain	1	14
Scorpaena_porcus	*Scorpaena porcus*	Özgül et al. (2019)	Turkey	13	52
Scorpaena_scrofa1	*Scorpaena scrofa*	Özgül et al. (2019)	Turkey	7	58
Scorpaena_scrofa2	*Scorpaena scrofa*	Bernat, unpublished	NE Med Spain	11	508
Seriola_dumerili	*Seriola dumerili*	Bernat, unpublished	NE Med Spain	8	356
Seriola_rivoliana	*Seriola rivoliana*	Fontes et al. (2014)	Azores	16	2787
Serranus_atricauda	*Serranus atricauda*	Afonso et al. (2016)	Azores	9	650
Serranus_cabrilla	*Serranus cabrilla*	Alós et al. (2011)	Balearic Islands	15	54
Serranus_scriba	*Serranus scriba*	March et al. (2010)	Balearic Islands	10	27
Solea_senegalensis	*Solea senegalensis*	Abecasis et al. (2014)	Portugal	22	237
Sparisoma_cretense	*Sparisoma cretense*	Afonso et al. (2008)	Azores	10	769
Sparus_aurata1	*Sparus aurata*	Özgül	Turkey	7	127
Sparus_aurata2	*Sparus aurata*	Bernat, unpublished	NE Med Spain	43	1133
Sphyraena_viridensis1	*Sphyraena viridensis*	Fontes and Afonso (2017)	Azores	13	1298
Sphyraena_viridensis2	*Sphyraena viridensis*	Bernat, unpublished	NE Med Spain	17	558
Spondyliosoma_cantharus	*Spondyliosoma cantharus*	Bernat, unpublished	NE Med Spain	21	663
Umbrina_cirrosa	*Umbrina cirrosa*	Bernat, unpublished	NE Med Spain	1	2
Xyrichtys_novacula	*Xyrichtys novacula*	Alós et al. (2012a)	Balearic Islands	12	44

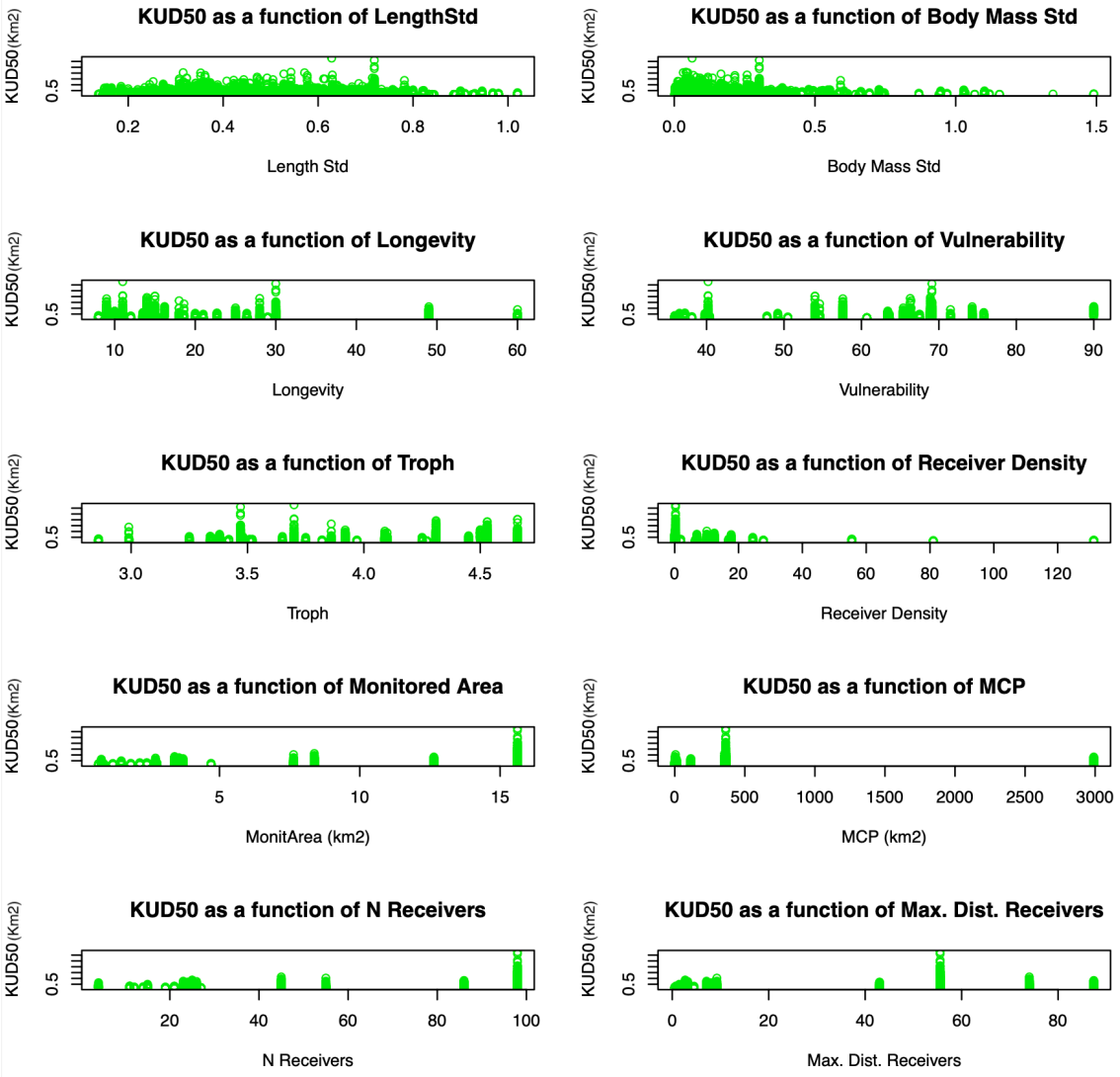
Statistical Analysis

Appendix 3: Example of the KUDs table. In this table are presented some species of different studies/sites, for which it is shown the KUD50 and KUD95 values and even the corresponding week (xx/yyyy - xx is the week of the yyyy year).

File	Species	Transmitter	KUD50	KUD95	Week
Dactylopterus_volitans	Dvol	A69-9006-11437	0.216	1.079	09/2022
Dactylopterus_volitans	Dvol	A69-9006-11437	0.180	0.839	10/2022
Dentex_dentex1	Dden	12	0.189	0.879	21/2007
Dentex_dentex1	Dden	12	0.222	0.990	22/2007
Dentex_dentex1	Dden	18	0.473	1.736	00/2008
Dentex_dentex1	Dden	18	0.490	1.943	01/2008
Dentex_dentex2	Dden	A69-9006-11484	0.360	1.448	24/2021
Dentex_dentex2	Dden	A69-9006-11484	0.324	1.327	25/2021
Dicentrarchus_labrax1	Dlab	A69-9006-3642	0.164	0.761	01/2019
Dicentrarchus_labrax1	Dlab	A69-9006-3642	0.165	0.762	06/2019
Dicentrarchus_labrax2	Dlab	A69-1602-40162	0.162	0.763	00/2021
Dicentrarchus_labrax2	Dlab	A69-1602-40162	1.049	5.602	01/2021
Diplodus_cervinus	Dcer	A69-1602-40706	0.376	1.468	00/2021
Diplodus_cervinus	Dcer	A69-1602-40706	0.345	1.362	01/2021
Diplodus_sargus1	Dsar	Sargo 02	0.290	1.146	19/2011
Diplodus_sargus1	Dsar	Sargo 02	0.279	1.102	20/2011
Diplodus_sargus2	Dsar	A69-1303-40931	0.306	1.219	07/2011
Diplodus_sargus2	Dsar	A69-1303-40931	0.238	1.134	08/2011
Diplodus_sargus3	Dsar	A69-1105-229	0.168	0.777	00/2010
Diplodus_sargus3	Dsar	A69-1105-229	0.164	0.768	01/2010
Diplodus_sargus4	Dsar	13	0.167	0.796	00/2008
Diplodus_sargus4	Dsar	13	0.167	0.793	01/2008
Diplodus_sargus5	Dsar	1511	0.164	0.761	30/2011
Diplodus_sargus5	Dsar	1511	0.163	0.762	33/2011
Diplodus_sargus6	Dsar	A69-9007-16172	0.165	0.762	00/2022
Diplodus_sargus6	Dsar	A69-9007-16172	0.162	0.764	08/2022



Appendix 4: Plots of KUD95 against all explanatory variables.



Appendix 5: Plots of KUD50 against all explanatory variables.

Appendix 6: Variable importance of each variable attributed to the random effects of the GLMM.

Variable Importance

	KUD95	**KUD50**
Transmitter	64	69
File	24	21
Species	12	10
	100	100

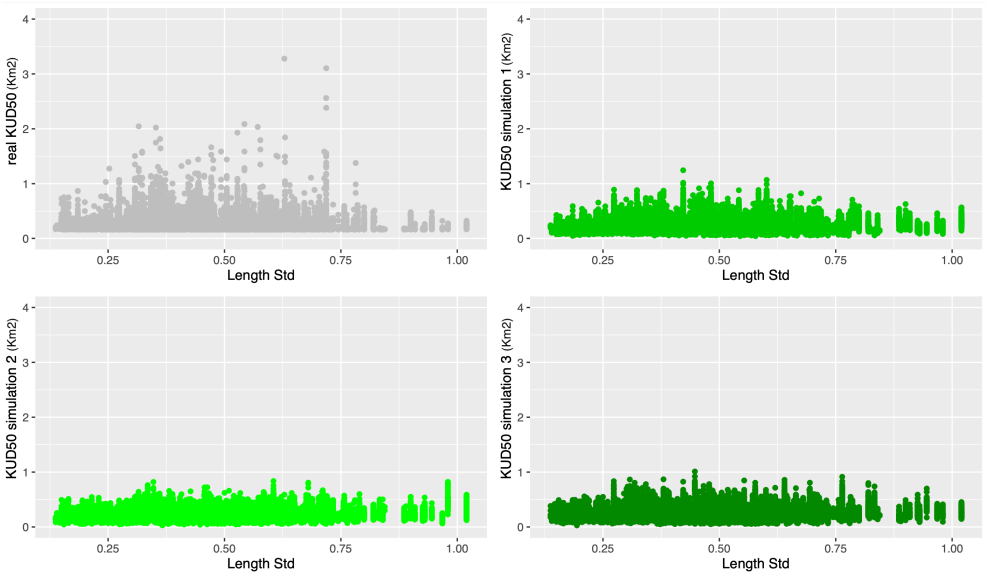
Appendix 7: Backward elimination process for KUD95 and KUD50. Initially, all variables are included in the model. Subsequently, variables are iteratively removed based on their statistical significance. In each step, the variable exhibiting the lowest statistical significance (i.e., the highest p-value) is eliminated from the model. This process continues until all remaining variables in the model demonstrate statistical significance at $\alpha = 0.05$.

Backward Elimination KUD95

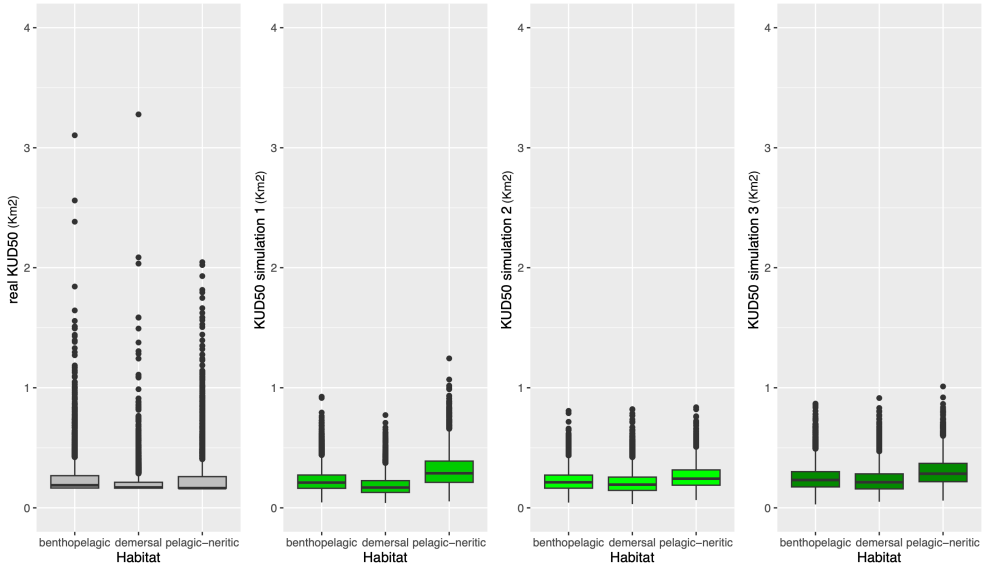
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.1752678	0.3500228	-0.501	0.6166
LengthStd	0.2872585	0.1560424	1.841	0.0656 .
BodyMassStd	-0.0693494	0.1132226	-0.613	0.5402
Longevity	-0.0020923	0.0033787	-0.619	0.5357
Vulnerability	-0.0050757	0.0041502	-1.223	0.2213
Troph	0.0774655	0.1068172	0.725	0.4683
Habitatdemersal	-0.0831834	0.0980336	-0.849	0.3961
Habitatpelagic-neritic	0.3914144	0.1531719	2.555	0.0106 *
Migrationoceanodromous	0.1246218	0.1224664	1.018	0.3089
ComImportmedium	-0.1587844	0.0704845	-2.253	0.0243 *
ComImportminor	-0.1796318	0.1124642	-1.597	0.1102
ReceiverDensity	0.0003696	0.0006197	0.596	0.5509
MonitArea_km2	0.0292048	0.0036728	7.952	1.84e-15 ***
(Intercept)	-0.181077	0.348111	-0.520	0.6029
LengthStd	0.291368	0.155748	1.871	0.0614 .
BodyMassStd	-0.075598	0.112526	-0.672	0.5017
Longevity	-0.002075	0.003358	-0.618	0.5366
Vulnerability	-0.005287	0.004115	-1.285	0.1988
Troph	0.085869	0.105323	0.815	0.4149
Habitatdemersal	-0.083594	0.097555	-0.857	0.3915
Habitatpelagic-neritic	0.376237	0.150171	2.505	0.0122 *
Migrationoceanodromous	0.131005	0.121277	1.080	0.2800
ComImportmedium	-0.155050	0.069822	-2.221	0.0264 *
ComImportminor	-0.176964	0.111865	-1.582	0.1137
MonitArea_km2	0.027850	0.002885	9.652	<2e-16 ***
(Intercept)	-0.133518	0.342009	-0.390	0.69625
LengthStd	0.296191	0.155860	1.900	0.05739 .
BodyMassStd	-0.075325	0.112935	-0.667	0.50479
Vulnerability	-0.006247	0.003826	-1.633	0.10250
Troph	0.075257	0.104813	0.718	0.47275
Habitatdemersal	-0.085480	0.098149	-0.871	0.38379
Habitatpelagic-neritic	0.420951	0.133563	3.152	0.00162 **
Migrationoceanodromous	0.124283	0.121847	1.020	0.30773
ComImportmedium	-0.156069	0.070363	-2.218	0.02655 *
ComImportminor	-0.175389	0.112620	-1.557	0.11939
MonitArea_km2	0.027645	0.002869	9.636	< 2e-16 ***
(Intercept)	-0.146276	0.345919	-0.423	0.67240
LengthStd	0.221190	0.107616	2.055	0.03984 *
Vulnerability	-0.005991	0.003856	-1.554	0.12022
Troph	0.078409	0.106094	0.739	0.45988
Habitatdemersal	-0.071999	0.097259	-0.740	0.45913
Habitatpelagic-neritic	0.423247	0.135309	3.128	0.00176 **
Migrationoceanodromous	0.130083	0.123292	1.055	0.29139
ComImportmedium	-0.162957	0.070556	-2.310	0.02091 *
ComImportminor	-0.187196	0.112550	-1.663	0.09627 .
MonitArea_km2	0.027935	0.002840	9.837	< 2e-16 ***
(Intercept)	0.062773	0.199574	0.315	0.7531
LengthStd	0.215352	0.107531	2.003	0.0452 *
Vulnerability	-0.004267	0.003137	-1.360	0.1737
Habitatdemersal	-0.089431	0.096039	-0.931	0.3518
Habitatpelagic-neritic	0.478812	0.115393	4.149	3.33e-05 ***
Migrationoceanodromous	0.102495	0.119771	0.856	0.3921
ComImportmedium	-0.159660	0.071631	-2.229	0.0258 *
ComImportminor	-0.162555	0.109263	-1.488	0.1368
MonitArea_km2	0.027990	0.002844	9.843	< 2e-16 ***
(Intercept)	0.058174	0.203120	0.286	0.7746
LengthStd	0.209921	0.107650	1.950	0.0512 .
Vulnerability	-0.003275	0.002981	-1.099	0.2718
Habitatdemersal	-0.136348	0.080599	-1.692	0.0907 .
Habitatpelagic-neritic	0.524241	0.105321	4.978	6.44e-07 ***
ComImportmedium	-0.158326	0.073068	-2.167	0.0302 *
ComImportminor	-0.149920	0.110231	-1.360	0.1738
MonitArea_km2	0.027880	0.002847	9.793	< 2e-16 ***
(Intercept)	-0.147187	0.082274	-1.789	0.0736 .
LengthStd	0.209248	0.108037	1.937	0.0528 .
Habitatdemersal	-0.119649	0.081491	-1.468	0.1420
Habitatpelagic-neritic	0.509418	0.107251	4.750	2.04e-06 ***
ComImportmedium	-0.157843	0.075486	-2.091	0.0365 *
ComImportminor	-0.094148	0.100718	-0.935	0.3499
MonitArea_km2	0.028404	0.002818	10.079	< 2e-16 ***
(Intercept)	-0.073815	0.070812	-1.042	0.2972
Habitatdemersal	-0.078908	0.076210	-1.035	0.3005
Habitatpelagic-neritic	0.498507	0.103783	4.803	1.56e-06 ***
ComImportmedium	-0.143869	0.072709	-1.979	0.0479 *
ComImportminor	-0.110944	0.097446	-1.139	0.2549
MonitArea_km2	0.029416	0.002759	10.66	< 2e-16 ***

Backward Elimination KUD50

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7156996	0.3143443	-5.458	4.81e-08 ***
LengthStd	0.2458631	0.1431988	1.717	0.0860 .
BodyMassStd	-0.0223650	0.1028706	-0.217	0.8279
Longevity	-0.0023743	0.0030162	-0.787	0.4312
Vulnerability	-0.0051734	0.0037397	-1.383	0.1665
Troph	0.0901222	0.0958287	0.940	0.3470
Habitatdemersal	-0.0783874	0.0880839	-0.890	0.3735
Habitatpelagic-neritic	0.2830693	0.1373694	2.061	0.0393 *
Migrationoceanodromous	0.0953247	0.1094802	0.871	0.3839
ComImportmedium	-0.1393443	0.0631055	-2.208	0.0272 *
ComImportminor	-0.2268145	0.1011106	-2.243	0.0249 *
ReceiverDensity	0.0002265	0.0005662	0.400	0.6891
MonitArea_km2	0.0225671	0.0033794	6.678	2.42e-11 ***
(Intercept)	-1.7193710	0.3151560	-5.456	4.88e-08 ***
LengthStd	0.2235565	0.0997664	2.241	0.0250 *
Longevity	-0.0023814	0.0030303	-0.786	0.4319
Vulnerability	-0.0050871	0.0037327	-1.363	0.1729
Troph	0.0907909	0.0961681	0.944	0.3451
Habitatdemersal	-0.0743212	0.0864098	-0.860	0.3897
Habitatpelagic-neritic	0.2842343	0.1378534	2.062	0.0392 *
Migrationoceanodromous	0.0968283	0.1097437	0.882	0.3776
ComImportmedium	-0.1414654	0.0626291	-2.259	0.0239 *
ComImportminor	-0.2304948	0.1000786	-2.303	0.0213 *
ReceiverDensity	0.0002374	0.0005648	0.420	0.6742
MonitArea_km2	0.0226927	0.0033333	6.808	9.90e-12 ***
(Intercept)	-1.723887	0.315364	-5.466	4.59e-08 ***
LengthStd	0.222754	0.099682	2.235	0.0254 *
Longevity	-0.002370	0.003034	-0.781	0.4347
Vulnerability	-0.005199	0.003729	-1.394	0.1633
Troph	0.096182	0.095435	1.008	0.3135
Habitatdemersal	-0.074140	0.086525	-0.857	0.3915
Habitatpelagic-neritic	0.275117	0.136285	2.019	0.0435 *
Migrationoceanodromous	0.101037	0.109428	0.923	0.3558
ComImportmedium	-0.139393	0.062532	-2.229	0.0258 *
ComImportminor	-0.229239	0.100178	-2.288	0.0221 *
MonitArea_km2	0.021836	0.002643	8.262	< 2e-16 ***
(Intercept)	-1.668849	0.311159	-5.363	8.17e-08 ***
LengthStd	0.228970	0.099385	2.304	0.02123 *
Vulnerability	-0.006287	0.003483	-1.805	0.07101 .
Troph	0.083786	0.095444	0.878	0.38002
Habitatdemersal	-0.076708	0.087434	-0.877	0.38031
Habitatpelagic-neritic	0.326777	0.121959	2.679	0.00738 ***
Migrationoceanodromous	0.093105	0.110542	0.842	0.39965
ComImportmedium	-0.140498	0.063356	-2.218	0.02658 *
ComImportminor	-0.227251	0.101329	-2.243	0.02492 *
MonitArea_km2	0.021585	0.002626	8.219	< 2e-16 ***
(Intercept)	-1.606929	0.309045	-5.200	2.00e-07 ***
LengthStd	0.221855	0.099357	2.233	0.025556 *
Vulnerability	-0.004912	0.003168	-1.551	0.121004
Troph	0.058957	0.093102	0.633	0.526568
Habitatdemersal	-0.120749	0.071785	-1.682	0.092549 .
Habitatpelagic-neritic	0.381715	0.106803	3.574	0.000352 ***
ComImportmedium	-0.138028	0.064823	-2.129	0.033228 *
ComImportminor	-0.208862	0.101139	-2.065	0.038914 *
MonitArea_km2	0.021511	0.002632	8.173	3.01e-16 ***
(Intercept)	-1.449973	0.183845	-7.887	3.10e-15 ***
LengthStd	0.219014	0.099391	2.204	0.0276 *
Vulnerability	-0.003803	0.002688	-1.415	0.1572
Habitatdemersal	-0.124199	0.072500	-1.713	0.0867 .
Habitatpelagic-neritic	0.414891	0.095007	4.367	1.26e-05 ***
ComImportmedium	-0.135728	0.065596	-2.069	0.0385 *
ComImportminor	-0.192866	0.099180	-1.945	0.0518 .
MonitArea_km2	0.021573	0.002634	8.191	2.58e-16 ***
(Intercept)	-1.689221	0.075732	-22.305	< 2e-16 ***
LengthStd	0.219283	0.100006	2.193	0.0283 *
Habitatdemersal	-0.104884	0.074901	-1.400	0.1614
Habitatpelagic-neritic	0.399792	0.098870	4.044	5.26e-05 ***
ComImportmedium	-0.134535	0.069372	-1.939	0.0525 .
ComImportminor	-0.128633	0.092506	-1.391	0.1644
MonitArea_km2	0.022167	0.002619	8.464	< 2e-16 ***
(Intercept)	-1.75314	0.07420	-23.626	< 2e-16 ***
LengthStd	0.21974	0.09990	2.199	0.027846 *
Habitatdemersal	-0.12543	0.07637	-1.642	0.100496
Habitatpelagic-neritic	0.36534	0.10438	3.500	0.000465 ***
MonitArea_km2	0.02219	0.00265	8.374	< 2e-16 ***



Appendix 8: Comparison plots between observed and simulated length values for KUD50.



Appendix 9: Comparison plots between observed and simulated habitat values for KUD50.

Final Models Equations

The final model for KUD95 can be written as follows:

$$\begin{aligned}\ln(\hat{\text{KUD95}}_{ijk}) = &- 0.073815 \\ &- 0.078908 \cdot \text{Habitat}_{\text{demersal}} \\ &+ 0.498507 \cdot \text{Habitat}_{\text{pelagic-neritic}} \\ &- 0.143869 \cdot \text{ComImport}_{\text{medium}} \\ &- 0.110944 \cdot \text{ComImport}_{\text{minor}} \\ &+ 0.029416 \cdot \text{MonitAreakm2} \\ &+ u_j \cdot Z_{ij} \\ &+ v_k \cdot W_{ik}\end{aligned} \tag{1}$$

Interpretation:
$\hat{\text{KUD95}}_{ijk}$ - represents the estimated value of KUD95 for observation i of Transmitter j and File k.
β_0 = -0.073815 - represents the value of the intercept. This is the average value expected for $\ln(\hat{\text{KUD95}}_{ijk})$ when all the numerical explanatory variables are zero, the individual is benthopelagic and of high commercial importance. In this conditions, it is expected that the home range has 0.928km^2 ($e^{-0.073815}$).
β_1 = -0.078908 - represents the expected variation in $\ln(\hat{\text{KUD95}}_{ijk})$ when the individual is demersal compared to when it is benthic. In this case, a demersal individual is expected to have a 7.6% ($e^{-0.078908} = 0.924$) decrease in KUD95 compared to a benthic individual, controlling for random effects.
β_2 = 0.498507 - represents the expected variation in $\ln(\hat{\text{KUD95}}_{ijk})$ when the individual is pelagic-neritic compared to when it is benthic. In this case, a pelagic-neritic individual is expected to have a 64.6% ($e^{0.498507} = 1.646$)increase in KUD95 compared to a benthic individual, controlling for random effects.
β_3 = -0.143869 - represents the expected variation in $\ln(\hat{\text{KUD95}}_{ijk})$ when the individual is of medium commercial importance compared to when it is of high commercial importance. In this case, an individual of medium commercial importance is expected to have a 13.4% ($e^{-0.143869} = 0.866$) decrease in KUD95 compared to an individual of high commercial importance, controlling for random effects.
β_4 = - 0.110944 - represents the expected variation in $\ln(\hat{\text{KUD95}}_{ijk})$ when the individual has low commercial importance compared to when it has high commercial importance. In this case, an individual with low commercial importance is expected to have a 10.5% ($e^{-0.110944} = 0.895$) decrease in KUD95 compared to an individual of high commercial importance, controlling for

random effects.
β_5 = 0.029416 represents the expected change in $\ln(\hat{\text{KUD95}}_{ijk})$ when the monitored area increases by 1 unit. In this case, it is expected that when the monitored area increases by 1km^2, the KUD will increase by an average of 3% ($e^{0.029416} = 1.03$), controlling for random effects.
u_j and v_k - represent the random effect for transmitter j and species k, respectively. Their values can be accessed via the ranef() function. For the sake of simplicity, these values will not be presented here, as 30 coefficients would have to be written for the Species and 850 for the Transmitter.
Z_{ij} and W_{ik} - represent the Transmitter and Species, respectively.

The final model for KUD50 can be written as follows:

$$\begin{aligned}\ln(\hat{\text{KUD50}}_{ijk}) = &-1.75314 \\ &+ 0.21974 \cdot \text{LengthStd} \\ &- 0.12543 \cdot \text{Habitatdemersal} \\ &+ 0.36534 \cdot \text{Habitatpelagic-neritic} \\ &+ 0.02219 \cdot \text{MonitAreakm2} \\ &+ u_j \cdot Z_{ij} \\ &+ v_k \cdot W_{ik}\end{aligned} \tag{2}$$

Interpretation:
KUD95$_{ijk}$ - represents the estimated value of KUD50 for observation i of Transmitter j and File k
β_0 = - 1.75314 - represents the value of the intercept. This is the average value expected for $\ln(\text{KUD50}_{ijk})$ when all the explanatory variables take the value of zero, which means that when all variables take the value 0, it is expected that the home range has 0.17km^2 ($e^{-1.75314}$).
β_1 = 0.21974 - represents the expected change in $\ln(\text{KUD50}_{ijk})$ when the length increases by 1 unit. In this case, it is expected that when the length increases by 1cm, the KUD will increase by an average of 24.6% ($e^{0.21974} = 1.246$), controlling for random effects.
β_2 = - 0.12543 - represents the expected variation in $\ln(\text{KUD50}_{ijk})$ when the individual is demersal compared to when it is benthic. In this case, a demersal individual is expected to have a 11.8% ($e^{-0.12543} = 0.882$) decrease in KUD95 compared to a benthic individual, controlling for random effects.

β_3 = 0.36534 - represents the expected variation in $\ln(\text{KUD50}_{ijk})$ when the individual is pelagic-neritic compared to when it is benthic. In this case, a pelagic-neritic individual is expected to have a 44% ($e^{0.36534} = 1.44$)increase in KUD95 compared to a benthic individual, controlling for random effects.
β_4 = 0.02219 - represents the expected change in $\ln(\text{KUD50}_{ijk})$ when the monitored area increases by 1 unit. In this case, it is expected that when the monitored area increases by 1km^2, the KUD50 will increase by an average of 2.2% ($e^{0.02219} = 1.022$), controlling for random effects.
u_j and v_k - represent the random effect for transmitter j and species k, respectively. Their values can be accessed via the ranef() function. For the sake of simplicity, these values will not be presented here, as 30 coefficients would have to be written for the Species and 850 for the Transmitter.
Z_{ij} and W_{ik} - represent the Transmitter and Species, respectively.

Note: When we say ”controlling for random effects,” we mean that even after accounting for the variability between Transmitters and Species, there remains a relationship of that proportion between the response variable and the explanatory variable.

Note: The variables Habitat, ComImport, Transmitter, and Species are categorical variables that have been transformed into dummy variables to be included in the models.

Required Packages

correlation - Methods for Correlation Analysis
ggplot2 - Create Elegant Data Visualisations Using the Grammar of Graphics
gridExtra - Miscellaneous Functions for ”Grid” Graphics
car - Companion to Applied Regression
rpart - Recursive Partitioning and Regression Trees
lme4 - Linear Mixed-Effects Models using ’Eigen’ and S4
Matrix - Sparse and Dense Matrix Classes and Methods
mgcv - Mixed GAM Computation Vehicle with Automatic Smoothness Estimation
glmmTMB - Generalized Linear Mixed Models using Template Model Builder
lmtest - Testing Linear Regression Models gamm4 - Generalized Additive Mixed Models using ’mgcv’ and ’lme4’
sjPlot - Data Visualization for Statistics in Social Science
sjmisc - Data and Variable Transformation Functions
sjlabelled - Labelled Data Utility Functions
performance - Assessment of Regression Models Performance
DHARMa - Residual Diagnosis for Hierarchical (Multi-Level/ Mixed) Regression Models

R Markdown Documents

R Markdown Documents of the dataset creation and statistical analysis can be found at the GitHub link:

- https://github.com/MatildeCorreia17/Tese-rmd