

# Optimization of a cetacean occurrence dataset: methods for controlling data bias, verification and validation

Cláudia Sofia Oliveira Rodrigues  
Master dissertation presented to  
Faculty of Sciences of the University of Porto  
Marine Ecology  
2021

MSC

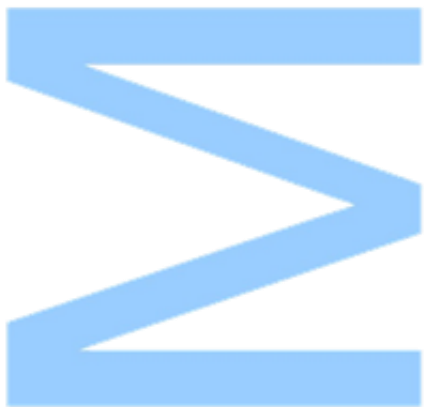
2.º  
CICLO

FCUP  
2021



Optimization of a cetacean occurrence dataset:  
methods for controlling data bias, verification and  
validation

Cláudia Sofia Oliveira Rodrigues





# Optimization of a cetacean occurrence dataset: methods for controlling data bias, verification and validation

Cláudia Sofia Oliveira Rodrigues

Master degree in Ecology and Environment

Department of Biology

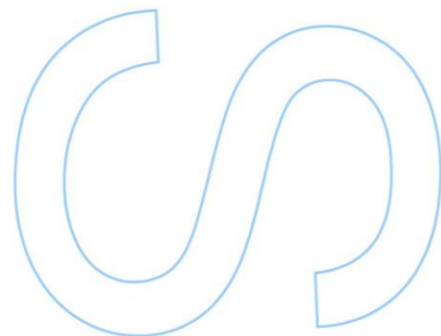
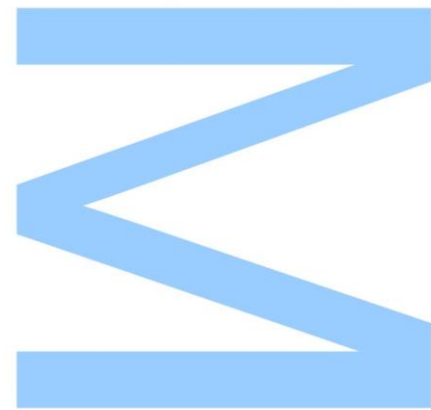
2021

## Supervisor

Dr. Ana Mafalda Correia, Researcher at CIIMAR – Porto, Portugal

## Co-supervisor

Prof. Dr. Isabel Sousa Pinto, Professor at Sciences Faculty, Porto University  
and Researcher at CIIMAR – Porto, Portugal



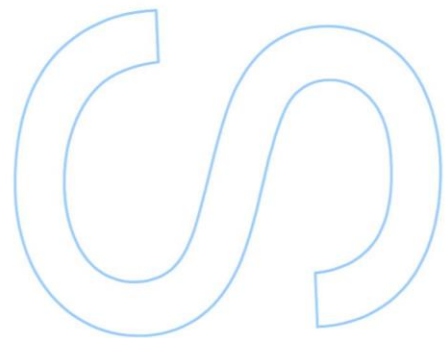
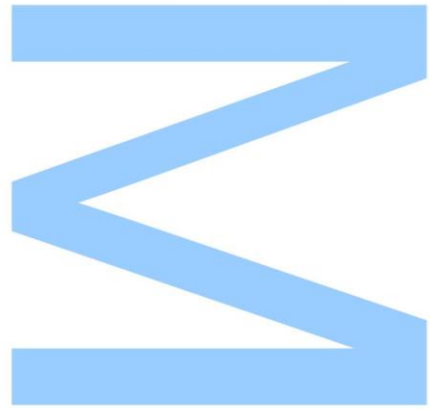




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, \_\_\_\_ / \_\_\_\_ / \_\_\_\_





“Optimism is a strategy for making a better future. Because unless you believe that the future can be better, it’s unlikely you will step up and take responsibility for making it so. If you assume that there’s no hope, you guarantee that there will be no hope. If you assume that there is an instinct for freedom, there are opportunities to change things, there’s a chance you may contribute to making a better world. The choice is yours.”

Noam Chomsky



# Acknowledgments

*To people,*

To my supervisor, Dr. Ana Mafalda Tomás Correia, for accepting me as her mentee and for giving me the chance to write a thesis on the thematic I wanted. But above all, for providing me guidance and inputs throughout this hard-working year, and for her support, encouragement, and patience. Without her, I would not have been able to overcome many of the obstacles I found. I am truly grateful to have her not only as an outstanding supervisor but also as a good friend.

To my co-supervisor, Prof. Dr. Isabel Sousa Pinto, for always being so approachable and for introducing me to all the amazing people who are part of the CETUS Project.

To Ágatha Gil, for always making me feel confident in my abilities after coming to her, and for her comforting words and contagious smile. But mostly, for being an inspiration.

To Raul Valente, for all the useful advice and for always being ready to help no matter the subject.

To Guilherme Estrela, for his enthusiasm with this work and for dedicating his time to providing me incredibly useful advice on species identification. I learned a lot from him.

To Luís Perat, for dedicating his time to read my thesis and for the attention he gave to the smallest details in grammar and sentence structuring in such a short time.

To former CETUS volunteers, for providing me their sighting photos and other additional information essential to the realization of this thesis. They were super available.

To my parents, Luís Rodrigues and Paula Rodrigues, who never stopped believing in me. For never raising an eyebrow when I chose to take this path, and for being there when I was at my lowest. They have been tireless over the years. I am who I am because of the love and support that my parents gave me, and I will never be able to thank them enough.

To my boyfriend, Vasco Freitas, for all his love and support, and for bearing with me on the days when I was the grumpiest person on earth. For being so understanding and for all the happy distractions to rest my mind outside of my research.

To all my friends and family, I cannot forget to thank them all for their unconditional support in this very intense academic year. I am lucky to have them.



*To institutions and organizations,*

To CETUS Project, for providing me the data for this thesis, for the amazing people I had the chance to meet, and for fulfilling my dream of becoming a marine mammal observer. But above all, for making me realize that this is what I want for my life.

To the Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), for welcoming me and giving me the best conditions to work.

To the Faculty of Sciences of the University of Porto (FCUP), for giving me the tools to thrive in the field of Ecology.

**Thank you!**

# Resumo

Os datasets de monitorização a longo prazo podem servir de base para melhor entender as respostas físicas e ecológicas às alterações ambientais do oceano, desempenhando um papel importante na gestão e conservação marinha. Logo, devem tornar-se públicos, seguindo os Princípios FAIR, de forma a potenciar a reutilização. Também é fundamental que os dados sejam confiáveis e as fontes de enviesamento sejam identificadas e quantificadas. O Projeto CETUS, um programa de monitorização de cetáceos no Atlântico Nordeste, em funcionamento desde 2012, conta com a participação internacional de biólogos para reunir dados em rotas de longos transectos a bordo de grandes embarcações utilizadas como plataformas de oportunidade. O dataset CETUS é disponibilizado em livre-acesso nos portais OBIS e EMODnet. Este trabalho teve como objetivo otimizar e permitir o uso adequado do dataset por meio de: i) aplicação de métodos de verificação/validação, com base na confirmação fotográfica das espécies identificadas; ii) criar critérios para a qualidade dos dados, com base na experiência do observador; e iii) avaliar a influência dos parâmetros de enviesamento, usando Modelos Aditivos Generalizados (MAG) para correlacionar o número de avistamentos com os quilómetros amostrados “em esforço”, as condições meteorológicas e a experiência dos observadores. Dos registos fotográficos reunidos, ~90,9% foram cruzados com as ocorrências registadas no dataset, embora correspondendo apenas a ~7.5% do dataset total. Dos registos cruzados, ~17.1% dos registos permitiram alcançar um táxon inferior, e em ~10.8% foi possível chegar à espécie. No total, ~59.2% avistamentos foram validados até à espécie e ~3.5% identificações erradas foram corrigidas. Isto revela a importância dos métodos de verificação/validação e a necessidade de aumentar os registos fotográficos durante a amostragem. O MAG revelou quais as variáveis que mais afetam a eficácia da monitorização. Em última análise, este trabalho contribuirá para um uso mais informado do dataset, e para o melhoramento do protocolo de monitorização do CETUS e de programas semelhantes.

**Palavras-chave:** Ocorrência de Cetáceos, Monitorização da Biodiversidade, Datasets, Verificação de Dados, Validação de Dados, Enviesamento Metodológico

# Abstract

Long-term monitoring datasets can provide a baseline to better understand physical and ecological responses to ocean environmental changes, playing an important role in marine management and conservation. Thus, they must become public following FAIR Data Principles to enhance their reusability. It is also fundamental that the data is reliable, and the sources of bias are identified and quantified. CETUS Project, a cetacean monitoring program in the NE Atlantic, ongoing since 2012, counts on international participation of biologists to collect data on long-transect routes from large vessels used as platforms of opportunity. The CETUS dataset is made available open-access at OBIS and EMODnet portals. This work aimed to optimize and allow the proper use of the dataset by: i) applying verification/validation methods, based on photographic confirmation of identified species; ii) creating criteria for the quality of the data, based on the observer's experience; and iii) assessing the influence of bias parameters, using Generalized Additive Models (GAM) to correlate the number of sightings with kilometres sampled "on-effort", weather conditions, and the experience of observers. From the collected photographic records, ~90.9% were matched with the dataset occurrences, although corresponding only to ~7.5% of the total dataset. Out of the matched records, ~17.1% records were able to reach a lower taxon, with ~10.8% up to the species. In total, ~59.2% sightings were validated to the species level and ~3.5% wrong identifications were corrected. This reveals the importance of verification/validation methods and the need to increase photographic registers during sampling. GAM allowed to assess which variables most affect monitoring efficiency. Ultimately, this work will contribute to a more informed use of the dataset, and an improvement of monitoring protocols of CETUS and similar programs.

**Key-words:** Cetacean Occurrence, Biodiversity Monitoring, Long-term Datasets, Data Verification, Data Validation, Methodological Bias



# Table of Contents

<b>List of Tables</b>	<b>ii</b>
<b>List of Figures</b>	<b>iv</b>
<b>List of Abbreviations</b>	<b>vi</b>
<b>1. Introduction</b>	<b>1</b>
1.1 Cetacean Monitoring Programs: Importance and Application	1
1.2 Reliability and Re-usability of Data: Verification and Validation Processes	10
1.3 Long-term Monitoring Datasets: Challenges and Opportunities	16
1.4 CETUS Project: Open-source Dataset	19
1.5 Aims	24
<b>2. Materials and Methods</b>	<b>25</b>
2.1 The CETUS dataset: Study Area, Data, Structure	25
2.2 Photographic Verification/Validation	30
2.3 Creating a Data Quality Criteria: MMOs Experience	33
2.4 Bias Modelling of Number of Sightings	35
<b>3. Results</b>	<b>37</b>
3.1 Photographic Verification/Validation	37
3.2 Creating a Data Quality Criteria: MMOs Experience	42
3.3 Bias Modelling of Number of Sightings	44
<b>4. Discussion</b>	<b>47</b>
4.1 Photographic Verification/Validation	47
4.2 Creating a Data Quality Criteria: MMOs Experience	51
4.3 Bias Modelling of Number of Sightings	52
<b>5. Conclusion</b>	<b>53</b>
<b>6. Bibliographic References</b>	<b>55</b>
<b>7. Appendix</b>	<b>66</b>

## List of Tables

**Table 1.** An overview of different cetacean monitoring approaches, techniques, and its applications.

**Table 2.** An overview of different cetacean monitoring platforms.

**Table 3.** An overview of active monitoring activities for cetaceans in Portugal. ENA – Eastern North Atlantic. OPO – Observation Platforms of Opportunity. ID – Identification. SAC – Special Area of Conservation.

**Table 4.** Examples of different types of technical processes applied to different monitoring datasets, from data collection to data processing, to ensure data quality. ENA – Eastern North Atlantic. MMO – Marine Mammal Observer. ID – Identification. QA – quality assurance. QC – quality control.

**Table 5.** The FAIR Guiding Principles (Wilkinson et al., 2016).

**Table 6.** Meteorologic variables assessed during CETUS Project surveys. Indication of meteorologic conditions when sampling is active (i.e., “on effort”, marked in green) and conditions when sampling is considered opportunistic (i.e. “off effort”, marked in red).

**Table 7.** Number of cetacean occurrences. The number of occurrences is presented by taxa recorded to the highest possible level. The table is organized by taxon rank of the records and alphabetically within.

**Table 8.** CETUS internal dataset structure with the information available for each datasheet (cruise, survey, segment, and positions). YY/YYYY – year. SL – ship’s letter code. NNN – number of the cruise in that ship, in that year. DD – day. MM – month. S – survey number of the day. SN – ship’s number code. EE – effort code. SSS – sighting number.

**Table 9.** Evaluation criteria created to evaluate the experience of the CETUS Project Marine Mammal Observers and generate a data quality criterion. ID – Identification.

**Table 10.** Number and percentage of the matched (M) and non-matched (NM) sightings by suborder.

**Table 11.** Number and percentage of the validation results of completed validated records by suborders and non-identified sightings (NI). Complete Validation represents all validations that reached the species level.

**Table 12.** Number and percentage of the validation results of non-validated/incomplete validated records by suborders and non-identified sightings (NI). Non-Validation / Incomplete Validation represents all validations that could not reach the species level.

**Table 13.** Updated number of cetacean occurrences of the CETUS dataset after the video/photographic verification and validation processes. Within parentheses is the difference in values when compared to table 7. The number of occurrences is presented by taxa recorded to the highest possible level. The table is organized by taxon rank of the records and alphabetically within.

**Table 14.** Results from the best final Generalized Additive Model (GAM) developed for assessing the bias on the number of sightings collected per survey. Sight – number of sightings per survey. MEO – Most Experienced Observer. Min\_Sea – minimums of the sea state in each survey. Min\_Wind – minimums of the wind state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey. Max\_Vis – maximums of the visibility in each survey.

**Table 15.** Results from the Mann-Whitney U tests performed between matched (M) and non-matched (NM) sightings for each group.

**Table 16.** Results from the Mann-Whitney U test performed between Odontoceti and Mysticeti matched (M) and non-matched (NM) sightings.

**Table 17.** Results from the Mann-Whitney U tests performed between complete validated (C-V) and incomplete/non-validated (IN-V) sightings for each group.

**Table 18.** Results from the Mann-Whitney U test performed between Odontoceti and Mysticeti complete validated (C-V) and incomplete/non-validated (IN-V) sightings.

**Table 19.** Variance Inflation Factor (VIF) results. LEO – evaluation score of Least Experienced Observers per survey. MEO – evaluation score of Most Experienced Observers per survey. Min\_Sea – minimums of the sea state in each survey. Max\_Sea – maximums of the sea state in each survey. Min\_Wind – minimums of the wind state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey. Max\_Vis – maximums of the visibility in each survey.

**Table 20.** Results of basis dimension (k) checking (with gam.check). GAM – Generalized Additive Model. EDF – Degrees of Freedom. MEO – evaluation score of Most Experienced Observers. Min\_Sea – minimums of the sea state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey.

## List of Figures

**Figure 1.** Whale carbon and oxygen flux. Retrieved from: GRID-Arendal (<https://www.grida.no/resources/14276>).

**Figure 2.** Examples of anthropogenic threats to cetaceans. **(a)** Bottlenose dolphin (*Tursiops truncatus*) shows signs of skin lesions associated with a deadly skin disease known as ulcerative dermatitis linked to extreme climate events. **(b)** Bottlenose dolphin (*Tursiops truncatus*) eats a plastic can holder. **(c)** Humpback whale (*Megaptera novaeangliae*), “Bladerunner”, shows extensive scarring from a boat propeller. **(d)** Humpback whale (*Megaptera novaeangliae*) entangled in a fishing net.

**Figure 3.** Study area and routes of the CETUS Project. IP – Iberian Peninsula; NWA – Northwest Africa; AZ – Azores; MAD – Madeira; CI – Canary Islands; CV – Cape Verde.

**Figure 4.** Monte da Guia vessel arriving to Port of Leixões with indication of the position of the Marine Mammal Observers (MMOs) during CETUS Project surveys, and the direction of the route.

**Figure 5.** Study area with cetacean occurrences. IP – Iberian Peninsula; NWA – Northwest Africa; AZ – Azores; MAD – Madeira; CI – Canary Islands; CV – Cape Verde.

**Figure 6.** Boxplot of the sighting distance of the matched (M) and non-matched (NM) records by suborder (Odontoceti and Mysticeti) and NI. NI – non-identified.

**Figure 7.** Boxplot of the sighting distance of the matched sightings by validation results. C-V – complete validation; IN-V – incomplete validation or non-validated sightings.

**Figure 8.** Boxplot of the validation results of the matched sightings by suborder. C-V – complete validation; IN-V – incomplete validation or non-validated sightings. NI – non-identified.

**Figure 9.** Histograms representing the frequency of cruises across the range of the Marine Mammal Observers’ experience, based on the MMOs evaluation performed, for the Most Experienced Observer (MEO) and the Least Experienced Observer (LEO).

**Figure 10.** Bar plot representing the evaluation score for the Most Experienced Observer (MEO) and for the Least Experienced Observer (LEO) for each combination of scores of the Marine Mammal Observers’ teams. For the surveys where there was only one Marine Mammal Observer, only the orange bar (MEO) is displayed.



**Figure 11.** Plots of the final Generalized Additive Model (GAM) developed for assessing the influence of bias of kilometres samples “on-effort”, meteorological conditions and the experience of observers. MEO – Most Experienced Observers.

**Figure 12.** Correlation Matrix. Results of Pearson correlations between all pairs of explanatory variables Results. Sight – number of sightings per survey. LEO – evaluation score of Least Experienced Observers per survey. MEO – evaluation score of Most Experienced Observers per survey. Mean – mean of the evaluation scores of the observers in each survey. Comb – accumulated evaluation scores of the observers in each survey. Min\_Sea – minimums of the sea state in each survey. Max\_Sea – maximums of the sea state in each survey. Min\_Wind – minimums of the wind state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey. Max\_Vis – maximums of the visibility in each survey.

**Figure 13.** GAM Check Plots. GAM – Generalized Additive Model.

**Figure 14.** GAM influence plot. GAM – Generalized Additive Model.

**Figure 15.** GAM effort residuals plot. GAM – Generalized Additive Model.

**Figure 16.** GAM MEO residuals plot. GAM – Generalized Additive Model. MEO – evaluation score of Most Experienced Observers per survey.

**Figure 17.** GAM Min\_Sea residuals plot. GAM – Generalized Additive Model. Min\_Sea – minimums of the sea state in each survey.

**Figure 18.** GAM Min\_Vis residuals plot. GAM – Generalized Additive Model. Min\_Vis – minimums of the visibility in each survey.

**Figure 19.** GAM Max\_Vis residuals plot. GAM – Generalized Additive Model. Max\_Vis – maximums of the visibility in each survey.

**Figure 20.** GAM Min\_Wind residuals plot. GAM – Generalized Additive Model. Min\_Wind – minimums of the wind state in each survey.

**Figure 21.** GAM Max\_Wind residuals plot. GAM – Generalized Additive Model. Max\_Wind – maximums of the wind state in each survey.

# List of Abbreviations

**AIC** – Akaike Information Criterion

**ANOVA** – Analysis of Variance

**AZ** – Azores

**CI** – Canary Islands

**C-V** – Complete Validation

**CV** – Cape Verde

**DD** - Day

**EE** – Effort code

**EEZ** – Economic Exclusion Zone

**EMODnet** – European Marine Observation and Data Network

**ENA** – Eastern North Atlantic

**EU** – European Union

**FAIR** - Findability, Accessibility, Interoperability, and Reusability

**GAM** – Generalized Additive Model

**GPS** – Global Positioning System

**ID** – Identification

**IN-V** – Non-validation / Incomplete Validation

**IP** – Iberian Peninsula

**LEO** – Least Experienced Observer

**M** – Matched

**MM** – Month

**MAD** – Madeira

**Max\_Vis** – Maximums of the visibility

**Max\_Wind** – Maximums of the wind State

**MEO** – Most Experienced Observer

**Min\_Sea** – Minimums of the sea State

**Min\_Vis** – Minimums of the visibility

**Min\_Wind** – Minimums of the wind State

**MMO** – Marine Mammal Observer

**NNN** – Number of the cruise in that ship, in that day

**NI** – Non-identified

**NM** – Non-Matched

**NWA** – Northwest Africa

**OBIS** – Ocean Biodiversity Information System

**OPO** – Observation Platforms of Opportunity

**QA** – Quality Assurance

**QC** – Quality Control

**S** – Survey number of the day

**SSS** – Sighting number

**SAC** – Special Area of Conservation

**Sight** – Number of sightings

**SL** – Ship's letter code

**SN** – Ship's number code

**UTC** - Universal Time Coordinated

**VIF** - Variance Inflation Factor

**YY / YYYY** – Year



# 1. Introduction

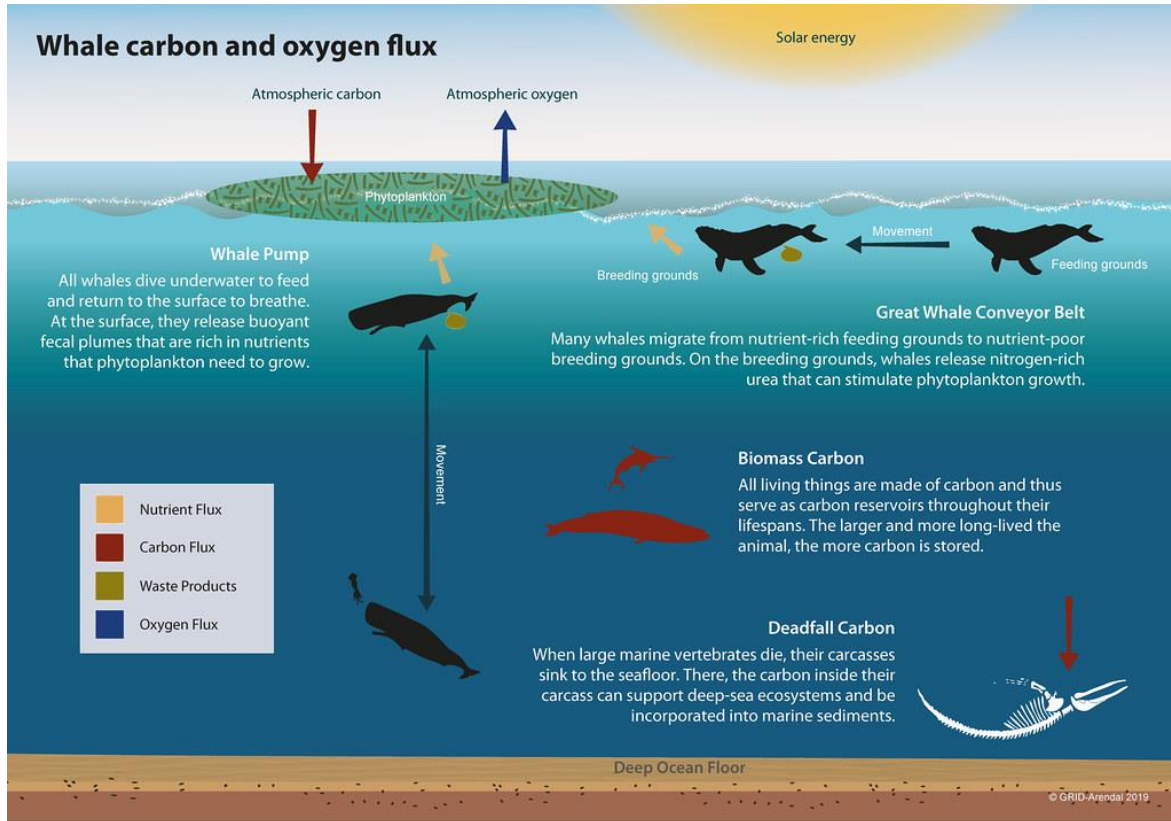
## 1.1 Cetacean Monitoring Programs: Importance and Application

Cetaceans play several roles in marine and other aquatic and nearshore ecosystems (Estes, 1998). As apex predators, the status of cetacean populations might reflect the state of an entire ecosystem. They act as sentinels of disturbances/risks for the environments they inhabit, mainly due to their longer lifespans, low densities, and low fecundity, which makes them very susceptible to changes in ecosystems; and they serve as good bioindicators of ecosystem contamination since they accumulate numerous compounds throughout their life (Durante et al., 2020; Sergio et al., 2008). Besides that, being on higher trophic levels, cetaceans may have a considerable impact on the structuring of various ecosystems, which makes them keystone species (Sergio et al., 2008). As a matter of fact, they are good vectors of nutrients, either vertical, when feeding in-depth and defecating higher in the water column – “whale pump” –, or horizontal, in their migrations between high-latitude feeding areas and low-latitude calving grounds – “great whale conveyor belt” (Figure 1; Roman et al., 2014; Roman & McCarthy, 2010). This promotes the production of phytoplankton and an increase in fish stocks (Roman et al., 2014). Along with all of this, they are charismatic species, attracting the public and media attention, which makes them flagship species (Parsons et al., 2015). As management actions addressing cetacean conservation will likely have positive effects on the conservation of the marine ecosystems, they truly can be focal species for conservation marketing (Sergio et al., 2008).

Addressing the Sustainable Development Goals established by the United Nations, more specifically Goal 14, that aims to “conserve and sustainably use the oceans, seas and marine resources”, it is necessary to rebuild the marine life-support systems (Duarte et al., 2020; United Nations, 2020). For all the reasons previously mentioned, cetaceans can play an important role in helping to achieve this goal while supporting Sustainable Blue Economy, which represents roughly 5.4 million jobs and generates a gross added value of almost €500 billion a year in the European Union (EU; European Commission, 2012). The whale-watching industry alone is thought to be worth over \$2.5 billion in yearly revenue and about 19,000 jobs all around the world (Cisneros-Montemayor et al., 2010). Nevertheless, a recent study estimates that a single whale is worth an average of \$2 million over its lifetime (Chami et al., 2020).

In addition, sinking cetacean carcasses provide a large amount of organic matter to deeper areas where it is used to support some ecosystems and species (Figure 1). Also, this carbon is removed from the atmosphere to the deep sea. If all whale populations were restored,

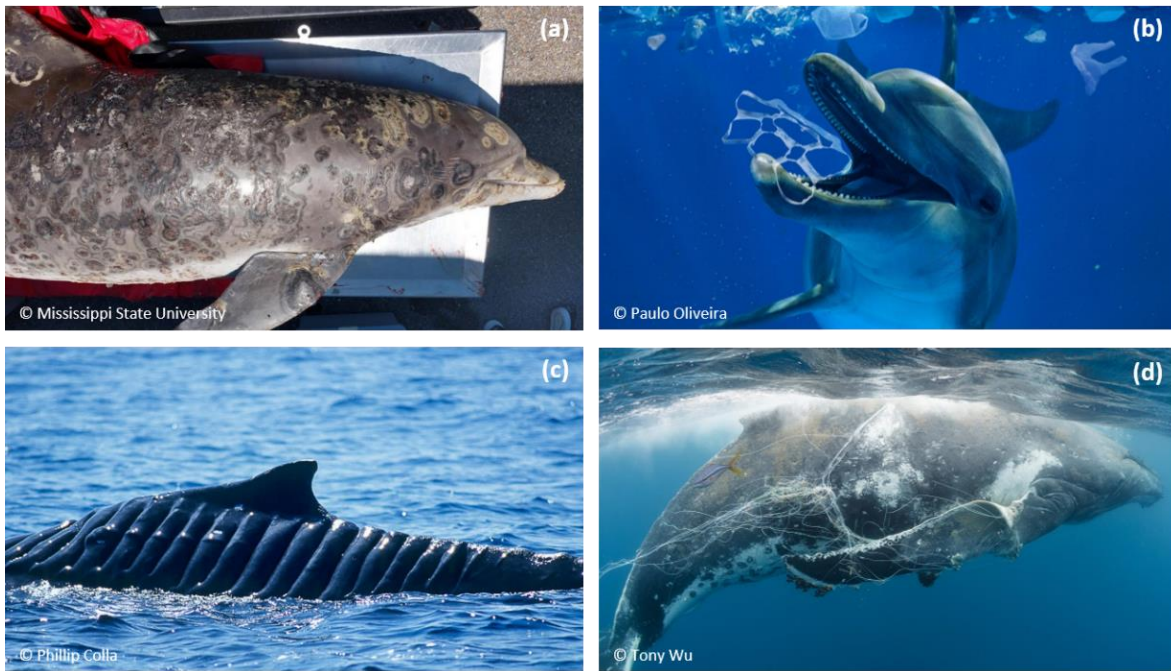
160,000 tons of carbon would be removed per year through their sinking. This flux would be equivalent to preserving 843 hectares of forest each year, which can have an important role in the mitigation of the climate change effects (Pershing et al., 2010). This emphasizes the high importance of cetacean species in marine ecosystems.



**Figure 1.** Whale carbon and oxygen flux. Retrieved from: GRID-Arendal (<https://www.grida.no/resources/14276>).

To safeguard cetacean conservation, many countries have legislation/laws that aim to protect them. Regarding international legislation, the Habitats Directive (92/43/EEC, 21 May 1992) stands out, which applies to the marine waters of Economic Exclusion Zones (EEZ) of the member states of the EU. Cetaceans are included in Annex IV, being considered species of community interest in need of strict protection. There is also Annex II, which includes species of community interest whose conservation requires the designation of a Special Area of Conservation (SAC). Harbour porpoises (*Phocoena phocoena*) and bottlenose dolphins (*Tursiops truncatus*) are part of this list.




However, cetaceans are nowadays quite exposed to a panoply of threats, most frequently related to anthropogenic activities, such as climate change effects (Duignan et al., 2020), marine and noise pollution (Parsons et al., 2008; Simmonds, 2017), collisions with boats (Peltier et al., 2019), ghost gear entanglements (Stelfox et al., 2016) or bycatch (Brownell Jr et al., 2019; Figure 2).





**Figure 2.** Examples of anthropogenic threats to cetaceans. (a) Bottlenose dolphin (*Tursiops truncatus*) shows signs of skin lesions associated with a deadly skin disease known as ulcerative dermatitis linked to extreme climate events. (b) Bottlenose dolphin (*Tursiops truncatus*) eats a plastic can holder. (c) Humpback whale (*Megaptera novaeangliae*), “Bladerunner”, shows extensive scarring from a boat propeller. (d) Humpback whale (*Megaptera novaeangliae*) entangled in a fishing net.



Consequently, for efficient cetacean conservation, it is crucial to monitor their distribution and abundance, determine population size and identify potential threats to populations and habitats. It is also determinant to maintain long-term monitoring programs to track changes over space and time (Evans & Hammond, 2003). Worldwide, several different types of cetacean monitoring approaches (Table 1) and platforms (Table 2) exist to inform conservation and management actions, each of them with its advantages and challenges.

**Table 1.** An overview of different cetacean monitoring approaches, techniques, and its applications.

Monitoring Approach	Techniques	Main Applications
 <b>Visual</b>	Marine Mammal Observers Photography Drone Citizen Science	Study of cetacean populations, distribution, abundance and habitat. Information on life history.
 <b>Acoustic</b>	Active Acoustic Methods Passive Acoustic Methods	Study cetacean acoustic communication. Study cetacean behavior. Better understand the anthropogenic noise impacts on cetaceans.
 <b>Sampling</b>	Tissue samples Blow samples	Genetics research. Toxicology research. Stable isotope research. Monitoring cetacean health.

 <p><b>Transmitters</b></p>	<p>Invasive tags Non-invasive tags</p>	<p>Better understand cetaceans physiology, biology, behaviour, ecology and habitat requirements.</p>
 <p><b>Stranding Networks</b></p>	<p>Tissue samples Necropsies</p>	<p>Genetics research. Toxicology research. Stable isotope research. Monitoring cetacean health. Assessment of death causes, including anthropogenic threats.</p>

**Table 2.** An overview of different cetacean monitoring platforms.

Monitoring Platforms		
 <p><b>Dedicated</b></p>	<p>Fixed survey stations Boats/ships Aircrafts</p>	 <p><b>Of Opportunity</b></p> <p>Fixed survey stations Boats/ships Aircrafts</p>

As an initial step, the approach of a monitoring program will depend on the goal of the study, the targeted species, and the resources available. For instance, when the goal is to gather information on abundance and distribution, cetacean researchers commonly use line-transect visual surveys. Conceptually, systematic surveys aboard dedicated platforms and dedicated Marine Mammal Observer (MMO) teams would be ideal, but to conduct such a design-unbiased survey can be extremely costly (Williams et al., 2006). Placing trained and dedicated observers on Observation Platforms of Opportunity (OPOs), such as inter-island ferries, whale-watching vessels, or cargo ships, can be a good alternative (Correia et al., 2019a; Williams et al., 2006; Kiszka et al., 2007). This approach can provide a wider spatial and temporal coverage; however, it leads to little or no control over survey design. Land-based surveys can also be quite useful to contribute to the understanding of a region's marine megafauna (Clarke et al., 2017). Usually, this approach is less intrusive, inexpensive, and not labour-intensive. However, it is limited to the narrow area near the coast, visible from the land-based point of observation.

Another way to counter financial barriers is by exploiting citizen science. Data can be collected by trained citizen scientists or the general public (Kosmala et al., 2016; Robbins et al., 2020). The Cetus project, presented in Table 3, is a citizen science project – the observers are volunteers, and the project use ships of opportunity. The project Sail & Whale is also a citizen science project (Table 3). This project relies on sailors aboard non-traditional observation platforms, sailboats, to collate cetacean occurrences from all around



the world's oceans. BioDiversity4All, a country-wide citizen science project in Portugal, is also a good example. This database already made it possible to publish articles with its records (e.g., to estimate climatic niches and species distributions; Tiago et al., 2017). These types of programmes are effective for data collection, raising awareness, and involving citizens in scientific endeavours. However, data validation efforts and improved verification tools, such as expert validation, accounting for random error and systematic bias, and skill-based statistical weighting of observer classifications, are necessary to produce high-quality ecological data derived from citizen science (Kosmala et al., 2016). The quality of data provided by citizen science can automatically be improved by trained data providers (e.g., training courses on data collection, species detection, and identification to fishing communities, sailors, or other sea users).

Besides collecting visual data on presence records [occurrence, species identification (ID), group size, and behaviour], other data/information on cetaceans may be collected in monitoring programmes. Photographic records, such as for photo-ID studies, are of great value to providing life history data and to estimate population parameters (Evans & Hammond, 2003; Hammond, 1990). However, it is limited to individuals with identifiable marks, like the presence of wounds and scars. Drone-based photo-ID is also a promising avenue of research, facilitating studies on individual history, site fidelity, or habitat use in locations with good visibility and flying conditions (Landeo-Yauri et al., 2020; Raoult et al., 2020). These can be operated from fixed stations if animals are close to the shore, or from boats. Besides photo-ID, they can also provide information on their ecology, behaviour, health, and movement patterns (Raoult et al., 2020).

Acoustic techniques such as passive acoustic monitoring provide coverage of areas that are otherwise difficult to observe for species presence, and acoustic habitat information, including anthropogenic sounds (Davis et al., 2017). Nevertheless, it essentially relies upon animals being vocal.

Biological samples can also provide various information on cetacean populations and, once again, drones can facilitate this task. For example, the use of drones to capture large whales' exhaled breath (blow) to examine the associated microbiome can provide a non-invasive method to remotely monitor their respiratory health (Apprill et al., 2017). However, in genetics, genomics, toxicology, or stable isotope research it is often necessary to collect tissue samples from the animals (Noren & Mocklin, 2012). This can ultimately present a risk to cetacean health and welfare since it is an invasive approach that implies intentionally breaking the skin. However, when well-applied, biopsy-sampling technique on live animals is minimally invasive and can provide numerous information on the animals. On the other

hand, stranding networks have at their disposal all types of samples from stranded animals, which can reveal a lot about their biology and health status (Arregui et al., 2017; Bento et al., 2019; Monteiro et al., 2020). Yet, stranding networks tend to provide biased data, as sick or stressed animals have a higher tendency to be sampled than healthy animals.

The deployment of satellite transmitters allows researchers to track movement patterns, habitat use, and other behavioural aspects that are otherwise difficult to monitor since they spend most of their time underwater (Andrews et al., 2019). This approach proves to be very useful for collecting information on cetacean physiology, behaviour, and ecology. The method of attachment can be either non-invasive (suction cup tag) or invasive (e.g., anchored, bolt-on, or consolidated). The latter, as with tissue sampling, can present a risk to cetacean health and welfare (Andrews et al., 2019).

Regardless of the chosen platform and monitoring technique, monitoring programmes often rely on the work of several researchers, students, and volunteers, since monitoring must be carried out throughout the year as many times as possible (Evans & Hammond, 2003). In Table 3 some examples of programmes and research centres are identified, with different goals and methodologies dedicated to studying cetaceans in Portugal.

Ultimately, it is also worth mentioning the importance of a good balance between the targets of a monitoring program, the resources available, and the methodologies applied. Most of the time, the chosen approach can make a difference in the program's endurance and sustainability.

**Table 3.** An overview of active monitoring activities for cetaceans in Portugal. ENA – Eastern North Atlantic. OPO – Observation Platforms of Opportunity. ID – Identification. SAC – Special Area of Conservation.

Program	Type	Study Area	Aims	Applications
<b>CETUS Project</b>	Program	ENA	Monitoring of marine megafauna in the ENA to provide cetacean occurrence data, using OPOs.  ( <a href="https://www2.ciimar.up.pt/projects.php?id=59">https://www2.ciimar.up.pt/projects.php?id=59</a> )	Provide new insights into distribution and abundance of cetaceans. Deliver habitat models to map habitat suitability. Explore and predict cetacean hotspots.
<b>AIMM - Marine Environment Research Association</b>	Program	Southern Coast of Portugal	Monitoring of marine megafauna in coastal waters of Algarve, Portugal. Characterization of whale-watching activities along the Southern coast of Portugal.  ( <a href="https://pt.aimmportugal.org/">https://pt.aimmportugal.org/</a> )	Determine which species live in southern Portuguese waters, why and how they use their habitat. Assess their conservation status and the potential threats they face. Understand how these animals use sound, and detect and locate them in their habitat. Contribute to the improvement of whale-watching activities.
<b>SOMAR</b>	Program	Algarve Coast	Study of cetaceans through acoustics. Develop environmental education actions, research, and promote community involvement.  ( <a href="https://somarbio.pt/">https://somarbio.pt/</a> )	Understand the mechanisms of production and reception of acoustic signals by marine organisms, how animals use these signals in their natural habitat and what interference can be related to anthropogenic impacts.
<b>CRAM - Centro de Reabilitação de Animais Marinhos</b>	Centre	Central and Northern Coast of Portugal	Rescue, collection, and rehabilitation of marine animals. Monitoring of marine megafauna through coastal censuses, on dedicated platforms and acoustics. Tracking of birds and sea turtles.  ( <a href="https://cram.org.pt/">https://cram.org.pt/</a> )	Rehabilitation and return of marine animals to their habitat. Research on veterinary, biological, and ecological aspects of marine animals. Assess the conservation status of threatened marine species and understand the consequences of economic and political activities on the marine environment. Determine marine animals' distribution.

<b>Whale Tales Project</b>	Project	Autonomous Region of Madeira	Monitoring of sperm whales ( <i>Physeter macrocephalus</i> ) through visual censuses, photo-ID, satellite biomarkers, and biopsies.  ( <a href="http://www.mare-centre.pt/pt/proj/whale-tales-project">www.mare-centre.pt/pt/proj/whale-tales-project</a> )	Increase scientific knowledge on habitat use and the physiological condition of sperm whales ( <i>Physeter macrocephalus</i> ) in the Macaronesian island waters. Deliver habitat models.
<b>META - Marine mammal and Ecosystem: Anthropogenic Threat Assessment</b>	Project	Autonomous Region of Madeira	Study behavioural changes in cetacean distribution and individual movement, and physiological changes.  ( <a href="https://meta.madeirawhalemuseum.org/">https://meta.madeirawhalemuseum.org/</a> )	Assess potential changes induced by anthropogenic threats within the studied resident cetacean population in Madeira. Evaluate the socio-economic impact and carrying capacity of whale watching activity. Support efficient management and cetacean conservation in Madeira waters.
<b>MONICET</b>	Project	Autonomous Region of the Azores	Collect, organize, and disseminate cetacean distribution data, and photo-ID images collected by whale watching companies in the Azores.  ( <a href="https://fgf.uac.pt/en/content/meemo-keep-expand-and-explore-monicet-platform-cetacean-watching-opportunity-science-0">https://fgf.uac.pt/en/content/meemo-keep-expand-and-explore-monicet-platform-cetacean-watching-opportunity-science-0</a> )	Analyse temporal trends of cetacean occurrence in the Azores in relation to oceanographic, atmospheric, and anthropogenic variables. Generate distribution estimates useful for species management.
<b>MARCET II</b>	Project	Macaronesian Region	Eco-sanitary studies of cetaceans. Create/Implement infrastructures for the cetaceans and other threatened marine species health surveillance. Promotion of tourism, business and science among whale watching in Macaronesia. Develop environmental education actions.  ( <a href="https://marcet-mac.eu/">https://marcet-mac.eu/</a> )	Assess anthropogenic impacts that affect the conservation of resident cetaceans in SACs of interest for whale watching activities. Determine oceanographic and anthropic use of the marine areas where cetacean species reside. Encourage the eco-tourist activity of whale watching as a model of sustainable economic development in the Macaronesian region.

<b>Sail &amp; Whale</b>	Program	Atlantic Ocean	Collate sightings of cetaceans using citizen science. Promote the dissemination of experiences and knowledge in favour of the conservation of the marine environment.	Studying the migration of cetaceans in the Atlantic.
-------------------------	---------	----------------	--	--

[\(http://sailandwhale.com/sail-whale-pt/\)](http://sailandwhale.com/sail-whale-pt/)

## 1.2 Reliability and Re-usability of Data: Verification and Validation Processes

To ensure that monitoring efforts can be rewarded, the data collected must be reliable. An editorial in the scientific periodical “Nature” argued that “an accurate and reliable record of what is going on can trump any particular strategy for trying to understand it” (Nature Publishing Group, 2007). Only in this way threats to the environment can be addressed, and policies and legislation can be developed to monitor and protect vulnerable areas, understand trends, and forecast future changes (Martín Míguez et al., 2019).

For this purpose, and to ensure the reusability of data, it is important to guarantee a rigorous and standardized method of collection. This deserves special attention when data is collected by several people, such as in the case of monitoring programs that rely on a network of researchers, students, and/or volunteers. Survey methods and training techniques can determine the success or failure of data collection.

When monitoring species occurrence, a representative spatio-temporal coverage at regular intervals is ideal. Animals are not distributed randomly in space; hence, the survey design must be homogeneous and representative, but this represents another challenge to marine surveys logistics (Evans & Hammond, 2003; Smith et al., 1986). Representational and homogeneous coverage of the marine environment is highly impractical due to the high logistical and economic costs. To work around this issue, targeted small-boat surveys and data collected aboard OPOs or non-randomized surveys can be cost-effective ways to fill knowledge gaps, but these approaches have their own sets of geographical distribution constraints (Kaschner et al., 2012). Ideally, sets of equally spaced parallel lines or a standard zigzag pattern design, starting from a random point along one edge of the survey area, should be taken (Evans & Hammond, 2003).

Likewise, it is important to have a seasonal survey design with data collected year-round, which is also a challenge. Specifically with programs relying on OPOs, having homogeneous effort coverage year-round is usually not possible due to financial and logistics reasons, and due to sampling designs not being dependent on researchers (Williams et al., 2006; Kiszka et al., 2007).

In addition, marine surveys are highly influenced by weather conditions such as visibility, sea state, and wind state, usually limiting survey effort. For these reasons, winter months are usually under-surveyed. Even if the weather conditions do not impede carrying out the surveys, they are factors of bias, likely leading to underestimations of species diversity and abundance (Evans & Hammond, 2003). A way to counter this problem is by collecting effort

data associated with the survey (Evans & Hammond, 2003; Williams et al., 2006). This will allow to calibrate and standardize the heterogeneous effort (spatial and temporal), and to ensure the correct use of data.

An appropriate methodology of data collection, supported by a strong survey design, can help to guarantee confidence in the legitimacy of findings on species abundance and density. Nevertheless, a careful verification and validation process of the data collected is required. Table 4, shows some examples of processes to ensure data quality, from data collection to processing, considering different types of monitoring datasets.

**Table 4.** Examples of different types of technical processes applied to different monitoring datasets, from data collection to data processing, to ensure data quality. ENA – Eastern North Atlantic. MMO – Marine Mammal Observer. ID – Identification. QA – quality assurance. QC – quality control.

Dataset	Study Area	Type of Data	Processes for Data Quality
A dataset of cetacean occurrences in the Eastern North Atlantic (Correia et al., 2019a) <sup>1</sup>	ENA	Cetacean's occurrence	<p>Intensive training of MMOs on both the sampling protocol and marine mammals' ID.</p> <p>Selection of MMOs according to their interest in participating in the project and previous experience on cetacean ID and fieldwork at sea.</p> <p>At least one of the two MMOs boarding must have experience in the survey protocol.</p> <p>Registration of IDs are only made to the taxonomic level MMOs are confident with.</p> <p>Collection of effort data.</p> <p>Checking and revision of incongruous data.</p> <p>Verification of the geographic information in ArcGIS (<a href="https://www.esri.com">https://www.esri.com</a>).</p> <p>Validation process in R (<a href="https://www.r-project.org/">https://www.r-project.org/</a>) to check specific errors derived from digitalisation.</p> <p>Annual verification and processing of data.</p>
Long-term surveys of age structure in 13 ungulate and one ostrich species in the Serengeti, 1926–2018 (Rogy & Sinclair, 2020)	Serengeti Ecosystem, Tanzania	Sample counts of 13 ungulate and one ostrich species	<p>Intensive training of observers.</p> <p>Sampling of all herds seen along transects was designed to provide an unbiased measurement of recruitment success in the populations relative to the number of females.</p> <p>Observations are based on a subset of data where the sexes cannot be distinguished, and on published research.</p> <p>Recording together as adults both males and females of zebras (<i>Equus quagga</i>) and warthogs (<i>Phacochoerus africanus</i>), since the sexes of these species cannot be identified with certainty.</p>
Long-term monitoring of the Iberian ibex population in the Sierra Nevada of the southeast Iberian Peninsula (Granados et al., 2020)	Sierra Nevada, Eastern South Iberian Peninsula	Data on the abundance and demographic structure of the Iberian ibex population	<p>Cross-checking of the sightings in situ during the sampling.</p> <p>Input masks control-data entry formats.</p> <p>Required fields are defined and lists of predefined values are made.</p> <p>Establishing of "control fields".</p> <p>Checking and revision of incongruous data.</p> <p>Verification of the geographic information in ArcGIS (<a href="https://www.esri.com">https://www.esri.com</a>).</p> <p>Validation process in R (<a href="https://www.r-project.org/">https://www.r-project.org/</a>) to check specific errors derived from digitalisation.</p>



<p>Fifteen-year record of soil temperature at the Bear Brook Watershed in Maine (Patel et al., 2018)</p>	<p>Temperate Forests, Bear Brook Watershed, Maine, United States</p>	<p>Data on soil temperature</p>	<p>QA procedures on data loggers.                  QC procedures on temperature data.                  Spatial consistency among sensors.                  Bias testing and assessment of the effect of replication.                  Consistency with National Oceanic and Atmospheric Administration station data.</p>
<p>Long-term dataset on aquatic responses to concurrent climate change and recovery from acidification (Leach et al., 2018)</p>	<p>Southwestern and South-central Adirondack Park, New York, United States</p>	<p>Record of physical, chemical, and biological measurements</p>	<p>Analysing of proficiency samples every six months to assure quality control.                  Recounting of samples, and consultation with outside experts for taxonomic verification.                  Photographic validation.                  Maximizing of the subsample-to-sample ratio for all zooplankton samples to limit multiplication errors.                  Performing of duplicates counts for all zooplankton samples from every tenth lake.                  Assessment of the analytical precision for all water chemistry data.                  Performing of QA/QC steps to verify that there are no data processing errors between the raw source files and final data tables.                  Manually checking of a random 1% of each datatype.                  Manually checking of all physical data.</p>

<sup>1</sup> The data quality processes of this dataset are described in more detail in subchapter “1.4 CETUS Project: Open-source Dataset”.

For cetacean, the importance of critical evaluation using visual records and/or descriptions cannot be overemphasized (Evans & Hammond, 2003). Specifically, the verification and validation of photographic records can be a good support for species ID since some species are difficult to differentiate and can lead to misidentifications, especially at sea and at a distance (Evans & Hammond, 2003; Smultea et al., 2010). Besides that, most observers tend to underestimate group sizes, and occasionally overestimate (Boyd et al., 2019).

Photographic verification/validation is a widely used process since photographs allow a sighting record to be analysed more objectively. However, a heavy reliance on photographs may introduce a bias in relative numbers because some species are easier than others to photograph or identify (Evans & Hammond, 2003). Moreover, this approach can only work if the subjects' photographs or videos can be effectively obtained under fieldwork conditions (Gordon, 2001).

Visual records, besides supporting the verification and validation of the collected data, can also help to infer the data quality of a dataset by assessing, for example, the percentage of accurate and wrong IDs. In this regard, it may also be useful to test the observers' survey skills, which can be done: i) a priori of the surveys, during a training period, by creating digital quizzes and testing ID skills with the support of photographs of a subset of animals; ii) during fieldwork, through the supervision of an expert; iii) or a posteriori of the surveys, through the detection rate and species ID success (Kosmala et al., 2016; Robbins et al., 2020). The bias related to the observer's experience is also important to acknowledge. Besides different capabilities in detecting and identify species, their experience with the sampling protocol and the environment of the survey (e.g., sampling marine offshore areas aboard large vessels is substantially different from near-shore surveys aboard small vessels) are also important. In this case, the selection of the observers is a key step. Nevertheless, it is almost impossible to guarantee homogeneous experience among different observers, so it is fundamental to assess the associated bias of heterogeneous experience among observers.

According to Greenland (2005), bias modelling should be part of the core training of researchers who are entrusted with observational data analysis. This applies even if the data is collected by experienced and trained observers, as most forms of bias observed in citizen-science datasets are also found in professional datasets, and can be mitigated using multiple-bias modelling (Greenland, 2005; Kosmala et al., 2016).

Detectability factors, such as weather conditions (e.g., sea state), the height of the observation platform, and distance of sighting to the vessels, are often included when developing ecological niche modelling of cetaceans (based on visual records) to account

for the influence of these variables. However, the assessment of the influence of observers' experience and reliability in the detection rate is still largely disregarded in ecological models of cetaceans (Correia et al., 2019b; Cominelli et al., 2013; Cominelli et al., 2014).

### 1.3 Long-term Monitoring Datasets: Challenges and Opportunities

Long-term datasets on species occurrence are compilations of a variety of data on one or more taxon, at one or more locations, over a long period of time. This type of data allows the identification of temporal trends, shifts in spatial distribution, abundance, and biodiversity hotspots, which makes them a very powerful conservation tool, providing baseline data to better understand physical and ecological responses to environmental changes (Magurran et al., 2010).

The Convention on Biological Diversity, currently signed by 168 countries, has highlighted the growing need for baseline data to reduce the rate of biodiversity loss, and therefore reinforcing the importance of long-term monitoring datasets (Convention on Biological Diversity, 2000; Magurran et al., 2010). For this reason, and to ensure good data quality and its re-usability, in 2016, the “FAIR Guiding Principles for scientific data management and stewardship” was published in the Journal “Scientific Data”, providing guidelines to improve the Findability, Accessibility, Interoperability, and Reusability of all research objects (Wilkinson et al., 2016). These will lead the resources along the continuum towards their optimal state (Table 5).

**Table 5.** The FAIR Guiding Principles (Wilkinson et al., 2016).

To be Findable	To be Accessible
F1. (meta)data are assigned a globally unique and persistent identifier.	A.1 (meta)data are retrievable by their identifier using a standardized communications protocol.
F2. data are described with rich metadata (defined by R1 below).	A1.1 the protocol is open, free, and universally implementable.
F3. metadata clearly and explicitly include the identifier of the data it describes.	A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
F4. (meta)data are registered or indexed in a searchable resource.	A2. metadata are accessible, even when the data are no longer available.
To be Interoperable	To be Reusable
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	R1. meta(data) are richly described with a plurality of accurate and relevant attributes.
I2. (meta)data use vocabularies that follow FAIR principles.	R1.1. (meta)data are released with a clear and accessible data usage license.
I3. (meta)data include qualified references to other (meta)data.	R1.2. (meta)data are associated with detailed provenance.
	R1.3. (meta)data meet domain-relevant community.

To be findable is the first step in (re)using data. Metadata and data should be easy to find. Once the user finds the required data, it is necessary to know how to access it and this is perhaps one of the biggest challenges in science. Universities, granting agencies, and

publishers each have different incentives for researchers and many of them lack clear motivation to be more open, even though the more data is made available, the more information can be accessed by the entire research community (Nosek et al., 2015). Collaborative efforts to involve all stakeholders to complement and coordinate incentives to drive research practices towards more open science is a necessary ground-breaking step. The AtlantOS initiative is a good example of that. This project managed to join 18 countries (13 EU & 5 non-EU) to improve and innovate Atlantic observing to obtain an international, more sustainable, more efficient, more integrated, and fit-for-purpose system (Deyoung et al., 2019).

To comply with the third principle, interoperability, usually, it is important to combine the required data with other information. Moreover, the data must interoperate with applications or workflows for analysis, storage, and processing. In other words, data must have the ability to integrate or work together with minimal effort (Wilkinson et al., 2016).

Lastly, the ultimate goal of FAIR is to optimize and enhance the reusability of data. To accomplish this, metadata and data should be well-described so that they can be reproduced and/or combined (Wilkinson et al., 2016).

Open science is a growing movement and databases such as the European Marine Observation and Data Network (EMODnet) and the Ocean Biodiversity Information System (OBIS) already work under FAIR principles. Both of these databases are open-access and store data from the marine environment, although EMODnet stores data at the European scale and OBIS at a global scale (Martín Míguez et al., 2019; Tanhua et al., 2019). These databases have gone beyond providing access to metadata and data. They have developed networks and communities of researchers/specialists who work to encourage the development and adoption of FAIR Principles, sharing best practices and promoting integration and interoperability between various systems (Martín Míguez et al., 2019).

Currently, there are over 5,000 datasets published in these two online databases. On EMODnet, the data provided ranges from data on bathymetry to data on marine litter or concentration of nutrients, among many other topics. OBIS provides data on the diversity, distribution, and abundance of all marine organisms, with a focus on georeferenced occurrence data. Datasets are available for many taxonomic groups, including data on cetaceans.

Long-term datasets can play an important role in marine management and conservation, so they must be public, following FAIR Data Principles, to enhance their reusability by stakeholders, from the scientific community to decision-makers. To achieve the ultimate goal of the FAIR principles, maximizing the use and value of the data collected, not only the

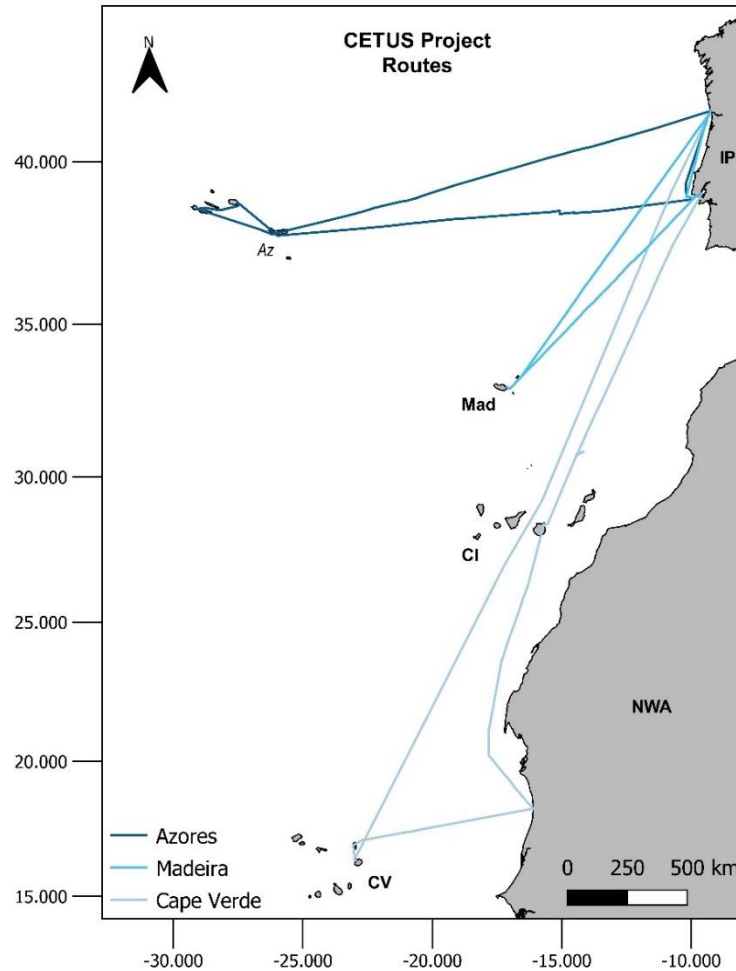
data should be made available, but also the quality of the data should be known, and the type and quality of the data should be consistent and comparable (Shampine, 1993). Thus, to get the most out of these datasets, it is necessary that the data is reliable and that the sources of bias are identified and quantified.

## 1.4 CETUS Project: Open-source Dataset

CETUS Project is a cetacean monitoring program, aiming at a long-term data frame within the ENA region. The backbone of the project is a partnership with Transinsular, a Portuguese company for maritime transport. Transinsular provides its cargo ships to be used as OPOs by the research team to survey the routes between Continental Portugal, Macaronesian archipelagos, and West Africa (Figure 3). In total, from 2012 to 2019, 61 MMOs from all over the world participated in the campaigns on a voluntary regime. The main goal of the monitoring program is to provide cetacean occurrence data in the ENA. In addition to cetacean sighting records, or the occurrence of other pelagic megafauna, data on survey effort, weather conditions, and marine traffic were also collected during the surveys (Correia et al., 2019a).

Despite that in 2020, due to the COVID-19 pandemic and its travel limitations, this monitoring program is still ongoing and collecting data, to build the first long-term, wide-range, open-source dataset on cetacean occurrence and distribution in the ENA (Correia et al., 2019a). The data has been used to study spatio-temporal distribution, species diversity, habitat characterization, ecological niche modelling and it is now being applied to predict cetacean habitat under climate change scenarios. The CETUS dataset has served as a baseline for scientific publications (Correia et al., 2015; Correia et al., 2019a; Correia et al., 2019b; Correia et al., 2020; Correia et al., 2021; Valente et al., 2019), academic thesis (Correia, 2013; Correia, 2020; Gil, 2018; Valente, 2017) and conservation reports (ASCOBANS, 2017; ICES, 2016).

The CETUS dataset is made available open-access at OBIS and EmodNET portals, presently with the data spanning from 2012 to 2017 (Correia et al., 2019a). Therefore, a series of concerns were raised to ensure it follows the FAIR Guiding Principles and, as a result, to be brought to its optimal state.



**Figure 3.** Study area and routes of the CETUS Project. IP – Iberian Peninsula; NWA – Northwest Africa; AZ – Azores; MAD – Madeira; CI – Canary Islands; CV – Cape Verde.

Monitoring programs like CETUS face many challenges in controlling the quality, reliability, and usability of data collected. The fieldwork involves a large network of MMOs, with teams changing every year, and data collection being undertaken by individuals with different levels of experience. Furthermore, there is limited control over the survey design (completely dependent on the OPOs schedules) which results in a heterogeneous survey effort in time and space (Correia et al., 2020; Evans & Hammond, 2003; Williams et al., 2006; Kiszka et al., 2007; Nerbonne, 2003). Ultimately, the height of the observation platform (usually varying between vessels), meteorological conditions and animal behaviour also bias and influence species detection (Bailey et al., 2004; Bas et al., 2008). These problems are especially critical when considering remote areas or unusual and cryptic species (Cominelli et al., 2015).

To deal with these challenges, the CETUS Project undertakes several processes to minimize potential bias and ensure data quality (Table 4). Since 2014, and every year, CETUS Project selects the MMOs through an international call that gives priority to



volunteers with previous experience in sea surveys and in the ID of marine mammals, as well as the motivation in the internship. Besides that, whenever possible, each vessel receives a team of two MMOs, one of whom is often active and comfortable with the CETUS methodology. All MMOs participate in an intensive course on line-transect survey protocol and marine mammals ID before they embark (Correia et al., 2019a).

Each MMO stands on one side of the ship with a field of view of approximately 90°, covering in total 180° and the survey is performed from sunrise to sunset. To avoid fatigue and biases related to the vessel side, they switch every 60 minutes. Furthermore, both take one-hour breaks for meals and two optional rests of up to 40 minutes, ideally in turns. During these periods, the lone MMOs survey the 180° (Figure 4; Correia et al., 2019a).



**Figure 4.** Monte da Guia vessel arriving to Port of Leixões with indication of the position of the Marine Mammal Observers (MMOs) during CETUS Project surveys, and the direction of the route.

In terms of data collection, the ship's route and the positions are recorded by the MyTracks application (<https://my-tracks.pt.aptoide.com>) installed on a tablet with an inbuilt Global Positioning System (GPS). Among other variables, this application registers the date and time, the speed and direction of the vessel, and the coordinates of the GPS (Correia et al., 2019a). However, due to battery life issues or other complications, now and then errors are generated in the date and time recording. For this reason, during data entry, careful verification processes are required. These are controlled manually by members of the team during the digitalization of the data to Microsoft Excel (<https://www.microsoft.com/pt-pt/microsoft-365/excel>).

Quantification of effort is required; this will allow the standardization of the data, and minimizing the impact of a heterogeneous effort coverage (Evans & Hammond, 2003; Williams et al., 2006). For this reason, the entire survey effort is recorded (Correia et al., 2019a). “On-effort” (i.e., periods of active survey) or “off-effort” (i.e., periods of interrupted survey effort) conditions are based on 4 meteorologic variables: sea state (using the Douglas scale), wind state (using the Beaufort scale), visibility (on a categorical scale of values from 1 – 10 estimated based on the definition of the horizon line and reference points at a known range (e.g., ships with an automatic identification system), and the occurrence of rain (Table 6). Moreover, whenever a proper dedicated survey effort of the 180° is not possible (e.g., upon a sighting, when it is not possible to stand in the observation deck, often due to cleaning or security drills), this period is considered “off effort”. All data collected “off effort”, is treated as opportunistically collected.

**Table 6.** Meteorologic variables assessed during CETUS Project surveys. Indication of meteorologic conditions when sampling is active (i.e., “on effort”, marked in green) and conditions when sampling is considered opportunistic (i.e. “off effort”, marked in red).

Sea State (Douglas scale)			Wind State (Beaufort scale)	
Code	Height (m)	Description	Code	Description
0	0	Calm (glassy)	0	Calm (oily/mirrored sea)
1	0 – 0.1	Calm (rippled)	1	Light air (smooth sea)
2	0.1 – 0.5	Smooth (wavelets)	2	Light breeze (looks like raining)
3	0.5 – 1.25	Slight	3	Gentle breeze (little white spots)
4	1.25 – 2.5	Moderate	4	Moderate breeze (several white spots)
5	2.5 – 4.0	Rough	5	Fresh breeze (white lines)
6	4.0 – 6.0	Very rough	6	Strong breeze (white lines w/ splashes)
7	6.0 – 9.0	High	7	Moderate gale (white lines w/ stretches)
8	9.0 – 14.0	Very high	8	Fresh gale
9	Over 14.0	Phenomenal	9	Strong gale
			10	Whole gale
			11	Storm
			12	Hurricane

Visibility (intuitive)		Rain	
Code	Distance (m)	Code	Description
1	< 50m	1	No rain
2	50 to 199 m	2	Little rain
3	200 to 499 m	3	Medium rain
4	500 to 999 m	4	Lots of rain
5	1000 to 1999m		
6	2000 to 3999 m		
7	4000 to 9 999 m		
8	10 000 to 19 999 m		
9	20 000 to 50 000 m		
10	> 50 000 m		

When a sighting occurs, the identity assigned is always at the taxonomic level at which the MMOs are very confident of their species ID. Sighting distance and angle are also recorded. These can then be used to calculate the approximate distance to the animals with an estimated observation height (Correia et al., 2019a). However, measuring distance at sea is notoriously difficult. Hence, many cetacean sighting surveys use binoculars marked with reticules, which is the case. Under good conditions and with adequate training, these measurements can provide good distance estimates. Otherwise, they can introduce a systematic bias that may vary among MMOs and OPOs (i.e., technique of using binoculars and reading distances, the height of the observation deck of the vessel, and height of the eye-level of the observer; Williams et al., 2007). Additionally, in the case of cargo ships such as those used in CETUS, the height of the vessel is highly dependent on the weight of the cargo being carried. All of these potential errors are acknowledged and identified in the information supporting the CETUS Project dataset (Correia et al., 2019a).

Moreover, as the compasses of the binoculars can be unreliable on platforms containing ferrous metals, when an animal is sighted, the vessel heading is also measured. This value is then compared with the direction of the route as calculated by the GPS to obtain the estimated compass error and correct the horizontal angle reported (during data processing) (Correia et al., 2019a).

Following data processing, all records are imported into MySQL database and restructured into the appropriate relational format using R (<https://www.r-project.org/>). ArcGIS (<https://www.esri.com>) is then used to better visualize the data, perform numerous verifications, and correct occasional inaccuracies in the coordinates (Correia et al., 2019a).

Despite CETUS verification and validation methods being rigorous, they are mostly related to data collection and processing, i.e.: i) embark of motivated, highly trained, and dedicated observers; ii) standard protocol for line-transect data collection; iii) recording of survey effort; iv) register of potential sources of bias during fieldwork; v) multiple verification steps in data entry and processing. None of these includes photographic verification/validation methodologies or the assessment of data bias in terms of species detection and identification.

## 1.5 Aims

This work aims to further optimize and allow the proper use of the CETUS dataset by (1) applying new verification and validation methods based on photographic confirmation of identified species using photographs obtained from MMOs; (2) creating quality criteria related to the MMOs experience (through *curricula vitae*); and (3) assessing the influence of bias parameters by modelling sightings using explanatory variables accounting for kilometres sampled “on-effort”, weather conditions and the experience of MMOs.

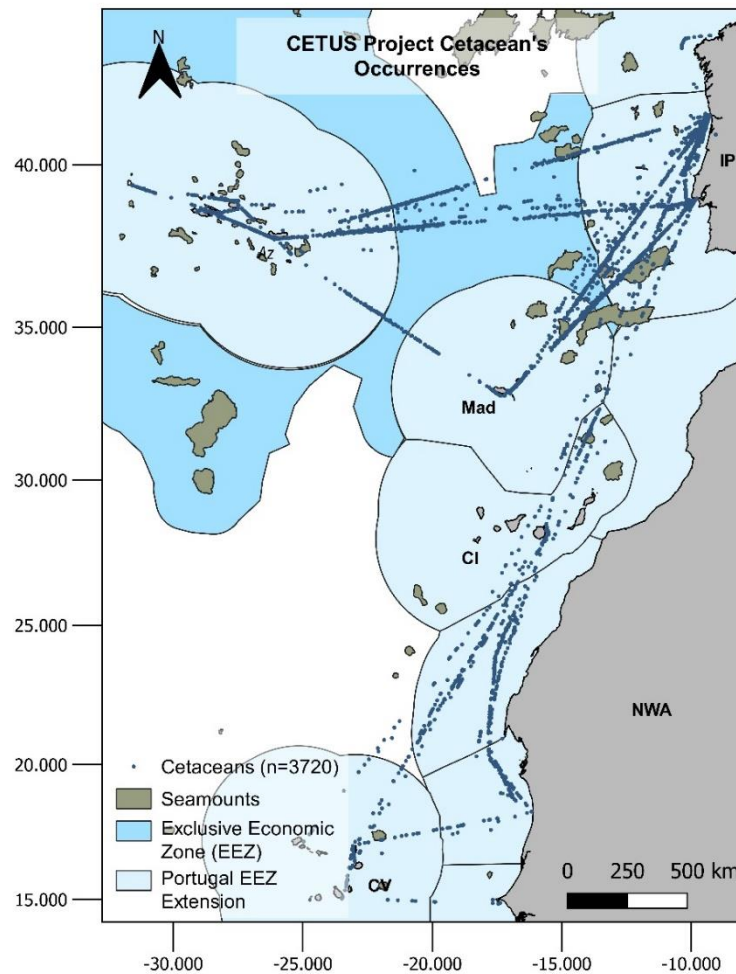
Results will also be used to determine useful improvements in the CETUS logistics, such as guidelines to i) improve the collection of photographic/video recordings for a better verification/validation process; ii) organize MMOs’ teams considering their experience; iii) define meteorological limits for monitoring. Other possible flaws found can be amended by identifying adequate and feasible procedures for the optimization of reliable data collection.

Ultimately, the work will contribute to an adequate and informed use of the CETUS dataset, expanding its application in science, and supporting marine management and conservation. It will also optimize and create verification and validation processes to be routinely applied not only in the CETUS Project but also in identical monitoring programs.

## 2. Materials and Methods

### 2.1 The CETUS dataset: Study Area, Data, Structure

To define the study area for this study, the following geographical limits were considered: the upper limit at 42° N (north), lower limit at 12° S (south), eastern limit at 6°, and western limit at 32° (Figure 5). This area was based in the CETUS Project surveyed area, within the ENA.



**Figure 5.** Study area with cetacean occurrences. IP – Iberian Peninsula; NWA – Northwest Africa; AZ – Azores; MAD – Madeira; CI – Canary Islands; CV – Cape Verde.

With archipelagos emerging from deep waters and a complex bathymetry, ENA is influenced by several main oceanographic features (e.g., coastal upwelling, ocean currents, mesoscale eddies; Mason, 2009; Sala et al., 2013). This results in a dynamic topographic and oceanographic system that supports a diverse cetacean community, including coastal and oceanic, transient and resident, and tropical and subtropical species, that in turn play

a major role in the biomass distribution (Caldeira & Sangrà, 2012; Correia et al., 2020; Redfern et al., 2006).

Between 2012 and 2019, 3720 occurrences of cetaceans were recorded. These sightings correspond to a minimum of 27 different species, identified, at least, to the genus (Table 7).

**Table 7.** Number of cetacean occurrences. The number of occurrences is presented by taxa recorded to the highest possible level. The table is organized by taxon rank of the records and alphabetically within.

Taxa	Taxon Rank	Number of Occurrences	Taxa	Taxon Rank	Number of Occurrences
Cetacea	Infraorder	430	<i>Megaptera novaeangliae</i>	Species	9
Mysticeti	Superfamily	359	<i>Mesoplodon densirostris</i>	Species	8
Delphinidae	Family	1014	<i>Orcinus orca</i>	Species	8
Ziphiidae	Family	182	<i>Peponocephala electra</i>	Species	4
<i>Globicephala</i>	Genus	67	<i>Phocoena phocoena</i>	Species	9
<i>Kogia</i>	Genus	7	<i>Physeter macrocephalus</i>	Species	178
<i>Balaenoptera acutorostrata</i>	Species	93	<i>Pseudorca crassidens</i>	Species	13
<i>Balaenoptera borealis</i>	Species	4	<i>Stenella attenuata</i>	Species	9
<i>Balaenoptera edeni</i>	Species	6	<i>Stenella clymene</i>	Species	17
<i>Balaenoptera musculus</i>	Species	3	<i>Stenella coeruleoalba</i>	Species	181
<i>Balaenoptera physalus</i>	Species	36	<i>Stenella frontalis</i>	Species	328
<i>Delphinus delphis</i>	Species	473	<i>Stenella longirostris</i>	Species	6
<i>Grampus griseus</i>	Species	13	<i>Steno bredanensis</i>	Species	4
<i>Hyperoodon ampullatus</i>	Species	5	<i>Tursiops truncatus</i>	Species	171
<i>Lagenodelphis hosei</i>	Species	1	<i>Ziphius cavirostris</i>	Species	79
<i>Lagenorhynchus albirostris</i>	Species	3	Total	31 taxa	3720 occurrences

Data spanning from 2012 to 2017 has already been published (Correia et al., 2019c). The data collected between 2018 and 2019 was compiled and processed, following all the procedures in the validation and technical verification of described in the data descriptor of the published CETUS dataset (Correia et al, 2019a), and was compiled into the internal dataset.

At the moment, the updated internal CETUS dataset consists of a Microsoft Excel (<https://www.microsoft.com/pt-pt/microsoft-365/excel>) document with the relevant information collected from MMOs, which is structured and compiled according to the dataset published at OBIS. This document is divided into 4 datasheets: i) cruise, ii) survey, iii) segment and iv) positions, i.e., all the georeferenced data recorded by the team of CETUS MMOs (Table 8).

In the cruise datasheet each register is a cruise that corresponds to a trip from one port to another. Information includes the route short description, the name of the ship, departure and arrival ports, and the name of the MMOs. The survey datasheet has all the information on each individual survey which corresponds to a continuous period of survey (i.e., a day from sunrise to sunset). Information includes the latitude and longitude points where the survey started and ended, the date, and the kilometres sampled “on-effort”. The segment datasheet includes the kilometres sampled per “on-effort” segment, which corresponds to a period of continuous “on-effort” sampling. Lastly, in the position datasheet is the information on cetaceans’ occurrence, as well as another marine megafauna, marine traffic, and weather waypoints.

**Table 8.** CETUS internal dataset structure with the information available for each datasheet (cruise, survey, segment, and positions). YY/YYYY – year. SL – ship's letter code. NNN – number of the cruise in that ship, in that year. DD – day. MM – month. S – survey number of the day. SN – ship's number code. EE – effort code. SSS – sighting number.

<b>Cruise</b>	<b>Route description</b> (e.g., Continental Portugal - Madeira - Continental Portugal)	<b>Cruise number</b> (i.e., YY SL – NNN)	Name of the <b>ship</b>	<b>Departure port and its latitude and longitude</b>	<b>Arrival port and its latitude and longitude</b>	Name of the <b>MMOs</b>
	<b>Cruise number</b>	<b>Survey number</b> (i.e., YYDDMM S SN)	<b>Latitude and longitude of the point where the survey started</b>	<b>Latitude and longitude of the point where the survey ended</b>	<b>Date</b> (i.e., DD MM YYYY)	<b>Effort</b> (i.e., kilometres sampled “on-effort” (Table 6 for “on/off-effort” meteorological conditions))
<b>Survey</b>	<b>Survey number</b>	<b>Segment number</b> (i.e., YYYYMMDD S SN EE)	<b>Distance</b> (i.e., kilometres sampled on each “on-effort” segment)			
<b>Segment</b>	<b>Survey number</b>	<b>Segment number</b> (available for “on-effort” positions only)	<b>Position number</b> (i.e., SN YY SSS)	<b>Effort</b> (i.e., “on” or “off”)	<b>Latitude and longitude</b>	<b>Platform speed and direction</b>
<b>Positions</b>						<b>Species, group size</b> (minimum, maximum, and best estimate), <b>bearing, reticules below the horizon and estimation method, behavior in relation to the ship</b>



Positions	<b>Number of small and big vessels</b> (smaller and bigger than 20 meters long)	<b>Sea, wind, visibility, and rain states</b> (Table 6 for scale information of the meteorological conditions)	<b>Comments</b>
-----------	--	---	-----------------

## 2.2 Photographic Verification/Validation

All 61 former MMOs who have integrated the CETUS Project between 2012 and 2019 were contacted and were asked to provide any available photographic/video of cetacean records they have collected during their internship.

Photographs/videos were organized in a folder hierarchy: by MMO, year, cruise and days monitored. Not all the records had metadata up to the day of recording, and these were inserted into the most appropriate folder (up to the cruise, year, or MMO). For each set of records corresponding to a sighting, those that allowed an easier ID were chosen as representative for that sighting. The remaining photos/videos were consulted in case of doubt (e.g., to look for some type of detail that would help with the ID).

Verification consisted of the process of matching the video/photographic records with the dataset sighting registers. As soon as the process started, it was noticed that often the date and/or time of the camera or mobile phone displayed in the metadata of the file, were wrong, non-existent, or with different time zones. To get around this problem, a conservative methodology was applied allowing the use of all available information to verify, match and validate as many sightings as possible without inaccuracies. This was achieved using the information documented upon the record of the sighting as well as the information provided by the MMOs themselves.

Thus, all the information provided by the file metadata (date and time) and indications provided by the MMOs (e.g., name of the file or folder with indication of the sighting ID or day of the record) started being considered. To determine the time difference between the timestamp in the metadata and the real time of the records, at least two obvious sightings were considered (e.g., unique sighting of the day, close to the boat, easy ID, calm sea state, or Universal Time Coordinated (UTC) /  $UTC \pm X$ ). The time difference of those two obvious sightings was then considered for the other photos/videos of the MMO on that cruise or vessel.

When the time difference between two sightings of the same species/group was less than the assessed time difference between metadata and sightings registers, the sequence of sightings vs. sequence of photos/videos, the time sequence of photos, and time of beginning and ending of sightings were considered.

When it was not possible to determine the time difference between the metadata timestamp of the files and the sighting registers (e.g., few pictures of that MMO for that cruise or vessel, photos/videos without time, meaningless hours, or too much variation), it was considered whether the sighting was too obvious and the complementary information assessed (e.g.,

the number of animals or the side of the sighting were unique for that sighting on that day and for that species/group).

After the verification process, the validation of the matched records was carried out, to confirm or correct the sightings ID of the CETUS dataset. A conservative methodology was followed, which involved, for more dubious ID, the discussion between four CETUS team members; and, if any doubts remained, an external expert on cetacean ID was consulted. Positive identifications from the photographic/video records required 100% certainty.

The matching sighting records of the dataset were grouped in the:

- “Complete Validation” (C-V) if it was possible to complete the photographic/video ID process and reach to the species level. In this group, we categorized the matched sightings as: “yes”, i.e., when the photographic/video ID corresponded to the MMOs ID; and “wrong”, i.e., when the MMOs ID was wrong.
- “Non-Validation/Incomplete Validation” (IN-V) if the records did not allow any ID or if the records only allowed an incomplete photographic/video ID process, not reaching the species level. In this group, we categorized the matched sightings as: “yes”, i.e., when the photographic/video ID corresponded to the MMOs ID; “no”, i.e., when the photographic/video records did not allow for the confirmation of the MMOs ID (e.g., due to poor-quality images); and “to the order” or “to the family”, i.e., when it was possible to validate a higher taxon than the one registered by MMOs.

The data on the photographic validation was compiled into the internal CETUS dataset (positions datasheet), onto the following columns: “validated” (with categories: “yes”, “wrong”, “no”, “to the order” and “to the family”); and “id after validation” for correct wrong IDs and for when it was possible to reach a lower taxon than the one registered by the MMOs. On the next CETUS dataset update, sightings will already have the correct ID (i.e., a delphinid sighting validated as common dolphin, will appear as common dolphin) with the information stating it has been photographically validated. This will allow future users to only use validated data, allowing a selection of data according to what the users want.

After verification/validation, a descriptive analysis was carried out to examine the success of obtaining the records, matching the sightings (i.e., verification), and validating them (i.e., validation). The aim was to discern whether this methodology yields valuable information and if it could be successfully applicable in the CETUS and other identical monitoring programs. For this analysis, results were compared for the suborders Odontoceti (i.e., toothed whales) and Mysticeti (i.e., baleen whales), and NI (non-identified) sightings.

To assess if the distance of the sightings to the vessel influenced the possibility to match or validate sightings from photographic/video records, boxplots of the relative distance (binoculars reticules) by suborder and NIs were created using the R software (Version 4.1.0). For comparisons of the median distances between matched and non-matched sightings, and C-V and IN-V sightings, Mann–Whitney tests were performed. For this, a significance level of 95% was considered (i.e., p-value of  $< 0.05$  deemed significant).

## 2.3 Creating a Data Quality Criteria: MMOs Experience

For the creation of the quality criteria, the experience of the MMOs was evaluated based on the information collected from their *curricula vitae* (CV). For this purpose, the following information was considered: (1) the experience at sea, (2) the experience with cetaceans' ID, (3) the number of species they have worked with, and (4) the experience working with the CETUS Project protocol. Each of these evaluation criteria was ranked from 0 to 5 with the maximum cumulative score, the quality criteria, being 20 (Table 9).

This evaluation was carried out considering the *curricula vitae* of the MMOs at the beginning of the sampling year. Thus, a recurring MMO can have the *curricula vitae* evaluated more than once if he/she participates in the CETUS Project in more than one year.

After all the MMOs were evaluated, the final score of the least experienced observer (LEO) and the most experienced observer (MEO) of the MMOs team was attributed for each cruise. These data were then compiled into the internal CETUS dataset (cruise datasheet) onto the following columns: MEO and LEO. When there was only one observer, the single MEO assessment was placed, leaving the data in the LEO column as NULL. On the next CETUS dataset update, this information will already be available.

By using R software (Version 4.1.0), histograms were generated representing the frequency of surveys across the range of the MMOs experience, based on the data quality criteria developed, for the LEO and MEO on each cruise. Moreover, a bar plot was generated in Microsoft Excel (Version 2106) to represent the evaluation score for the LEO and MEO (and cumulative score) for each combination of scores of the MMOs teams. For cruises with one MMO, only one bar is represented (as the MEO).

**Table 9.** Evaluation criteria created to evaluate the experience of the CETUS Project Marine Mammal Observers and generate a data quality criterion. ID – Identification.

Experience at sea		Experience with CETUS Project protocol	
Owns diving/lifeguard courses	1	Up to 1 month on board	1
Has up to 3 months of experience on small boats	2	More than 1 month and up to 3 months on board	2
Has more than 3 months of experience on small boats	3	More than 3 months and up to 5 months on board	3
Has up to 3 months of experience on ships	4	More than 5 months and up to 7 months on board	4
Has more than 3 months of experience on ships	5	More than 7 months on board	5

Number of species MMOs have worked		Experience with cetaceans' ID	
Research directed to 1 species only with apparently no contact with other species	1	Up to 3 months of experience with cetaceans (e.g., whale watching, cetaceans monitoring or cetacean's rehabilitation)	1
Research directed to 1 or 2 species in areas with a medium number of species (e.g., Mediterranean)	2	More than 3 months and up to 6 months of experience with cetaceans	2
Research directed at 1 or 2 species in areas with a high number of species (e.g., Canaries or Azores)	3	More than 6 months and up to 1 year of experience with cetaceans	3
Generalized research of species in areas with a medium number of species	4	More than 1 year and up to 3 years of experience with cetaceans	4
Generalized research of species in areas with a high number of species	5	More than 3 years of experience with cetaceans	5

## 2.4 Bias Modelling of Number of Sightings

To assess the bias parameters on the number of sightings recorded per survey, generalized additive models (GAM) were performed in R (Version 4.1.0). The following detectability factors were considered as explanatory variables: weather conditions (i.e., the minimums and maximums of the sea state, wind state and visibility), experience of MMOs (i.e., the evaluation scores of LEOs and MEOs, as well as the mean and cumulative scores of the MMOs teams) and kilometres sampled “on-effort”. For this, only “on-effort” records of occurrence were used since “off-effort” segments are considered “off” not only due to poor weather conditions, but also when it is not possible to stand in the observation deck, often due to cleaning or security drills. This would result in a bias towards “off-effort” segments. If there were no MMOs monitoring, it is obvious that they could not detect any animals.

Prior to modelling, Pearson correlations were assessed between all pairs of explanatory variables and highly correlated variables were excluded, considering a threshold of 0.75 (Correia et al., 2020; Correia et al., 2021; Marubini et al., 2009). Moreover, multicollinearity among explanatory variables was measured through the Variance Inflation Factor (VIF), with a threshold of 3 (Correia et al., 2020; Correia et al., 2021; Zuur et al., 2010).

Considering that the response variables were counts, the saturated model including remaining all explanatory variables was first tested with a Poisson distribution (with a log link function; Correia et al., 2020). Then the model was checked for overdispersion through the Pearson estimator, i.e., the (weighted) sum of squares of the Pearson residuals, divided by the effective residual degrees of freedom. If the result is greater than 1, it tests positive for overdispersion. Since it tested positive for overdispersion (1.975499), a negative binomial distribution (with a log link function) was fitted (Correia et al., 2020).

To attain the best model, we started with a saturated model, followed by a backward selection (Correia et al., 2015; Qian, 2017; Correia et al., 2019b; Correia et al., 2020; Correia et al., 2021). The Akaike Information Criterion (AIC) was used as a measure of adequation of fitness, choosing the model with the lowest AIC value at each step of the model fitting process, i.e., comparing otherwise identical models with or without a specific explanatory variable. If the AIC-difference between two models was less than 2, the Analysis of Variance (ANOVA) chi-square test was used to check if the AIC-difference was significant (Zuur et al., 2007; Correia et al., 2020; Correia et al., 2021). If this difference was not statistically significant, the simplest model was kept. All GAMs were fitted in package “mgcv”.

A maximum of four splines ( $k = 4$ ) was chosen to limit the complexity of smoothers describing the effects of the explanatory variables (Correia et al., 2021; Quian, 2017). If a

spline was close to linear (with estimated degrees of freedom of  $\sim 1$ ), the smooth term was removed, and a linear function was fitted.

To interpret the final model “gam.check” function and multiple plots of interest were performed.



## 3. Results

### 3.1 Photographic Verification/Validation

Out of the 61 volunteers who took part in the CETUS monitoring programme, 54 (~88.5%) responded to our email. From those, 21 (~34.4%) provided photographic/video records. Hereupon, it was possible to gather photographs and videos of sightings recorded by 26 MMOs (~42.6%), since often there are two MMOs monitoring.

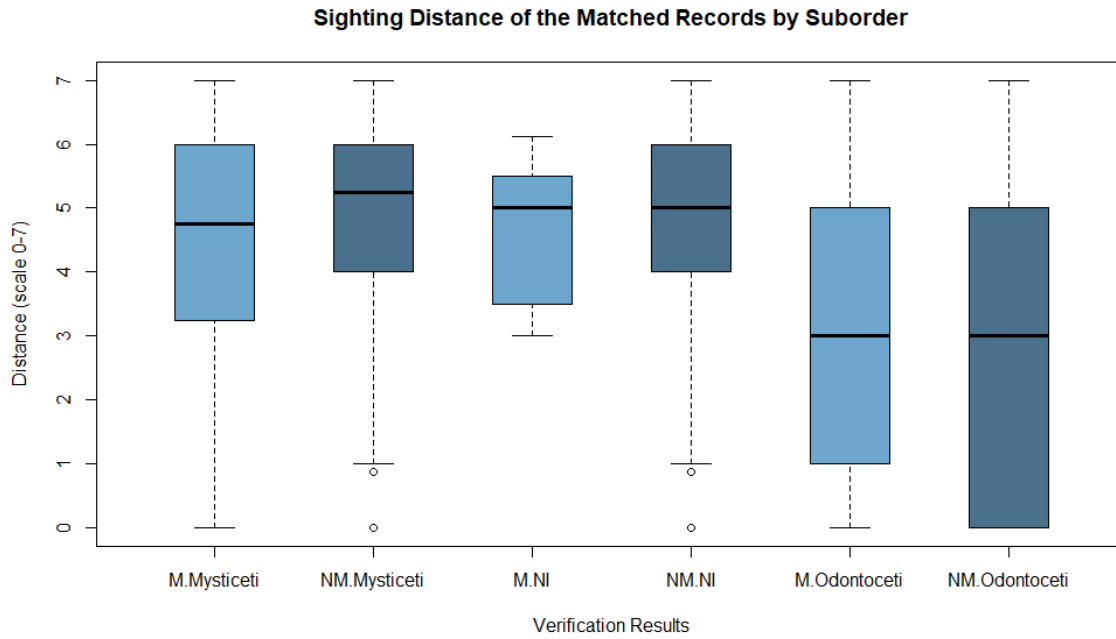
In total, were collated 9,970 photographic/video records, corresponding to 307 sightings, and we were able to match 279 (~90.9%) of them with the CETUS dataset records, corresponding to ~7.5% of the total sightings in the dataset.

Since 8 sightings were from two associated taxon (e.g., *Tursiops truncatus* sighted in association with *Globicephala* sp.), the total 279 matched (M) sighting records correspond to 287 single-taxon sightings. Best matching results correspond to the suborder Odontoceti (Table 10).

**Table 10.** Number and percentage of the matched (M) and non-matched (NM) sightings by suborder.

Matched Results		
	M	NM
<b>Odontoceti</b>	254 (9.15%)	2522 (90.85%)
<b>Mysticeti</b>	20 (3.89%)	494 (96.11%)
<b>NI</b>	13 (3.02%)	417 (96.98%)
<b>Total</b>	287 (7.72%)	3433 (92.28%)

When relating the distance of the sightings to the vessel of the matched and non-matched records by suborder (Figure 6), there were no significant differences (Table 15 in Appendix). However, when the sighting distance median from matched sightings between the suborders Odontoceti and Mysticeti were compared, this difference was significant ( $w = 1659.5$ ;  $p\text{-value} = 0.009557$ ). The same happened when the distance median from non-matched sightings between these two suborders were compared ( $w = 334814$ ;  $p\text{-value} = <2.2 \times 10^{-16}$ ), with odontocetes having a lower sighting distance median for matched and non-matched records.



**Figure 6.** Boxplot of the sighting distance of the matched (M) and non-matched (NM) records by suborder (Odontoceti and Mysticeti) and NI. NI – non-identified.

After the process (i.e., process of matching the video/photo records with registers in the CETUS dataset), the validation results (i.e., process of confirming/correcting the ID of the matched records with the video/photographs), yielded the highest percentage of completed validations for the suborder Odontoceti (Table 11). In total, there were 10 wrong identifications, which represent ~3.5% of the matched records. 9 within odontocetes (e.g., *Stenella attenuata* mistaken for *Stenella frontalis*) and 1 within an odontocete and mysticete (*Physeter macrocephalus* mistaken for *Balaenoptera borealis*).

**Table 11.** Number and percentage of the validation results of completed validated records by suborders and non-identified sightings (NI). Complete Validation represents all validations that reached the species level.

Validation Results					
		Odontoceti	Mysticeti	NI	Total
Complete Validation (n=170)	Yes	156 (94.5%)	3 (75%)	1 (100%)	160 (94.1%)
	Wrong	9 (5.5%)	1 (25%)	-	10 (5.9%)

Out of the sightings that could not reach the species level, Mysticetes had the highest percentage of non-validated (“no”) sightings (Table 12).

**Table 12.** Number and percentage of the validation results of non-validated/incomplete validated records by suborders and non-identified sightings (NI). Non-Validation / Incomplete Validation represents all validations that could not reach the species level.

Validation Results					
		Odontoceti	Mysticeti	NI	Total
Non-Validation / Incomplete Validation (n=117)	Yes	41 (46.6%)	5 (29.4%)	7 (58.3%)	53 (45.3%)
	No	21 (23.9%)	11 (64.7%)	5 (41.7%)	37 (31.6%)
	To the Family	26 (29.5%)	-	-	26 (22.2%)
	To the Order	-	1 (5.9%)	-	1 (0.9%)

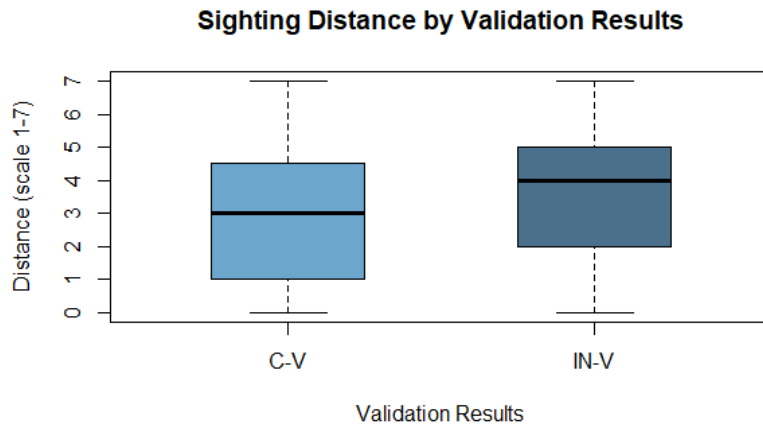
Out of all of the matched records, 170 (~59.2%) validations reached the species level (Table 11). As a result of the validation process, it was possible to reach a lower taxon in 49 sightings (~17.1%), out of which, 31 (~10.8%) up to the species level. Table 13 displays the updated number of cetacean occurrences, which includes 3 sightings of a new species: *Mesoplodon europaeus*. Besides that, 2 of the Mysticeti occurrences were registered as “*Balaenoptera borealis* or *Balaenoptera edeni*”.

**Table 13.** Updated number of cetacean occurrences of the CETUS dataset after the video/photographic verification and validation processes. Within parentheses is the difference in values when compared to table 7. The number of occurrences is presented by taxa recorded to the highest possible level. The table is organized by taxon rank of the records and alphabetically within.

Taxa	Taxon Rank	Number of Occurrences	Taxa	Taxon Rank	Number of Occurrences
Cetacea	Infraorder	428 (-2)	<i>Megaptera novaeangliae</i>	Species	9
Mysticeti	Superfamily	358 (-1)	<i>Mesoplodon densirostris</i>	Species	8
Delphinidae	Family	1000 (-14)	<i>Mesoplodon europaeus</i>	Species	3 (+3)
Ziphiidae	Family	178 (-4)	<i>Orcinus orca</i>	Species	8
<i>Globicephala</i>	Genus	68 (+1)	<i>Peponocephala electra</i>	Species	4
<i>Kogia</i>	Genus	8 (+1)	<i>Phocoena phocoena</i>	Species	9
<i>Balaenoptera acutorostrata</i>	Species	93	<i>Physeter macrocephalus</i>	Species	178
<i>Balaenoptera borealis</i>	Species	4	<i>Pseudorca crassidens</i>	Species	13
<i>Balaenoptera edeni</i>	Species	6	<i>Stenella attenuata</i>	Species	8 (-1)

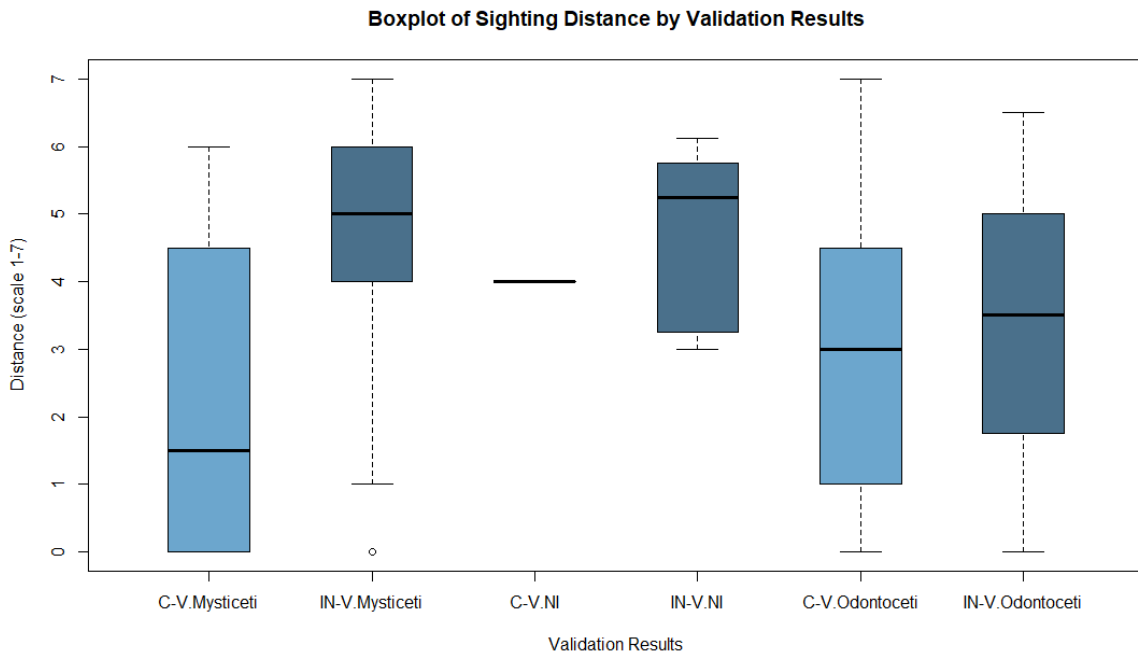
Taxa	Taxon Rank	Number of Occurrences	Taxa	Taxon Rank	Number of Occurrences
<i>Balaenoptera musculus</i>	Species	3	<i>Stenella clymene</i>	Species	17
<i>Balaenoptera physalus</i>	Species	37 (+1)	<i>Stenella coeruleoalba</i>	Species	183 (+2)
<i>Delphinus delphis</i>	Species	477 (+4)	<i>Stenella frontalis</i>	Species	336 (+8)
<i>Grampus griseus</i>	Species	13	<i>Stenella longirostris</i>	Species	6
<i>Hyperoodon ampullatus</i>	Species	5	<i>Steno bredanensis</i>	Species	4
<i>Lagenodelphis hosei</i>	Species	1	<i>Tursiops truncatus</i>	Species	171
<i>Lagenorhynchus albirostris</i>	Species	3	<i>Ziphius cavirostris</i>	Species	81 (+2)
			Total	32 taxa	3720 occurrences

Concerning the validation results, incomplete and non-validated sightings (“IN-V”) had a significantly higher sighting distance median ( $w = 8501$ ;  $p\text{-value} = 0.03585$ ) when compared with those who reached correctly the species level (“C-V”; Figure 7).



**Figure 7.** Boxplot of the sighting distance of the matched sightings by validation results. C-V – complete validation; IN-V – incomplete validation or non-validated sightings.

C-V and IN-V sightings within each group had no significant differences (Table 17 in Appendix), although all of the IN-V sightings had a higher sighting distance median. Mysticeti and Odontoceti when compared did not have a significant difference for the C-V sightings (Table 18 in Appendix), however this difference was significant for the IN-V sightings ( $w = 461$ ;  $p\text{ value} = 0.01242$ ).

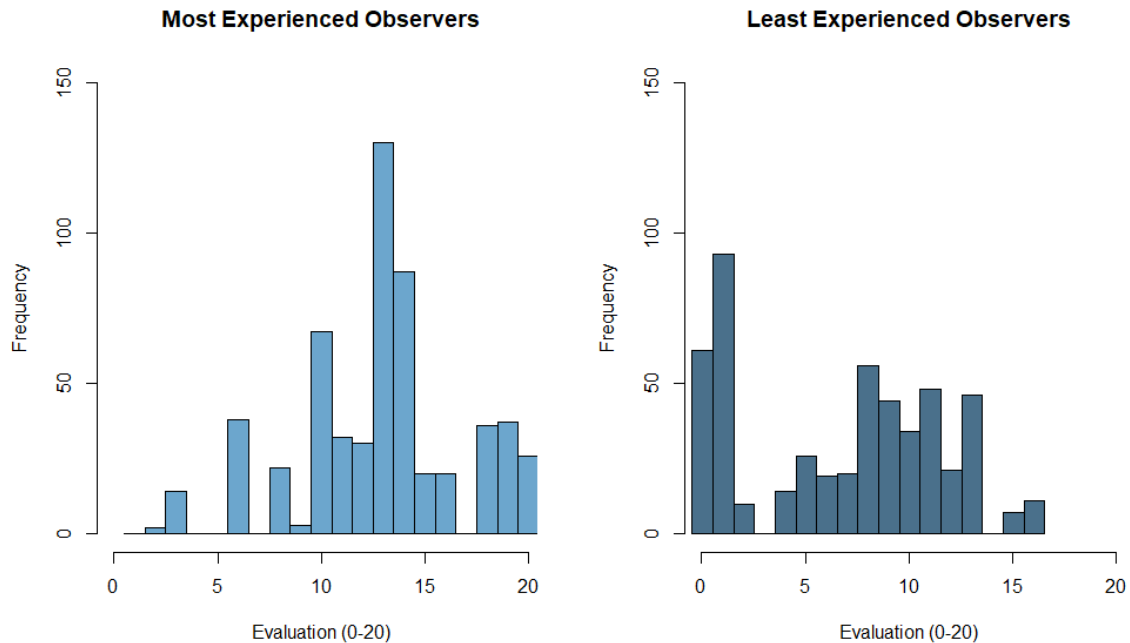


**Figure 8.** Boxplot of the validation results of the matched sightings by suborder. C-V – complete validation; IN-V – incomplete validation or non-validated sightings. NI – non-identified.

### 3.2 Creating a Data Quality Criteria: MMOs Experience

There were 61 MMOs that participated in the data collection, but 80 *curricula vitae* were evaluated given the alumni MMOs.

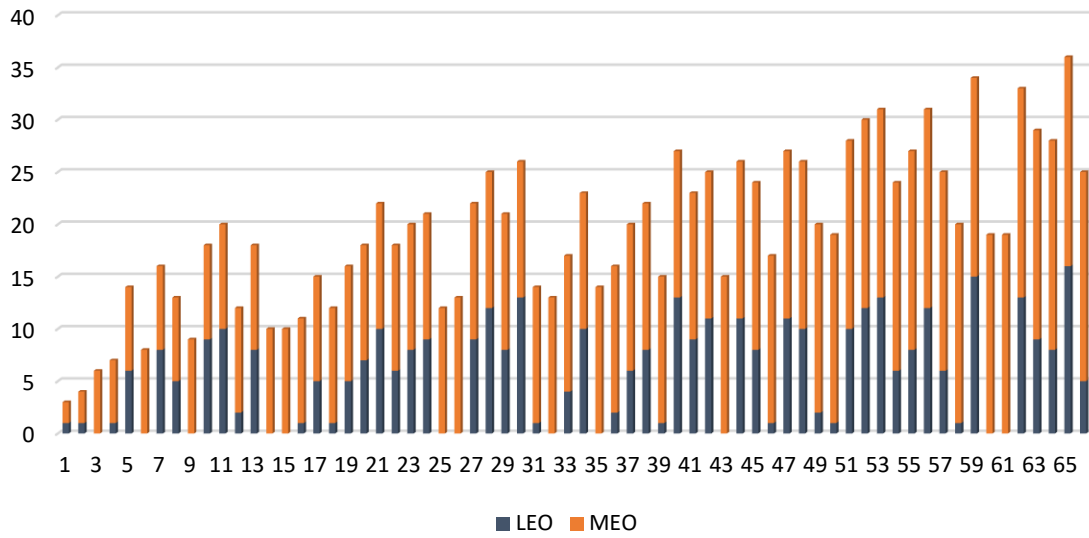
Regarding the MMOs' teams, in a total of 564 cruises, only ~14% had their most experienced observer with a score of less than 10. The minimum score recorded for the most experienced observer was 2 (Figure 9).



**Figure 9.** Histograms representing the frequency of cruises across the range of the Marine Mammal Observers' experience, based on the MMOs evaluation performed, for the Most Experienced Observer (MEO) and the Least Experienced Observer (LEO).

The lack of experience of the least experienced MMO was generally compensated by a more experienced MMO, with generally balanced MMOs' teams. Out of 66 MMO combined evaluations, only 4 scored less than 10 of accumulated experience (between the two MMOs). On over half of the surveys, accumulated experience was higher than 20, with only ~8% of the surveys having less than 10 of accumulated experience (Figure 10).

### Experience of the MMO teams – combination of evaluations



**Figure 10.** Bar plot representing the evaluation score for the Most Experienced Observer (MEO) and for the Least Experienced Observer (LEO) for each combination of scores of the Marine Mammal Observers' teams. For the surveys where there was only one Marine Mammal Observer, only the orange bar (MEO) is displayed.

### 3.3 Bias Modelling of Number of Sightings

As a result of the Pearson correlations between all pairs of explanatory variables, the mean and cumulative scores of the LEOs and MEOs were excluded as they were highly correlated with the individual scores of the LEOs e MEOs (Pearson correlation >0.75; Figure 12 in Appendix). After excluding those two variables, all VIF values were lower than the threshold considered (i.e., lower than 3), so no additional variables were removed (Table 19 in Appendix).

By means of backward selection, the best final GAM model ended with the following explanatory variables: kilometres sampled “on-effort”, evaluation score of the MEOs, minimums of the sea state, and minimums and maximums of the wind state and visibility (Table 14). Abbreviations of the explanatory variables presented in Table 14. The minimums of the wind and the maximums of visibility variables had the smooth terms removed, as splines were close to linear.

**Table 14.** Results from the best final Generalized Additive Model (GAM) developed for assessing the bias on the number of sightings collected per survey. Sight – number of sightings per survey. MEO – Most Experienced Observer. Min\_Sea – minimums of the sea state in each survey. Min\_Wind – minimums of the wind state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey. Max\_Vis – maximums of the visibility in each survey.

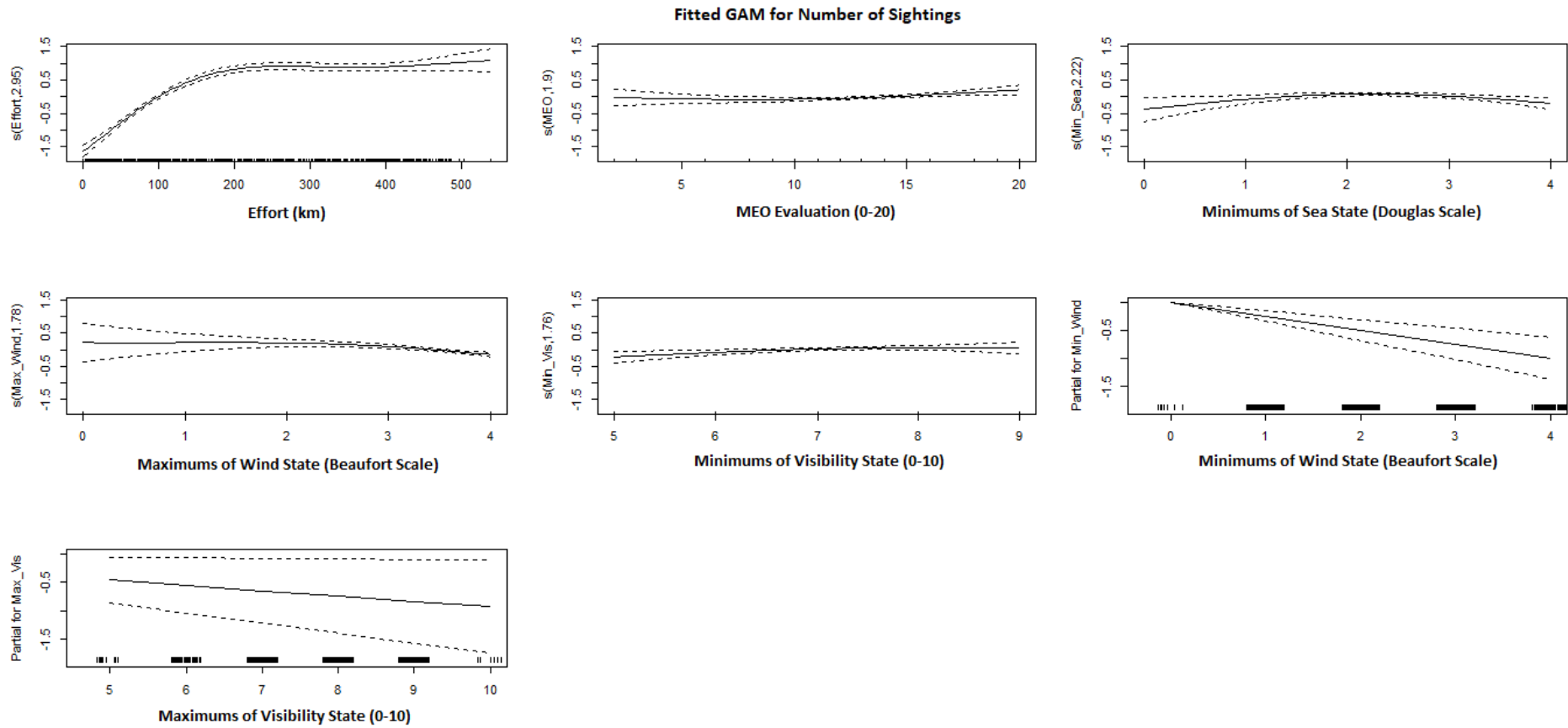
GAM				
Variables	Degrees of Freedom	Reference Degrees of Freedom	Chi-Square	P-value
s(Effort)	2.954	2.998	493.145	<2x10 <sup>-16</sup>
s(MEO)	1.905	2.275	8.544	0.016300
s(Min_Sea)	2.215	2.600	10.617	0.007561
s(Max_Wind)	1.785	2.111	18.352	0.000134
s(Min_Vis)	1.759	2.157	6.385	0.051609
Variables	Estimate	Standard Error	Z-value	Pr(> z )
Min_Win	-0.24648	0.04652	-5.298	1.17x10 <sup>-07</sup>
Max_Vis	-0.09264	0.04081	-2.270	0.0232

Number of sightings per survey increased with the sampling efforts, stabilizing after the 200 kilometres sampled. In other words, the more is the number of kilometres sampled “on-effort” per survey, the more sightings are expected to occur, stabilizing from the 200



kilometres sampled. Number of sightings peaked when the minimum sea state of the survey was around 2 and 3 (in the Douglas scale). Overall, number of sightings decreases with the increase of the wind state (both minimum and maximum) registered for the survey (in the Beaufort scale), and increases with better visibility conditions (considering the minimum value registered) and with higher scores for the MEO. Although the maximum visibility variable remained in the best final model, the confidence intervals were too wide, and no obvious tendency can be verified (Figure 11).

With “gam.check” function, and multiple plots of interest, we could state that: there are no highly influential data points (“hat” values lower than 0.25; Figure 13 in Appendix; Correia et al., 2020; Zuur et al., 2007); an adequate number of knots was defined [adequacy between number of knots and estimated degrees of freedom as retrieved by “gam.check” – although, in all cases but one, p-value is very low (significant), the  $k'$  is not too close to the estimated degrees of freedom, so it is worth maintaining a limit to the number of knots as to have more meaningful smooth functions; Table 20 in Appendix]; model residuals do not follow a normal distribution (“gam.check” histogram of residuals; Figure 13 in Appendix) and they are not correlated to any of the explanatory variables (plot between model residuals and explanatory variables; Figures 14-21 in Appendix).



**Figure 11.** Plots of the final Generalized Additive Model (GAM) developed for assessing the influence of bias of kilometres samples “on-effort”, meteorological conditions and the experience of observers. MEO – Most Experienced Observers.

## 4. Discussion

### 4.1 Photographic Verification/Validation

#### *Gathering of Photographic and Video Material*

Since capturing sightings on video or photography is not part of the CETUS Project monitoring protocol, all the records used in this work were opportunistic, and most of the records were not from the CETUS team, but obtained by alumni volunteer observers. Therefore, all 61 volunteers had to be contacted. As some time went by since their internship at CETUS, it is possible that those 7 volunteers that did not respond to our request, do not have the same email anymore. In fact, they all volunteered within the first 5 years of the CETUS Project.

Although photographs and videos were gathered from 26 MMOs (~42.6%), these corresponded to only 307 sightings - ~8.3% of the CETUS dataset. For future improvement, it is essential to add the collection of cetacean visual records to the monitoring protocol. This will automatically increase the percentage of verified and validated sightings, and therefore ensure data quality (Evans & Hammond, 2003; Kosmala et al., 2016).

#### *Photographic Verification*

The wrong information associated to the files provided by MMOs was the biggest obstacle to the verification process. More specifically, the date and/or time of the camera or mobile phone displayed in the metadata of the file. The majority was either wrong, non-existent or with different time zones. All things considered, ~90.9% of the sightings recorded were able to be matched with the CETUS dataset records as a result of the conservative methodology adopted.

Photographs or videos that could not be matched were due to the lack of photographs/videos of that MMO for that route (i.e., which made it impossible to determine the time difference between the metadata timestamp of the files and the sighting registers) or to the lack of information (e.g., photographs/videos without time, meaningless hours, or too much variation). In this regard, the majority of the videos could not be used since almost none of them had the information regarding the date and time. To prevent this from interfering with the verification process in the future, it is necessary that, before each survey, the date and time of the cameras or mobile phones of all observers are checked and configured to UTC. This way, it will make the matching process a lot simpler and less susceptible to errors.

The best matching results correspond to the suborder Odontoceti, which represents ~9.15% of the total Odontoceti sightings, while only ~3.89% of mysticetes were matched. In other words, MMOs tend to photograph more this suborder. This can be explained by the fact that the majority of Odontoceti sightings are delphinids. Delphinids are avid riders of the bow wakes of ships, and some species even regularly perform aerial stunts, such as high leaps and flips, which makes them a lot easier to photograph (LeDuc, 2009; Gowans, et al. 2007).

Although there were no significant differences between the matched and non-matched or incomplete sightings distance, the latter were never lower. Moreover, both Odontoceti matched and non-matched sightings distances were significantly lower when compared to the Mysticeti results. This may be due to the fact that it is easier for MMOs to photograph or record videos if the animals are closer to the ship, which is common among the delphinids. This is even more pronounced if MMOs do not have a professional camera. Besides that, sightings at a closer distance are usually longer in time, allowing for better opportunities to photograph (especially, when the priority at sea is to register the coordinates of the sighting and only then it is possible to photograph the animals). A professional camera per ship, along with a long-range zoom lens, could increase the number of photographic/video records of sightings at a longer distance (which in turn could support the ID of sightings that are usually harder to identify given the distance to the observer). Also, the availability of such equipment could serve as an incentive for the observers to photograph the animals.

#### *Photographic Validation*

Suborder Odontoceti had the highest percentage of validated sightings with correct ID (“yes”) when compared to Mysticeti (Table 11). This can also be explained by delphinids behaviour, which, as previous mentioned, makes them a lot easier to photograph. Being easier to be photographed, it is also easier to get better quality photographs that in turn enable better validation results.

Out of all the 10 wrong identifications, 9 were between delphinid species, with 4 of them in the same genus: 2 sightings of *Stenella frontalis* mistaken for *Stenella attenuata*, 1 sighting of *Stenella frontalis* mistaken for *Stenella coeruleoalba*, and 1 sighting of *Stenella coeruleoalba* mistaken for *Stenella frontalis*. Actually, these 3 species have plenty of visual similarities. The only major visual difference between *Stenella attenuata* and *Stenella frontalis* is the dark ventral spots: in *Stenella attenuata* these tend to merge and fade as the animals get older resulting in a slightly mottled or uniform grey belly; whereas in *Stenella frontalis* they remain clearly defined with the original white underlying surface visible between them (Perrin et al., 1994; Jefferson et al., 2015). *Stenella coeruleoalba* can be

easily distinguished by their dark grey lateral stripe that starts in the eye and ends in the anus, however both *Stenella coeruleoalba* and *Stenella frontalis* have a pale grey blaze that sweeps up towards the dorsal fin (Perrin et al., 1994; Jefferson et al., 2015). To prevent misidentifications theoretical must be enhanced. Presenting the wrong IDs to future volunteers can help to prevent it from happening again.

Furthermore, there were also 3 sightings of *Stenella frontalis* mistaken for *Tursiops truncatus*, 1 sighting of *Stenella coeruleoalba* mistaken for *Tursiops truncatus*, and 1 sighting of *Tursiops truncatus* mistaken for *Delphinus delphis*. Within mysticetes, only a *Balaenoptera borealis* was mistaken for a *Physeter macrocephalus*.

The validation process resulted in about 3.5% of misidentifications (out of the matched records), which if we extrapolate the result to the entire CETUS dataset can sum up to ~130 sightings with the wrong ID. However, most of these misidentifications are within the small delphinids. This indicates that for certain analyses, it may be wiser to merge sightings and perform the analyses for the group and not at the species level. Alternatively, it may also be useful to select species less prone to be misidentified with enough sightings to guarantee a lower error rate. This is often done by the CETUS team (Correia et al., 2015; Correia et al., 2019b).

As a result of the validation process, it was possible to reach a lower taxon in 49 sightings, out of which 31 up to the species level. The validation process also allowed for the inclusion of a new species in in the CETUS dataset: *Mesoplodon europaeus*.

While the verification process had a high success rate (with ~90.9% of the recorded sightings matched), the same did not apply to the validation process (only ~59.2% C-V sightings), which means that photographic quality hindered the identification up to the species level. Also, distance of the sighting was significantly higher in records with incomplete validation (i.e., ID did not reach the species level) or non-validated. This corroborates the need to improve the visual recording of the sighting, not only by increasing the number of photographic/video records (i.e., to increase the number of sightings in the dataset with photographic/video records) but also the quality of the records (i.e., to increase the number of validated sightings). This can be done, as said before, by providing a professional camera plus long-range zoom lens in each ship. However, this procedure can translate into high costs, potentially impeditive for most long-term monitoring programmes, as these cameras are expensive and have a reduced life span when being manipulated by several users in a moving ship at sea. Nonetheless, with the improvement of mobile cameras, cell phones and/or tablets may be sufficient and can become fundamental equipment in this task, yielding excellent results if used properly. As such, it would also be

useful to include, in the monitoring programme, a training on photography basis and how to photograph the animals (i.e., body parts that increase possibilities of identification). This would automatically improve image quality and it could also serve as an incentive for the volunteers to photograph, which in turn would increase the number of sightings photographed and therefore the number of possible validated sightings.

Although C-V sightings had no significant differences between mysticetes and odontocetes, this difference was significant for IN-V sightings. This also corroborates what has been discussed, highlighting the need to improve the visual recording of the sightings.

The verification and validation proved to be fundamental processes, not only in the CETUS Project, but also in identical monitoring programs. Making the information on whether the sightings were validated or not through photographic/video records is a valuable information to open-access datasets on occurrences. It allows the user to decide on whether to use the entire dataset or just the validated sightings (which may also depend on the data analysis intended).

## 4.2 Creating a Data Quality Criteria: MMOs Experience

Regarding the data quality criteria, the evaluation of the MMOs experience was used to inform the dataset user on possible data bias. MMOs experience was evaluated based on their CVs at the time of the internship. In the future, it could be useful to evaluate MMOs based on their performance in the theoretical trainings, and if possible, with a practical training at sea. This would possibly provide a better indication of the quality of the data collected.

With the evaluation process it was possible to verify that in a total of 564 cruises, only ~14% had its most experienced observer with a score of less than 10, and only ~8% of the surveys had less than 10 of accumulated experience. This indicates that, overall, the cruises had balanced teams of observers in order to prevent the cetacean detectability and identification from being influenced.

In addition, just like the verification/validation processes, this type of information also gives the possibility of giving future users the choice of excluding, for example, data from surveys that had a cumulative experience of less than 10 or data that had the MEO with an evaluation score below 10.

### 4.3 Bias Modelling of Number of Sightings

As expected, the number of species sighted per survey increased with kilometres sampled “on-effort”, and stabilized at a high number of kilometres sampled per survey (~200 km). Therefore, a homogeneous effort coverage across surveys or sampling over 200 km per survey is recommended. Another solution is to standardize the data with the sampling effort, as done in previous studies with this dataset (e.g., Correia et al., 2020; Correia et al., 2021).

The number of sightings increases with wave height, peaking around 2 (Douglas scale), and decreasing thereafter. Since the majority of sightings in CETUS dataset belongs to the family Delphinidae (~62.2%), this can be due to the fact that delphinids tend to wave-ride and thus may be visible at the surface for longer if the waves are higher and wider. This behaviour is believed to provide additional benefits in terms of speed and in terms of saving energy (Williams et al., 1992). This can also explain the increase of number of sightings with the increase in wind speed, for maximum values (from 0 to 1, in Beaufort scale), as when wind blows over water, it leads to surface waves (Ardhuin and Orfila, 2018). However, confidence intervals of the smoothing function for maximum wind speed are very wide in the limits (from 0 to 1, Beaufort scale), preventing from any robust conclusions.

GAM results show that weather conditions influence detectability, with an overall decreased number of sightings per survey with poor weather conditions (strong wind speed and poor visibility). Therefore, these factors must be considered when analysing occurrence and abundance data.

GAM results also indicate that the number of sightings increases with the experience of the MEO, especially at high values of the MEO experience (i.e., higher than 10). The experience of the LEO was dropped from the final model. This may emphasize the need to have, at least, a very experienced observer on-board. To combine a less experienced observer with a very experienced one may be a good strategy to provide a quality training to the LEO, without losing data quality.

Overall, the variables included in the model explained a great amount of deviance in the number of sightings (~49.1%). This shows how relevant it is to apply bias modelling in data from visual records. This specific model provides additional information to future dataset users, enabling them to use CETUS data more accurately, depending on the needs of the analysis intended: i.e., if a very conservative approach is deemed necessary, a user may choose to use only data from surveys with over 200 km, sea state of 2, wind state below 3, visibility higher than 7, MEO with a score higher than 10.



## 5. Conclusion

This thesis aimed to optimize and allow the proper use of a long-term monitoring cetacean dataset. The main conclusions are presented below:

- Long-term monitoring datasets can provide a baseline to better understand physical and ecological responses to ocean environmental changes, playing an important role in marine management and conservation.

- Data must become public following FAIR Data Principles to enhance their reusability. It is also fundamental that the data is reliable, and the sources of bias are identified and quantified.

- From the collected photographic records, ~90.9% were able to be matched with the dataset occurrences, although corresponding only to ~7.5% of the total dataset. This emphasizes the need to include the collection of photographic records in the protocol and encourage observers to do so. To optimize the matching process, it is necessary that, before each survey, the date and time of the cameras or mobile phones of all observers are checked and configured to UTC. This will make the matching process a lot simpler and less susceptible to errors.

- While the verification process had a high successful rate (with ~90.9% of the recorded sightings matched), the same did not apply to the validation process (only ~59.2% C-V sightings), which means that photographic quality hindered the identification up to the species level. This reveals the need to improve image quality of photographic / video records.

- On ~17.1% of the matched records, we were able to reach a lower taxon with the validation process, ~10.8% up to the species. Besides that, the validation process allowed ~3.5% wrong identifications being corrected and the inclusion of a new species in in the CETUS dataset: *Mesoplodon europaeus*. This reveals the importance of verification/validation methods and the need to increase photographic registers during sampling.

- The suborder Odontoceti had the best verification/validation results. This can be explained by the fact that it is easier for MMOs to photograph or record videos if the animals are closer to the ship, which is common among the delphinids. Being easier to be photographed, it is also easier to get better quality photographs, that in turn enable better validation results.

- The median distance sighting was significantly higher in records with incomplete validation or non-validated. This corroborates the need to improve the visual recording of the sighting,

not only by increasing the number of photographic/video records but also the quality of the records. A professional camera per ship, along with a long-range zoom lens, could help to achieve this.

- It would be useful to include in the monitoring programme a training on photography basics and how to photograph the animals. This would automatically improve image quality and it could also serve as an incentive for the volunteers to photograph, which in turn would increase the number of sightings photographed and therefore the number of possible validated sightings.

- Regarding the experience of observers, results shown that, overall, the cruises had balanced teams of observers in order to prevent the cetacean detectability and identification from being influenced.

- In the future, it could be useful to evaluate MMOs based on their performance in the theoretical trainings, and if possible, with a practical training at sea. This would provide better evaluations of the MMOs, allowing for a more informed decision on MMO teams and retrieving better indicators for the data quality criteria.

- Model results show that weather conditions influence detectability, with an overall decreased number of sightings per survey with poor weather conditions (strong wind speed and poor visibility). Therefore, these factors must be considered when analysing occurrence and abundance data.

- Model results indicate that the number of sightings increases with the experience of the MEO, especially at high values of the MEO experience (i.e., higher than 10). On the other hand, the experience of the LEO was dropped from the final model, which may emphasize the need to have, at least, a very experienced observer on-board. To combine a less experienced observer with a very experienced one may be a good strategy to provide a quality training to the LEO, without losing data quality.

- Overall, the variables included in the model explained a great amount of deviance in the number of sightings (~49.1%). This shows how relevant it is to apply bias modelling in data from visual records.

- Ultimately, this work provides additional information to future dataset users, enabling them to use CETUS data more accurately, depending on the needs of the analysis intended, and contributes to an improvement of monitoring protocols of CETUS and similar programs.

## 6. Bibliographic References

- Andrews, R. D., Baird, R. W., Calambokidis, J., Goertz, C. E. C., Gulland, F. M. D., Heide-Jorgensen, M. P., Hooker, S. K., Johnson, M., Mate, B., Mitani, Y., Nowacek, D. P., Owen, K., Quakenbush, L. T., Raverty, S., Robbins, J., Schorr, G. S., Shpak, O. V., Townsend Jr, F. I., Uhart, M., ... & Zerbini, A. N. (2019). Best practice guidelines for cetacean tagging. *Journal of Cetacean Research and Management*, 20, 27-66. <https://doi.org/10.47536/jcrm.v20i1.237>
- Apprill, A., Miller, C. A., Moore, M. J., Durban, J. W., Fearnbach, H., & Barrett-Lennard, L. G. (2017). Extensive Core Microbiome in Drone-Captured Whale Blow Supports a Framework for Health Monitoring. *mSystems*, 2(5), e00119-17. <https://doi.org/10.1128/msystems.00119-17>
- Ardhuin, F. & Orfila, A. (2018). Wind Waves. Florida State University Libraries. Retrieved from <https://diginole.lib.fsu.edu/islandora/object/fsu:602129/datastream/PDF/view>
- Arregui, M., Josa, M., Aguilar, A., & Borrell, A. (2017). Isotopic homogeneity throughout the skin in small cetaceans. *Rapid Communications in Mass Spectrometry*, 31(18), 1551-1557. <https://doi.org/10.1002/rcm.7936>
- ASCOBANS. (2017). Intersessional Working Group on Research and Conservation Actions Undertaken in the Extended Agreement Area: Update for the Period September 2014 to August 2015. Retrieved from [https://www.ascobans.org/sites/default/files/document/AC22\\_5.5.a\\_ExtensionArea\\_WGReport.pdf](https://www.ascobans.org/sites/default/files/document/AC22_5.5.a_ExtensionArea_WGReport.pdf)
- Bailey, L. L., Simons, T. R., & Pollock, K. H. (2004). Estimating Site Occupancy and Species Detection Probability Parameters for Terrestrial Salamanders. *Ecological Applications*, 14(3), 692-702. <https://doi.org/10.1890/03-5012>
- Bas, Y., Devictor, V., Moussus, J.-P., & Jiguet, F. (2008). Accounting for weather and time-of-day parameters when analysing count data from monitoring programs. *Biodiversity and Conservation*, 17, 3403-3416. <https://doi.org/10.1007/s10531-008-9420-6>
- Bento, M., Canha, R., Eira, C., Vingada, J., Nicolau, L., Ferreira, M., Domingo, M., Tavares, L., & Duarte, A. (2019). Herpesvirus infection in marine mammals: A retrospective molecular survey of stranded cetaceans in the Portuguese coastline. *Infection, Genetics and Evolution*, 67, 222-233. <https://doi.org/10.1016/j.meegid.2018.11.013>

- Bose, N., & Lien, J. (1990). Energy Absorption from Ocean Waves: A Free Ride for Cetaceans. *Proceedings of the Royal Society B: Biological Sciences*, 240(1299), 591–605. <https://doi:10.1098/rspb.1990.0054>
- Boyd, C., Hobbs, R. C., Punt, A. E., Shelden, K. E. W., Sims, C. L., & Wade, P. R. (2019). Bayesian estimation of group sizes for a coastal cetacean using aerial survey data. *Marine Mammal Science*, 35(4), 1322-1346. <https://doi.org/10.1111/mms.12592>
- Brownell Jr., R. L., Reeves, R. R., Read, A. J., Smith, B. D., Thomas, P. O., Ralls, K., Amano, M., Berggren, P., Chit, A. M., Collins, T., Currey, R., Dolar, M. L. L., Genov, T., Hobbs, R. C., Krebs, D., Marsh, H., Zhigang, M., Perrin, W. F., Phay, S., ... & Wang, J. Y. (2019). Bycatch in gillnet fisheries threatens Critically Endangered small cetaceans and other aquatic megafauna. *Endangered Species Research*, 40, 285-296. <https://doi.org/10.3354/esr00994>
- Caldeira, R. M. A., & Sangrà, P. (2012). Complex geophysical wake flows. *Ocean Dynamics*, 62, 683-700. <https://doi.org/10.1007/s10236-012-0528-6>
- Chami, R., Fullenkamp, C., Berzaghi, F., Español-Jiménez, S., Marcondes, M., & Palazzo, J. (2020). On Valuing Nature-Based Solutions to Climate Change: A Framework with Application to Elephants and Whales. *Economic Research Initiatives at Duke*, 297. <http://dx.doi.org/10.2139/ssrn.3686168>
- Cisneros-Montemayor, A. M., Sumaila, U. R., Kaschner, K., & Pauly, D. (2010). The global potential for whale watching. *Marine Policy*, 34(6), 1273-1278. <https://doi.org/10.1016/j.marpol.2010.05.005>
- Clarke, R. H., Gales, R., & Schulz, M. (2017). Land-based observations of cetaceans and a review of recent strandings at subantarctic Macquarie Island. *Australian mammalogy*, 39(2), 248-253. <https://doi.org/10.1071/AM16007>
- Cominelli, S., Moulins, A., Rossi, V., Arcangeli, A., David, L., Di-Meglio, N., & Tepsich, P. (2013). Assessing the Consistency of Data Collected Using Ferries as Platforms of Opportunity for Cetacean Monitoring Programs. Poster presented at: 27th Conference of the European Cetacean Society; Setúbal, Portugal.
- Cominelli, S., Moulins, A., Rossi, V., Rosso, M., & Tepsich, P. (2014). A new process for developing an effective index to assess variability in cetacean presence. Poster presentation. Poster presented at: 28th Conference of the European Cetacean Society; Liège, Belgium.

- Cominelli, S., Moulins, A., Rosso, M., & Tepsich, P. (2015). Fin whale seasonal trends in the Pelagos Sanctuary, Mediterranean Sea. *The Journal of Wildlife Management*, 80(3), 490-499. <https://doi.org/10.1002/jwmg.1027>
- Convention on Biological Diversity. (2000). How the Convention on Biological Diversity promotes nature and human well-being. Sustaining Life on Earth. Retrieved 27 December 2020 from <https://www.cbd.int/convention/guide/?id=web>
- Correia, A. M. (2013). Cetacean monitoring in Northeastern Atlantic Ocean: Occurrence and distribution of cetacean species in the Canary Basin. (Master's thesis). Retrieved from <https://hdl.handle.net/10216/68792>
- Correia, A. M., Tepsich, P., Rosso, M., Caldeira, R., & Sousa-Pinto, I. (2015). Cetacean occurrence and spatial distribution: Habitat modelling for offshore waters in the Portuguese EEZ (NE Atlantic). *Journal of Marine Systems*, 143, 73-85. <https://doi.org/10.1016/j.imarsys.2014.10.016>
- Correia, A. M., Gandra, M., Liberal, M., Valente, R., Gil, Á., Rosso, M., Pierce, G. J., & Sousa-Pinto, I. (2019a). A dataset of cetacean occurrences in the Eastern North Atlantic. *Scientific Data*, 6(177). <https://doi.org/10.1038/s41597-019-0187-2>
- Correia, A. M., Gil, Á., Valente, R., Rosso, M., Pierce, G. J., & Sousa-Pinto, I. (2019b). Distribution and habitat modelling of common dolphins (*Delphinus delphis*) in the eastern North Atlantic. *Journal of the Marine Biological Association of the United Kingdom*, 99(06), 1443-1457. <https://doi.org/10.1017/s0025315419000249>
- Correia, A. M., Gandra, M., Liberal, M., Valente, R., Gil, A., Rosso, M., Pierce, G.J. & Sousa-Pinto, I, CIIMAR - UP. (2019c). CETUS: Cetacean monitoring surveys in the Eastern North Atlantic. *Marine Data Archive*. <https://doi.org/10.14284/350>
- Correia, A. M. (2020). Distribution and habitat modelling for cetacean species in the eastern north Atlantic Ocean. (Doctoral dissertation). Retrieved from <https://hdl.handle.net/10216/125661>
- Correia, A. M., Gil, Á., Valente, R., Rosso, M., Sousa-Pinto, I., & Pierce, G. J. (2020). Distribution of cetacean species at a large scale - Connecting continents with the Macaronesian archipelagos in the eastern North Atlantic. *Diversity and Distributions*, 26, 1234-1247. <https://doi.org/10.1111/ddi.13127>
- Correia, A. M., Sousa-Guedes, D., Gil, Á., Valente, R., Rosso, M., Sousa-Pinto, I., Sillero, N. & Pierce, G.J. (2021) Predicting Cetacean Distributions in the Eastern North

- Atlantic to Support Marine Management. *Frontiers in Marine Science*, 8(643569). <https://doi.org/10.3389/fmars.2021.643569>
- Davis, G. E., Baumgartner, M. F., Bonnell, J. M., Bell, J., Berchok, C., Bort Thornton, J. B., Brault, S., Buchanan, G., Charif, R. A., Cholewiak, D., Clark, C. W., Corkeron, P., Delarue, J., Dudzinski, K., Hatch, L., Hildebrand, J., Hodge, L., Klinck, H., Kraus, S., ... & Van Parijs, S. M. (2017). Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014. *Scientific Reports*, 7(13460). <https://doi.org/10.1038/s41598-017-13359-3>
- Deyoung, B., Visbeck, M., De Araujo Filho, M. C., Baringer, M. O. N., Black, C., Buch, E., Canonico, G., Coelho, P., Duha, J. T., Edwards, M., Fischer, A., Fritz, J.-S., Ketelhake, S., Muelbert, J.-H., Monteiro, P., Nolan, G., O'Rourke, E., Ott, M., Le Traon, P. Y., ... & Willis, Z. (2019). An Integrated All-Atlantic Ocean Observing System in 2030. *Frontiers in Marine Science*, 6(428). <https://doi.org/10.3389/fmars.2019.00428>
- Duarte, C. M., Agusti, S., Barbier, E., Britten, G. L., Castilla, J. C., Gattuso, J.-P., Fulweiler, R. W., Hughes, T. P., Knowlton, N., Lovelock, C. E., Lotze, H. K., Predragovic, M., Poloczanska, E., Roberts, C., & Worm, B. (2020). Rebuilding marine life. *Nature*, 580, 39-51. <https://doi.org/10.1038/s41586-020-2146-7>
- Duignan, P. J., Stephens, N. S., & Robb, K. (2020). Fresh water skin disease in dolphins: a case definition based on pathology and environmental factors in Australia. *Scientific Reports*, 10, 21979. <https://doi.org/10.1038/s41598-020-78858-2>
- Durante, C. A., Reis, B. M. M., Azevedo, A., Crespo, E. A., & Lailson-Brito, J. (2020). Trace elements in trophic webs from South Atlantic: The use of cetaceans as sentinels. *Marine Pollution Bulletin*, 150(110674). <https://doi.org/10.1016/j.marpolbul.2019.110674>
- Estes, J. A., Tinker, M. T., Williams, T.M., & Doak, D. F. (1998). Killer Whale Predation on Sea Otters Linking Oceanic and Nearshore Ecosystems. *Science*, 282(5388), 473-476. <https://doi.org/10.1126/science.282.5388.473>
- European Commission. (2012). Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions. Blue Growth opportunities for marine and maritime sustainable growth. Retrieved 10 January from <https://eur-lex.europa.eu/legalcontent/EN/TXT/HTML/?uri=CELEX:52012DC0494&from=EN>

- Evans, P. G. H., & Hammond, P. S. (2003). Monitoring cetaceans in European waters. *Mammal Review*, 34(1-2), 131-156. <https://doi.org/10.1046/j.0305-1838.2003.00027.x>
- Gil, A. (2018). Cetáceos na Zona Económica Exclusiva Continental Portuguesa: distribuição espaço-temporal e registo de novas ocorrências. (Master's thesis). Retrieved from <https://hdl.handle.net/10216/118804>
- Gordon, J. (2001). Measuring the range to animals at sea from boats using photographic and video images. *Journal of Applied Ecology*, 38, 879-887. <https://doi.org/10.1046/j.1365-2664.2001.00615.x>
- Gowans, S., Würsig, B., & Karczmarski, L. (2007). The Social Structure and Strategies of Delphinids: Predictions Based on an Ecological Framework. *Advances in Marine Biology*, 53, 195-294. [https://doi.org/10.1016/S0065-2881\(07\)53003-8](https://doi.org/10.1016/S0065-2881(07)53003-8)
- Granados, J. E., Ros-Candeira, A., Pérez-Luque, A. J., Moreno-Llorca, R., Cano-Manuel, F. J., Fandos, P., Soriguer, R. C., Cerrato, J. E., Jiménez, J. M. P., Ramos, B., & Zamora, R. (2020). Long-term monitoring of the Iberian ibex population in the Sierra Nevada of the southeast Iberian Peninsula. *Scientific Data*, 7(203). <https://doi.org/10.1038/s41597-020-0544-1>
- Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2), 267-306. <https://doi.org/10.1111/j.1467-985x.2004.00349.x>
- Hammond, P. S. (1990). Capturing whales on film—estimating cetacean population parameters from individual recognition data. *Mammal Review*, 20(1), 17-22. <https://doi.org/10.1111/j.1365-2907.1990.tb00099.x>
- ICES. (2016). Report of the Working Group on Marine Mammal Ecology (WGMME). Retrieved from [http://www.ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/acom/2016/WGMME/wgmme\\_2016.pdf](http://www.ices.dk/sites/pub/Publication%20Reports/Expert%20Group%20Report/acom/2016/WGMME/wgmme_2016.pdf)
- Jefferson, T. A., Webber, M. A. & Pitman, R. L. (2015). 4 - Cetaceans. *Marine Mammals of the World (Second Edition)*, 24-357. <https://doi.org/10.1016/B978-0-12-409542-7.50004-4>
- Kaschner, K., Quick, N. J., Jewell, R., Williams, R., & Harris, C. M. (2012). Global Coverage of Cetacean Line-Transsect Surveys: Status Quo, Data Gaps and Future Challenges. *PLoS ONE*, 7(9), e44075. <https://doi.org/10.1371/journal.pone.0044075>

- Kiszka, J., Macleod, K., Canneyt, O. V., Walker, D., & Ridoux, V. (2007). Distribution, encounter rates, and habitat characteristics of toothed cetaceans in the Bay of Biscay and adjacent waters from platform-of-opportunity data. *ICES Journal of Marine Science*, 64(5), 1033-1043. <https://doi.org/10.1093/icesjms/fsm067>
- Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10), 551-560. <https://doi.org/10.1002/fee.1436>
- Landeo-Yauri, S. S., Ramos, E. A., Castelblanco-Martínez, D. N., Niño-Torres, C. A. & Searle, L. (2020). Using small drones to photo-identify Antillean manatees: a novel method for monitoring an endangered marine mammal in the Caribbean Sea. *Endangered Species Research*, 41, 79-90. <https://doi.org/10.3354/esr01007>
- Leach, T. H., Winslow, L. A., Acker, F. W., Bloomfield, J. A., Boylen, C. W., Bukaveckas, P. A., Charles, D. F., Daniels, R. A., Driscoll, C. T., Eichler, L. W., Farrell, J. L., Funk, C. S., Goodrich, C. A., Michelena, T. M., Nierzwicki-Bauer, S. A., Roy, K. M., Shaw, W. H., Sutherland, J. W., Swinton, M. W., ... & Rose, K. C. (2018). Long-term dataset on aquatic responses to concurrent climate change and recovery from acidification. *Scientific Data*, 5, 180059. <https://doi.org/10.1038/sdata.2018.59>
- LeDuc, Rick. (2009). Delphinids, Overview. *Encyclopedia of Marine Mammals (Second Edition)*. 298-302. <https://doi.org/10.1016/B978-0-12-373553-9.00072-9>
- Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution*, 25(10), 574-582. <https://doi.org/10.1016/j.tree.2010.06.016>
- Martín Míguez, B., Novellino, A., Vinci, M., Claus, S., Calewaert, J.-B., Vallius, H., Schmitt, T., Pititto, A., Giorgetti, A., Askew, N., Iona, S., Schaap, D., Pinardi, N., Harpham, Q., Kater, B. J., Populus, J., She, J., Palazov, A. V., McMeel, O., ... & Hernandez, F. (2019). The European Marine Observation and Data Network (EMODnet): Visions and Roles of the Gateway to Marine Data in Europe. *Frontiers in Marine Science*, 6. <https://doi.org/10.3389/fmars.2019.00313>
- Marubini, F., Gimona, A., Evans, P. G. H., Wright, P. J., & Pierce, G. J. (2009). Habitat preferences and interannual variability in occurrence of the harbour porpoise *Phocoena phocoena* off northwest Scotland. *Marine Ecology Progress Series*. 381, 297–310. <https://doi.org/10.3354/meps07893>



- Mason, E. (2009). High-resolution modelling of the Canary Basin oceanic circulation. (Doctoral dissertation). Retrieved from [https://www.researchgate.net/publication/50600114\\_High-resolution\\_modelling\\_of\\_the\\_Canary\\_Basin\\_oceanic\\_circulation](https://www.researchgate.net/publication/50600114_High-resolution_modelling_of_the_Canary_Basin_oceanic_circulation)
- Monteiro, S. S., Bozzetti, M., Torres, J., Tavares, A. S., Ferreira, M., Pereira, A. T., Sá, S., Araújo, H., Bastos-Santos, J., Oliveira, I., Vingada, J. V. & Eira, C. (2020). Striped dolphins as trace element biomonitoring tools in oceanic waters: Accounting for health-related variables. *Science of The Total Environment*, 699, 134410. <https://doi.org/10.1016/j.scitotenv.2019.134410>
- Nature Publishing Group. (2007). Patching together a world view. *Nature*, 450(7171), 761-761. <https://doi.org/10.1038/450761a>
- Nerbonne, J. (2003). Volunteer macroinvertebrate monitoring: assessing training needs through examining error and bias in untrained volunteers. *Journal of the North American Benthological Society*, 22(1), 152-163. <https://doi.org/10.2307/1467984>
- Noren, D. P., & Mocklin, J. A. (2012). Review of cetacean biopsy techniques: Factors contributing to successful sample collection and physiological and behavioral impacts. *Marine Mammal Science*, 28(1), 154–199. <https://doi.org/10.1111/j.1748-7692.2011.00469.x>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... & Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425. <https://doi.org/10.1126/science.aab2374>
- Parsons, E. C. M., Dolman, S. J., Wright, A. J., Rose, N. A., & Burns, W. C. G. (2008). Navy sonar and cetaceans: Just how much does the gun need to smoke before we act? *Marine Pollution Bulletin*, 56(7), 1248-1257. <https://doi.org/10.1016/j.marpolbul.2008.04.025>
- Parsons, E. C. M., Baulch, S., Bechshoft, T., Bellazzi, G., Bouchet, P., Cosentino, A. M., Godard-Coding, C. A. J., Gulland, F., Hoffmann-Kuhnt, M., Hoyt, E., Livermore, S., Macleod, C. D., Matrai, E., Munger, L., Ochiai, M., Peyman, A., Recalde-Salas, A., Regnery, R., Rojas-Bracho, L., ... & Sutherland, W. J. (2015). Key research questions of global importance for cetacean conservation. *Endangered Species Research*, 27(2), 113-118. <https://doi.org/10.3354/esr00655>

- Patel, K. F., Nelson, S. J., Spencer, C. J., & Fernandez, I. J. (2018). Fifteen-year record of soil temperature at the Bear Brook Watershed in Maine. *Scientific Data*, 5. <https://doi.org/10.1038/sdata.2018.153>
- Peltier, H., Beaufils, A., Cesarini, C., Dabin, W., Dars, C., Demaret, F., Dhermain, F., Doremus, G., Labach, H., Canneyt, O. V., & Spitz, J. (2019). Monitoring of Marine Mammal Strandings Along French Coasts Reveals the Importance of Ship Strikes on Large Cetaceans: A Challenge for the European Marine Strategy Framework Directive. *Frontiers in Marine Science*, 6(486). <https://doi.org/10.3389/fmars.2019.00486>
- Perrin, W. F., Caldwell, D. K., & Caldwell, M. C. (1994) Atlantic spotted dolphin *Stenella frontalis* (G. Cuvier, 1829). *Handbook of marine mammals*, 5, 173-190.
- Pershing, A. J., Christensen, L. B., Record, N. R., Sherwood, G. D., & Stetson, P. B. (2010). The Impact of Whaling on the Ocean Carbon Cycle: Why Bigger Was Better. *PLoS ONE*, 5(8), e12444. <https://doi.org/10.1371/journal.pone.0012444>
- Qian, S. S. (2017). *Environmental and Ecological Statistics with R*. Second Edition. Chapman & Hall/CRC.
- Raoult, V., Colefax, A. P., Allan, B. M., Cagnazzi, D., Castelblanco-Martínez, N., Ierodiaconou, D., Johnston, D. W., Landeo-Yauri, S., Lyons, M., Pirotta, V., Schofield, G., & Butcher, P. A. (2020). Operational Protocols for the Use of Drones in Marine Animal Research. *Drones*, 4(4), 64. <https://doi.org/10.3390/drones4040064>
- Redfern, J. V., Ferguson, M. C., Becker, E. A., Hyrenbach, K. D., Good, C., Barlow, J., Kaschner, K., Baumgartner, M. F., Forney, K. A., Ballance, L. T., Fauchald, P., Halpin, P., Hamazaki, T., Pershing, A. J., Qian, S. S., Read, A., Reilly, S. B., Torres, L. & Werner, F. (2006). Techniques for cetacean-habitat modeling. *Marine Ecology Progress Series*, 310, 271-295.
- Robbins, J. R., Babey, L., & Embling, C. B. (2020). Citizen science in the marine environment: estimating common dolphin densities in the north-east Atlantic. *PeerJ*, 8, e8335. <https://doi.org/10.7717/peerj.8335>
- Rogy, P., & Sinclair, A. R. E. (2020). Long-term surveys of age structure in 13 ungulate and one ostrich species in the Serengeti, 1926–2018. *Scientific Data*, 7(359). <https://doi.org/10.1038/s41597-020-00701-0>

- Roman, J., & McCarthy, J. (2010). The Whale Pump: Marine Mammals Enhance Primary Productivity in a Coastal Basin. *PLoS ONE*, 5(10), e13255. <https://doi.org/10.1371/journal.pone.0013255>
- Roman, J., Estes, J. A., Morissette, L., Smith, C., Costa, D., McCarthy, J., Nation, J., Nicol, S., Pershing, A., & Smetacek, V. (2014). Whales as marine ecosystem engineers. *Frontiers in Ecology and the Environment*, 12(7), 377-385. <https://doi.org/10.1890/130220>
- Sala, I., Caldeira, R. M. A., Estrada-Allis, S. N., Froufe, E., & Couvelard, X. (2013). Lagrangian transport pathways in the northeast Atlantic and their environmental impact. *Limnology and Oceanography: Fluids and Environments*, 3, 40-60. <https://doi.org/10.1215/21573689-2152611>
- Sergio, F., Caro, T., Brown, D., Clucas, B., Hunter, J., Ketchum, J., McHugh, K., & Hiraldo, F. (2008). Top Predators as Conservation Tools: Ecological Rationale, Assumptions, and Efficacy. *Annual Review of Ecology, Evolution, and Systematics*, 39(1), 1-19. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173545>
- Shampine, W. J. (1993). Quality assurance and quality control in monitoring programs. *Environmental Monitoring and Assessment*, 26, 143-151. <https://doi.org/10.1007/bf00547492>
- Simmonds, M. P. (2017). Of Poisons and Plastics: An Overview of the Latest Pollution Issues Affecting Marine Mammals. *Marine Mammal Welfare*, 17, 27-37. [https://doi.org/10.1007/978-3-319-46994-2\\_3](https://doi.org/10.1007/978-3-319-46994-2_3)
- Smith, R. C., Dustan, P., Au, D., Baker, K. S., & Dunlap, E. A. (1986). Distribution of cetaceans and sea-surface chlorophyll concentrations in the California Current. *Marine Biology*, 91, 385-402. <https://doi.org/10.1007/bf00428633>
- Smultea, M. A., Jefferson, T. A., & Zoidis, A. M. (2010). Rare Sightings of a Bryde's Whale (*Balaenoptera edeni*) and Sei Whales (*B. borealis*) (Cetacea: Balaenopteridae) Northeast of O'ahu, Hawai'i. *Pacific Science*, 64(3), 449-457. <https://doi.org/10.2984/64.3.449>
- Stelfox, M., Hudgins, J., & Sweet, M. (2016). A review of ghost gear entanglement amongst marine mammals, reptiles and elasmobranchs. *Marine pollution bulletin*, 111(1-2), 6-17. <https://doi.org/10.1016/j.marpolbul.2016.06.034>
- Tanhua, T., Pouliquen, S., Hausman, J., O'Brien, K., Bricher, P., De Bruin, T., Buck, J. J. H., Burger, E. F., Carval, T., Casey, K. S., Diggs, S., Giorgetti, A., Graves, H.,

- Harscoat, V., Kinkade, D., Muelbert, J. H., Novellino, A., Pfeil, B., Pulsifer, P. L., ... & Zhao, Z. (2019). Ocean FAIR Data Services. *Frontiers in Marine Science*, 6(440). <https://doi.org/10.3389/fmars.2019.00440>
- Tiago, P., Pereira, H. M., & Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions. *Basic and Applied Ecology*, 20, 75-85. <https://doi.org/10.1016/j.baae.2017.04.001>
- United Nations. (2020). Conserve and sustainably use the oceans, seas and marine resources for sustainable development. Retrieved 27 December 2020, from <https://sdgs.un.org/goals/goal14>
- Valente, R. (2017). Looking for the migratory whales: Routes of the baleen whales in the Macaronesia. (Master's thesis). Retrieved from <https://hdl.handle.net/10216/108013>
- Valente, R., Correia, A. M., Gil, Á., González García, L., & Sousa-Pinto, I. (2019). Baleen whales in Macaronesia: occurrence patterns revealed through a bibliographic review. *Mammal Review*, 49(2), 129-151. <https://doi.org/10.1111/mam.12148>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B. D., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R. ... & Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>
- Williams, R., Hedley, S. L. & Hammond P. S. (2006). Modeling Distribution and Abundance of Antarctic Baleen Whales Using Ships of Opportunity. *Ecology and Society*, 11(1), 1. <https://doi.org/10.5751/ES-01534-110101>
- Williams, R., Leaper, R., Zerbini, A. N., & Hammond, P. S. (2007). Methods for investigating measurement error in cetacean line-transect surveys. *Journal of the Marine Biological Association of the United Kingdom*, 87(1), 313-320. <https://doi.org/10.1017/s0025315407055154>
- Williams, T. M., Friedl, W. A., Fong, M. L., Yamada, R. M., Sedivy, P. & Haun, J. E. (1992) Travel at low energetic cost by swimming and wave-riding bottlenose dolphins. *Nature*, 355, 821–823. <https://doi.org/10.1038/355821a0>
- Zuur, A. F., Ieno, E., & Smith, G. M. (2007). *Analysing Ecological Data*. Springer-Verlag New York. <https://doi.org/10.1007/978-0-387-45972-1>

Zuur, A. F., Ieno, E., & Elphick, C. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1(1), 3–14.  
<http://www.respond2articles.com/MEE/>

## 7. Appendix

**Table 15.** Results from the Mann-Whitney U tests performed between matched (M) and non-matched (NM) sightings for each group.

Mann-Whitney U Tests Results		
Group	W	P-value
M vs. NM	460857	0.08419
Odontoceti (M vs. NM)	328903	0.4239
Mysticeti (M vs. NM)	4366.5	0.3997
NI (M vs. NM)	2433.5	0.5491

**Table 16.** Results from the Mann-Whitney U test performed between Odontoceti and Mysticeti matched (M) and non-matched (NM) sightings.

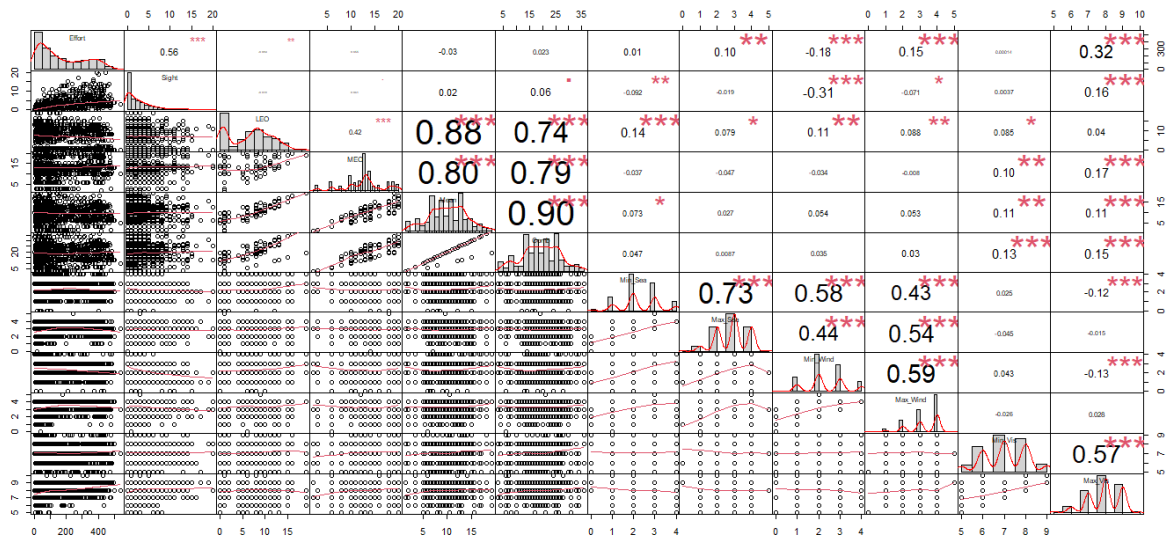
Mann-Whitney U Tests Results		
Group	W	P-value
M (Odontoceti vs. Mysticeti)	1659.5	0.009557
NM (Odontoceti vs. Mysticeti)	334814	<2.2x10 <sup>-16</sup>

**Table 17.** Results from the Mann-Whitney U tests performed between complete validated (C-V) and incomplete/non-validated (IN-V) sightings for each group.

Mann-Whitney U Tests Results		
Group	W	P-value
C-V vs. IN-V	8501	0.03585
Odontoceti (C-V vs. IN-V)	6970	0.5998
Mysticeti (C-V vs. IN-V)	16	0.1139
NI (C-V vs. IN-V)	4	0.6848

**Table 18.** Results from the Mann-Whitney U test performed between Odontoceti and Mysticeti complete validated (C-V) and incomplete/non-validated (IN-V) sightings.

Mann-Whitney U Tests Results		
Group	W	P-value
C-V (Odontoceti vs. Mysticeti)	405.5	0.4355
IN-V (Odontoceti vs. Mysticeti)	461.5	0.01242



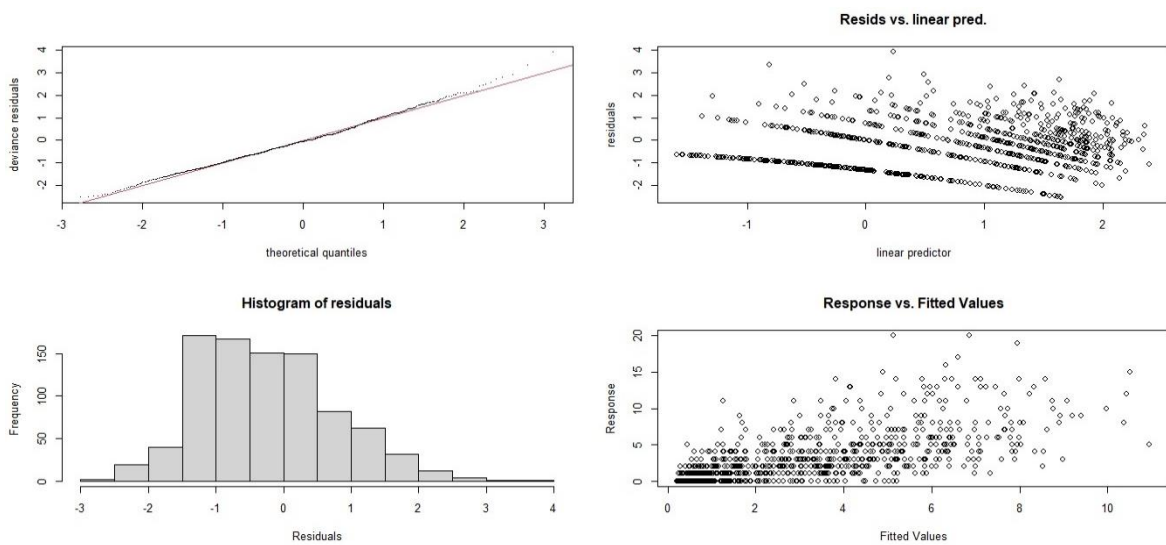
**Figure 12.** Correlation Matrix. Results of Pearson correlations between all pairs of explanatory variables Results. Sight – number of sightings per survey. LEO – evaluation score of Least Experienced Observers per survey. MEO – evaluation score of Most Experienced Observers per survey. Mean – mean of the evaluation scores of the observers in each survey. Comb – accumulated evaluation scores of the observers in each survey. Min\_Sea – minimums of the sea state in each survey. Max\_Sea – maximums of the sea state in each survey. Min\_Wind – minimums of the wind state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey. Max\_Vis – maximums of the visibility in each survey.

**Table 19.** Variance Inflation Factor (VIF) results. LEO – evaluation score of Least Experienced Observers per survey. MEO – evaluation score of Most Experienced Observers per survey. Min\_Sea – minimums of the sea state in each survey. Max\_Sea – maximums of the sea state in each survey. Min\_Wind – minimums of the wind state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey. Max\_Vis – maximums of the visibility in each survey.

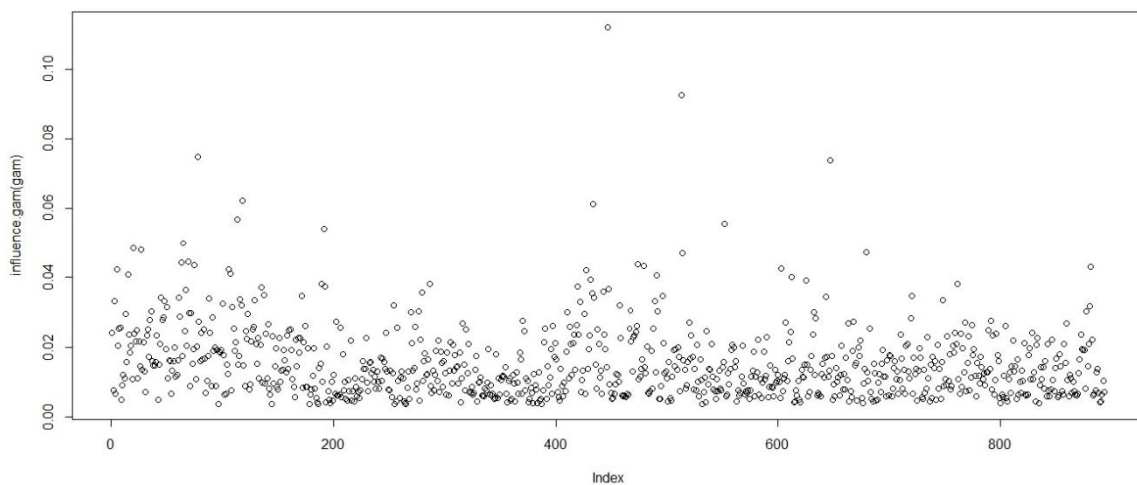
VIF Test Results		
Effort	LEO	MEO
1.343425	1.280023	1.266583
Min_Sea	Max_Sea	Min_Wind
2.808774	2.585341	2.230470
Max_Wind	Min_Vis	Max_Vis
2.007390	1.674973	1.910774

**Table 20.** Results of basis dimension (k) checking (with gam.check). GAM – Generalized Additive Model. EDF – Degrees of Freedom. MEO – evaluation score of Most Experienced Observers. Min\_Sea – minimums of the sea state in each survey. Max\_Wind – maximums of the wind state in each survey. Min\_Vis – minimums of the visibility in each survey.

GAM Check				
Variables	K'	EDF	K-index	P-value
s(Effort)	3.00	2.95	0.91	0.13
s(MEO)	3.00	1.90	0.80	$<2 \times 10^{-16}$
s(Min_Sea)	3.00	2.22	0.81	$<2 \times 10^{-16}$
s(Max_Wind)	3.00	1.78	0.84	$<2 \times 10^{-16}$
s(Min_Vis)	3.00	1.76	0.81	$<2 \times 10^{-16}$

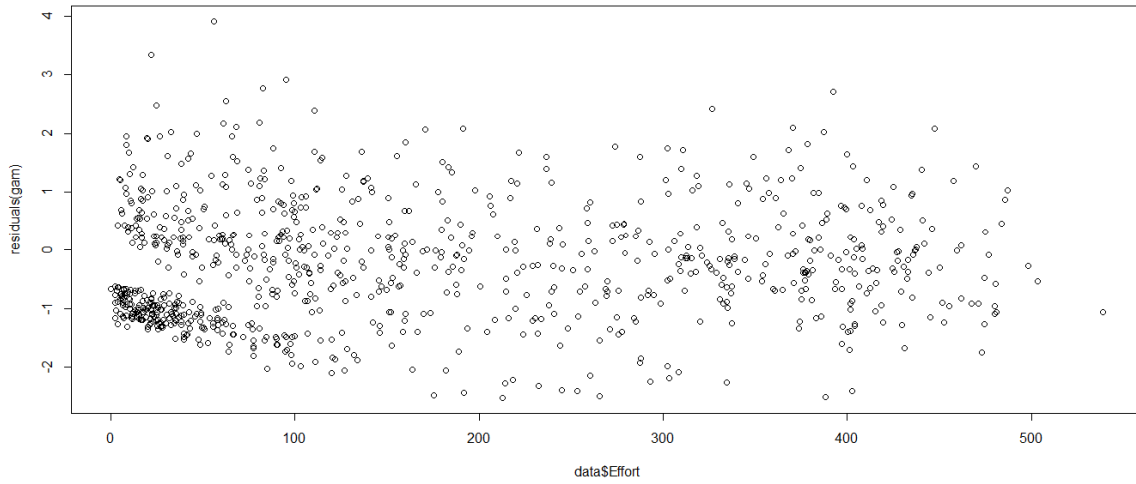


**Figure 13.** GAM Check Plots. GAM – Generalized Additive Model.

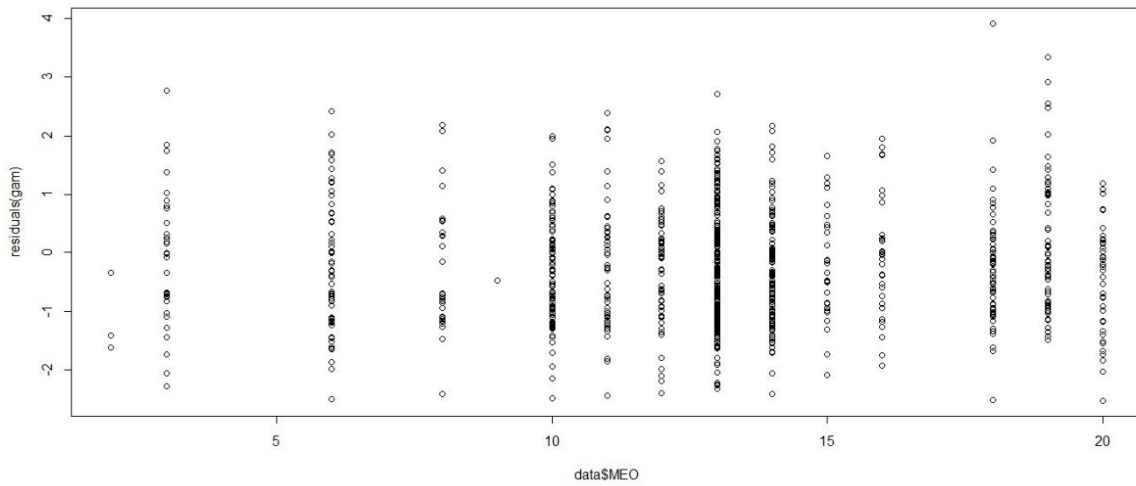


**Figure 14.** GAM influence plot. GAM – Generalized Additive Model.

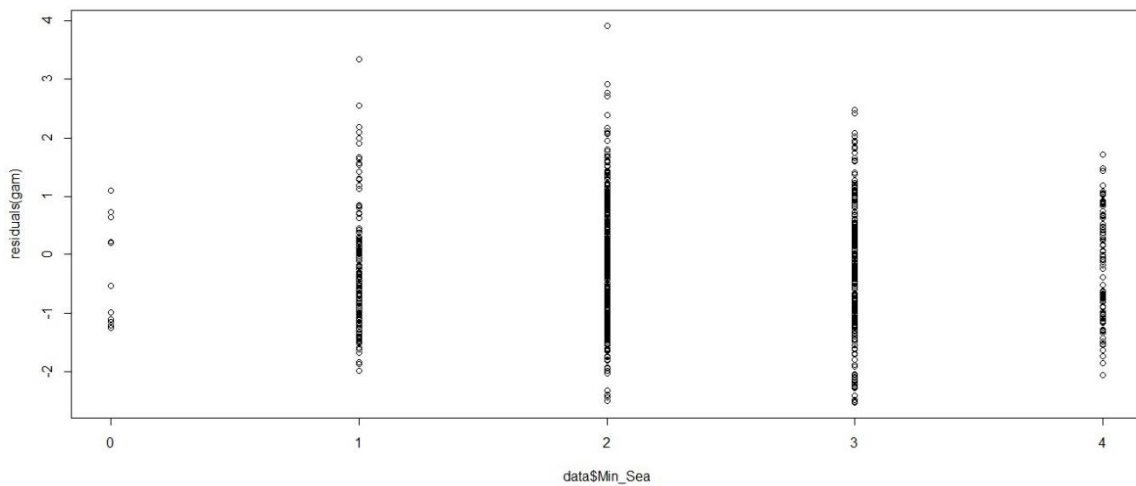




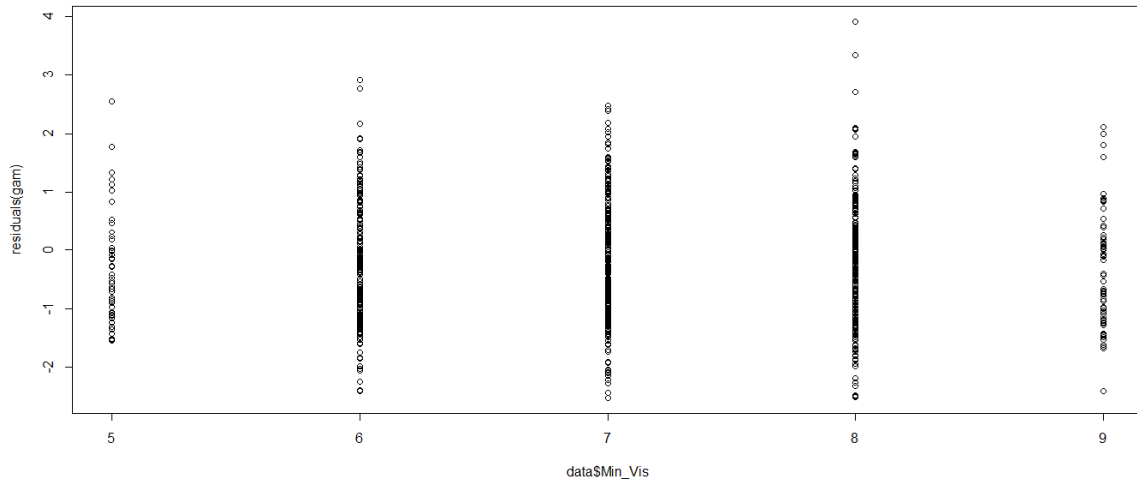
**Figure 15.** GAM effort residuals plot. GAM – Generalized Additive Model.



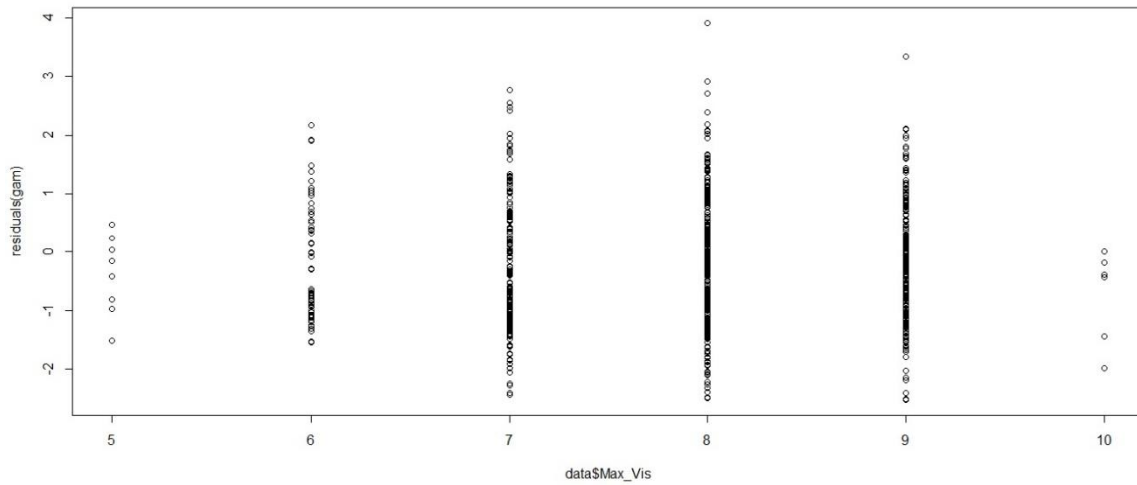
**Figure 16.** GAM MEO residuals plot. GAM – Generalized Additive Model. MEO – evaluation score of Most Experienced Observers per survey.



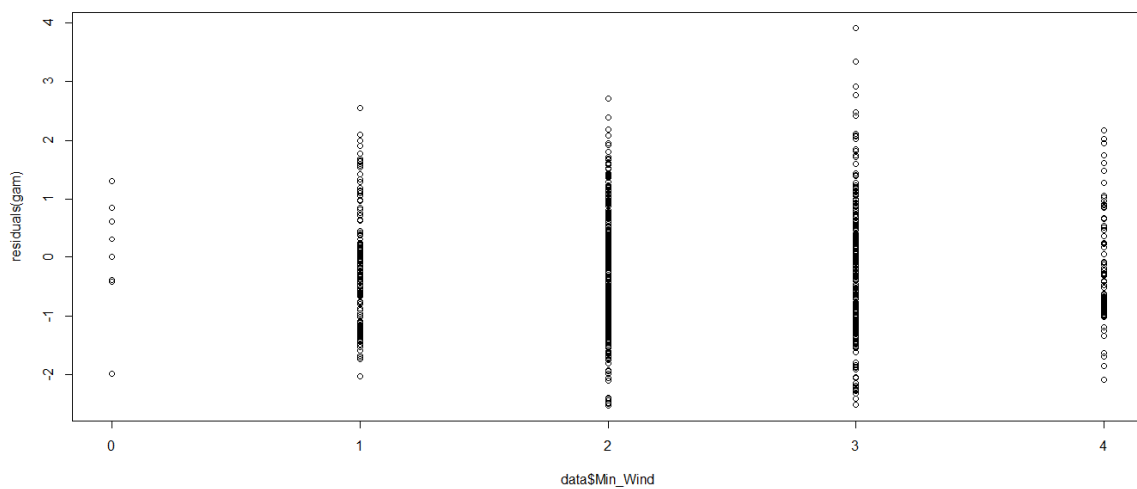
**Figure 17.** GAM Min\_Sea residuals plot. GAM – Generalized Additive Model. Min\_Sea – minimums of the sea state in each survey.



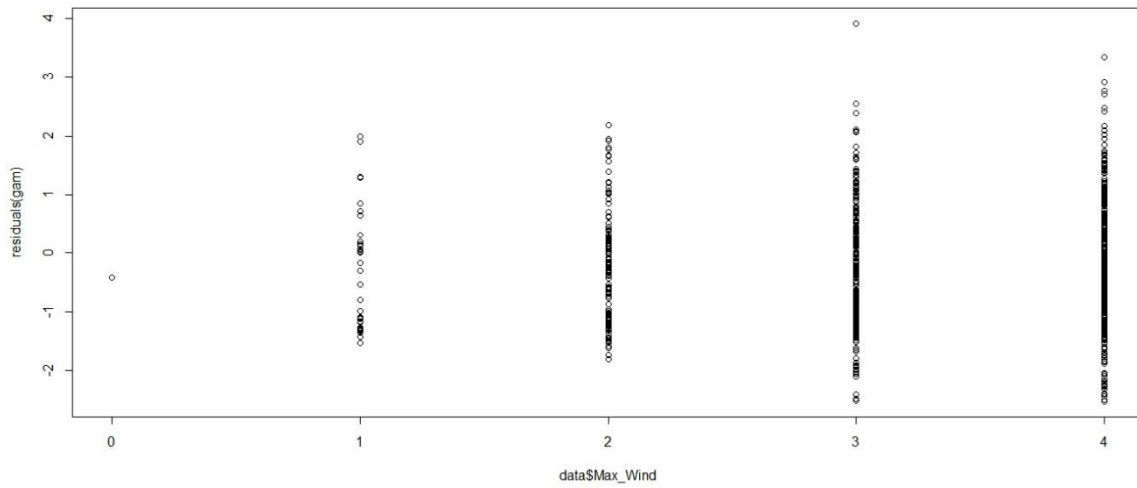
**Figure 18.** GAM Min\_Vis residuals plot. GAM – Generalized Additive Model. Min\_Vis – minimums of the visibility in each survey.



**Figure 19.** GAM Max\_Vis residuals plot. GAM – Generalized Additive Model. Max\_Vis – maximums of the visibility in each survey.



**Figure 20.** GAM Min\_Wind residuals plot. GAM – Generalized Additive Model. Min\_Wind – minimums of the wind state in each survey.



**Figure 21.** GAM Max\_Wind residuals plot. GAM – Generalized Additive Model. Max\_Wind – maximums of the wind state in each survey.