remote sensing

MDPI

Article

Reduction of Species Identification Errors in Surveys of Marine Wildlife Abundance Utilising Unoccupied Aerial Vehicles (UAVs)

Eyal Bigal [1,*], Ori Galili [1], Itai van Rijn [1], Massimiliano Rosso [2], Christophe Cleguer [3], Amanda Hodgson [4], Aviad Scheinin [1] and Dan Tchernov [1]

[1] Morris Kahn Marine Research Station, Department of Marine Biology, Leon H. Charney School of Marine Sciences, University of Haifa, Haifa 3498838, Israel
[2] CIMA Research Foundation, 17100 Savona, Italy
[3] Centre for Tropical Water and Aquatic Ecosystem Research, College of Marine and Environmental Sciences, James Cook University, Townsville, QLD 4810, Australia
[4] Harry Butler Institute, Centre for Sustainable Aquatic Ecosystems, Murdoch University, Murdoch, WA 6150, Australia
* Correspondence: ebigal@staff.haifa.ac.il

Abstract: The advent of unoccupied aerial vehicles (UAVs) has enhanced our capacity to survey wildlife abundance, yet new protocols are still required for collecting, processing, and analysing image-type observations. This paper presents a methodological approach to produce informative priors on species misidentification probabilities based on independent experiments. We performed focal follows of known dolphin species and distributed our imagery amongst 13 trained observers. Then, we investigated the effects of reviewer-related variables and image attributes on the accuracy of species identification and level of certainty in observations. In addition, we assessed the number of reviewers required to produce reliable identification using an agreement-based framework compared with the majority rule approach. Among-reviewer variation was an important predictor of identification accuracy, regardless of previous experience. Image resolution and sea state exhibited the most pronounced effects on the proportion of correct identifications and the reviewers' mean level of confidence. Agreement-based identification resulted in substantial data losses but retained a broader range of image resolutions and sea states than the majority rule approach and produced considerably higher accuracy. Our findings suggest a strong dependency on reviewer-related variables and image attributes, which, unless considered, may compromise identification accuracy and produce unreliable estimators of abundance.

Keywords: aerial surveys; cetaceans; dolphins; drones; false positive detections; marine mammals; misclassification; trial experiments

Citation: Bigal, E.; Galili, O.; van Rijn, I.; Rosso, M.; Cleguer, C.; Hodgson, A.; Scheinin, A.; Tchernov, D. Reduction of Species Identification Errors in Surveys of Marine Wildlife Abundance Utilising Unoccupied Aerial Vehicles (UAVs). *Remote Sens.* **2022**, *14*, 4118. https://doi.org/10.3390/rs14164118

Academic Editors: Luis González Vilas, Jesus Torres Palenzuela and Laura González García

Received: 26 May 2022
Accepted: 7 August 2022
Published: 22 August 2022

1. Introduction

Aerial surveys are recognised as a standard technique for estimating wildlife abundance across large spatial scales. They typically employ light aircraft from which trained observers record animal sightings along pre-designed search paths; if complying with the critical assumptions of line-transect sampling [1], the densities of species within the covered area can then be scaled up to the broader study region. Although applicable to various terrestrial and marine species, aerial surveys may prove hazardous, costly, and logistically challenging to implement [2]. Thus, it is essential to explore the potential use of new techniques for abundance estimation at a wide variety of scales.

In the past decade, unoccupied aerial vehicles (UAVs, or drones) have been increasingly employed for numerous civilian applications and touted as a revolutionary tool for wildlife conservation, e.g., [3]. These versatile platforms offer high data quality and accessibility to remote environments [4–6]. In addition, their utility for abundance estimation has been demonstrated for various taxa, including elephants, marine mammals, sea

turtles, sharks, jellyfish, and birds [7–13], yet most studies have relied on either raw counts or corrected indices without accounting for detection errors in the new type of data, i.e., digital imagery in place of direct human observations [14]. Nonetheless, evaluating the factors that influence animal detectability is a prerequisite for solid ecological inference and environmental conservation [15,16].

The probability of detection incorporates two classes of errors: false negative and false positive sightings. False negatives are observations that are missed due to the proportion of time an animal is present along the transect but unexposed for detection, i.e., *availability bias*. Contributing factors may include diving episodes or visual obstructions from the aerial perspective. Alternatively, sightings may be missed due to the acuity of observers, i.e., *perception bias* [17]. For instance, fatigue, experience, or competency in the sampling methodology could potentially affect the rate of missed detections; in UAV-based surveys, the ground sample distance (GSD), which results from the flight altitude and resolution of the camera, may also influence perception probability [14]. False negatives are commonly addressed in the literature, and the solutions available in the conventional approach are generally applicable to UAV-based studies [10,14]. For example, satellite telemetry may provide auxiliary data on diving behaviour in marine taxa and, thus, on the magnitude of availability bias, i.e., regardless of the survey platform, e.g., [18]. Similarly, comparing the detections from two or more observers situated on the same side of the aircraft or, instead, image reviewers, may inform perception probabilities [19].

The second class of errors, namely false positives, are records of animals at locations where they are not truly present, typically due to *double counting* or *species misidentification*. The former refers to the repeated sampling of individuals moving between transects or appearing in consecutive images due to spatial overlap. The latter relates to the erroneous classification of other objects, including animals and background features, as the survey's focal species. In UAV-based surveys, avoiding those errors may depend upon knowledge of the minimum GSD required to identify target taxa to species, particularly where those are morphologically similar or occur in mixed groups. Nevertheless, although false positives have been shown to induce substantial bias in estimates, even in small probabilities e.g., [20,21], they are commonly assumed insignificant, and only a few studies have addressed them compared to false negatives [14,22]. Published efforts have primarily been devoted to accommodating those errors in occupancy models [20,23–28], and only a small body of literature has focused on developing equivalent frameworks for abundance estimators, e.g., [21,29,30]. The utility of statistical models for aerial transect data may depend on their specific analytical requirements and the capacity of the sampling platform employed.

For example, an important limitation of the modelling approach in occupancy studies is the difficulty of distinguishing the heterogeneity in true positive probabilities from false positive probabilities [20]. Nevertheless, estimators that incorporate unambiguous records for a subset of the survey data, e.g., based on the distinction between multiple degrees of certainty in an observation or by using several detection methods, may overcome this limitation [20,31,32]. In aerial transect surveys, which typically use a single detection technique and, therefore, no verification scheme concerning false positives (but see [10] regarding double counting), the requirement for unambiguous records may be met using multiple-observer protocols. Here, identification certainty is based on the degree of concordance between observers or their integrated level of confidence, and highly-ranked observations are assumed to be correct, e.g., [21,30]. However, the capacity of such records to produce accurate species identification has yet to be explored. Moreover, recent work has highlighted the potential of identification- and confidence-mismatches to result in various abundance estimates under different data-filtering scenarios [33]. Arguably, this finding highlights an important caveat of the conventional methodology; we are unaware of any previous attempts to determine whether reviewing images in consultation with multiple experts may increase identification accuracy.

An alternative approach to inform statistical models on the misidentification process is to conduct trial experiments whereby the true abundance and species are known and error

probabilities are assessed as a function of covariates. For example, McClintock et al. [27] demonstrated the effects of distance, time, ambient noise, and observer abilities on false positive detections using simulated anuran calls; Miller et al. [34] used the same recordings to establish practical predictors of among-observer and among-species variation in error rates. An additional advantage of this approach is the utility of prior information to optimise data collection procedures ([35] and references therein) and, hence, the potential to reduce misidentification probabilities during survey implementation. However, this has yet to be attempted for false positives in aerial survey data, presumably due to the difficulty of simulating observations made in passing mode. In this regard, rotary-wing UAVs, capable of vertical flights and focal follow missions, may present a solution.

In the current study, we combined shipboard and UAV-based surveys to produce imagery of unambiguous species identifications. We investigated the effects of reviewer-related variables and image attributes on identification accuracy and the degree of certainty in observations to produce informative priors for future applications. Furthermore, we compared the capacity of multiple-reviewer frameworks, i.e., agreement- and majority-based identification, to produce accurate records during post-survey image processing. Finally, we evaluated the potential of agreement-based identifications to produce unambiguous data for the modelling approach. We focused our investigation on Mediterranean and Black Sea dolphins, which are characterised by small body size (<3.5 m), morphological resemblance, and occurrence in potentially large, mixed groups, which we considered the most challenging conditions for multi-species surveys. This study presents the first attempt to conduct trial experiments for species identification errors in aerial surveys.

2. Materials and Methods

2.1. Data Collection

We utilised rotary-wing UAVs (Supplementary S1, Table S1) to obtain aerial imagery of *Stenella coeruleoalba* (striped dolphin), *Delphinus delphis* (short-beaked common dolphin), and *Tursiops truncatus* (common bottlenose dolphin) during designated shipboard surveys. The imagery of *S. coeruleoalba* was obtained in July and August 2017 in the northern Ligurian Sea, using a modified DJI Phantom 1 quadcopter (DJI Co., Shenzhen, China) mounted with a GoPro Hero3+ Black edition camera (GoPro Inc., San Mateo, CA, USA; Supplementary S1, Figure S1). Surveys of *D. delphis* and *T. truncatus* were conducted between February 2019 and May 2020 in the eastern Levantine Sea, using four quadcopter models (DJI Co., Shenzhen, China): DJI Mavic Pro, DJI Phantom 3 Advanced, DJI Phantom 4 Advanced, and DJI Phantom 4 Pro. Flights were carried out across various sea conditions and included focal follows in video and still modes. Ground truth species identifications and sea states according to the Beaufort scale were documented per encounter by the shipboard surveyors and added into the metadata; image resolution was calculated based on flight altitude and camera sensor dimensions (Supplementary S1, Equations (S1)–(S5) and Table S2). In total, 15 dolphin encounters were documented in this study (*S. coeruleoalba*, $n = 4$; *D. delphis*, $n = 5$; *T. truncatus*, $n = 6$; Supplementary S1, Table S3). We were not required to obtain an ethics permit for conducting observational surveys of marine mammals; further details on the fieldwork are reported in Supplementary S1 according to the standardised protocol by Barnas et al. [36].

2.2. Data Curation

We arranged the still imagery from each encounter via three binning levels (Table 1): true species identity (ID; *S. coeruleoalba*, *D. delphis* and *T. truncatus*), Beaufort sea state (BSS; 0, 1, 2), and ground sample distance (GSD), i.e., image resolution (cm/pixel; <1, 1–2, 2–3). From each bin, a maximum of three images with at least one dolphin at the surface was selected per encounter. Images that were not captured at nadir were only used if the animals appeared in the bottom third of the frame.

Table 1. Number of images by ground sample distance (GSD), Beaufort sea state (BSS) and true species identity (ID; Dd = *Delphinus delphis*; Sc = *Stenella coeruleoalba*; Tt = *Tursiops truncatus*).

ID	BSS	GSD 0–1 cm/pixel	GSD 1–2 cm/pixel	GSD 2–3 cm/pixel	Total
Dd	0	0	0	0	0
	1	9	9	2	20
	2	9	7	6	22
Sc	0	0	3	3	6
	1	0	3	2	5
	2	0	0	0	0
Tt	0	3	2	0	5
	1	7	8	7	22
	2	6	5	3	14
Total	-	34	37	23	94

In the next step, we uploaded the selected images (n = 94) to an online survey platform (https://surveylegend.com/, accessed on 20 May 2020). Then, we distributed the survey amongst an international team of aerial observers (henceforth, 'reviewers'), all of whom had been trained at a data collection workshop in France, May 2018, held by the Agreement for the Conservation of Cetaceans of the Black Sea, Mediterranean Sea, and contiguous Atlantic Area (ACCOBAMS). Initially, participating reviewers (total, n = 13; Slovenia, n = 1; Spain, n = 4; Great Britain, n = 1; Romania, n = 2; Turkey, n = 1; France, n = 1; Italy, n = 2; Bulgaria, n = 1) were requested to specify their level of experience in conventional aerial surveys as 'training only' (zero surveys), 'novice' (1–10 surveys), or 'expert' (10+ surveys). Then, they were required to ascribe a single species to each image from a list of all small cetaceans occurring in the ACCOBAMS Agreement Area—*S. coeruleoalba*; *D. delphis*; *T. truncatus*; *Grampus griseus* (Risso's dolphin); *Steno bredanensis* (rough-toothed dolphin); and *Phocoena phocoena* (harbour porpoise). Alternatively, images could be classed as 'unidentifiable.' Finally, for each selection, they were requested to choose between two confidence levels: 'guess' or 'definite.' To prevent species identification based on feature resemblance between images from the same encounter, e.g., luminosity or water clarity, the options of revisiting previous selections and resubmitting the survey were disabled. Full-size viewing and zooming were possible in all images. In each survey answer, we considered the assigned species identity as a correct or incorrect selection, i.e., regardless of the associated level of confidence, and treated all unidentified images as incorrect. In addition, each survey answer included the metadata of the subject image, i.e., encounter number, true species identity, sea state, image resolution, and dolphin count (Figure 1), as well as the reviewer's number and level of experience in conventional surveys.

2.3. Data Analysis

We performed a series of analytical procedures to explore the factors affecting identification accuracy in the manual review of images and produce practical recommendations for UAV-based surveys. First, we aimed to determine the importance of reviewer training and the conditions in which UAV-based surveys are conducted. Therefore, we assessed the relative significance of reviewer-related variables and image attributes as predictors of identification accuracy. Here, we employed a linear modelling approach whereby each species identification by a reviewer constituted a standalone data point (n = 1222; Section 2.3.1). Next, we assessed the hierarchy of variables affecting identification accuracy *and* the reviewers' mean level of confidence through a series of random forest models. This analysis considered a single identification per image based on the majority rule approach (n = 94; Section 2.3.2). Furthermore, we generated confusion matrices to determine whether some species were more likely to be mistaken for others present in the data or listed on the survey legend (Section 2.3.3). We repeated Sections 2.3.2 and 2.3.3 for high-confidence selections, i.e., identifications classed by the reviewers as definite, to determine whether these images

yielded more accurate results than the dataset in its entirety. Finally, we performed a Monte Carlo simulation to assess the effect of the number of reviewers on identification accuracy via the agreement- and majority-based frameworks (Section 2.3.4). We performed the analysis within the R statistical environment [37], version 4.3.1 (R Core Team, Vienna, Austria; Supplementary Materials, Data S1).

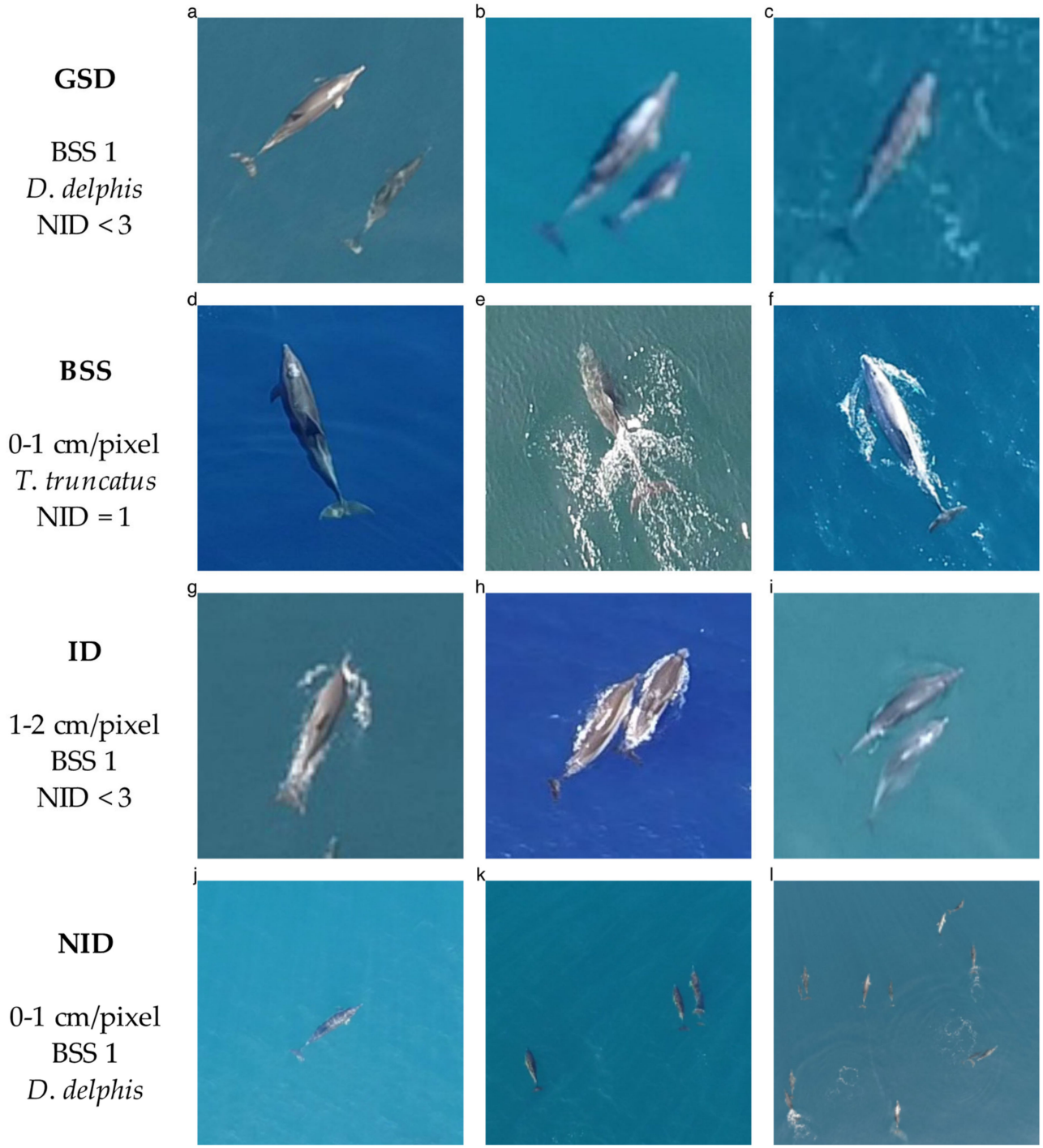

Figure 1. Example effects of image attributes on species identifiability. Each row represents a different variable (**top to bottom:** Ground sample distance, GSD; Beaufort sea state, BSS; true species identity, ID; number of individual dolphins, NID) where all other attributes are constant. The columns represent three different levels of each variable: (**a**–**c**) 0–1 cm/pixel, 1–2 cm/pixel, 2–3 cm/pixel; (**d**–**f**) BSS-0, BSS-1, BSS-2; (**g**–**i**) *Delphinus delphis*, *Stenella coeruleoalba*, *Tursiops truncatus*; (**j**–**l**) Single animal, small group, large group. All images were resized and cropped for illustration purposes.

2.3.1. Reviewer and Image Attributes Effects

We implemented a series of generalised linear mixed models (GLMMs) to assess the relative importance of reviewer-related variables and image attributes as predictors of identification accuracy. This analysis treated each survey answer by a reviewer as a binomial response variable, i.e., correct or incorrect selection (Rev_ans; n = 1222). Initially, we constructed two linear models (Table 2) with image identifying number as a random variable (Img_ID). Model 1 did not incorporate any explanatory variables; in model 2, we included reviewer experience as a pseudo-numerical variable (EXP; 0–2, training only, novice, or expert). Then, we evaluated the relative performance of those models based on Akaike's Information Criterion (AIC) and a one-way analysis of variance (ANOVA).

Table 2. Description of statistical models used to predict the accuracy and certainty of species identification in the manual review of images. Abbreviations: REV_ans, reviewer answer; PCS_img, proportion of correct selections per image; CNF_img, mean level of confidence per image; EXP, experience level; GSD, ground sample distance; BSS, Beaufort sea state; ID, true species identity; NID, number of individual dolphins; IMG_ID, image identifying number; REV, reviewer number; ENC, encounter number.

Model	Effect of Interest	Response Variable	Explanatory Variables	Random Variables
1	Null	Rev_ans	None	Img_ID
2	Previous experience	Rev_ans	EXP	Img_ID
3	Image attributes	Rev_ans	GSD, BSS, ID, NID, EXP	None
4	Among-reviewer variation	Rev_ans	GSD, BSS, ID, NID, EXP	REV
5	Encounter	Rev_ans	GSD, BSS, ID, NID, EXP	REV, ENC
6	Image attributes as predictors of accuracy (all data)	PCS_img	GSD, BSS, ID, NID	None
7	Image attributes as predictors of accuracy (high-confidence selections)	PCS_img	GSD, BSS, ID, NID	None
8	Image attributes as predictors of certainty	CNF_img	GSD, BSS, ID, NID	None

Next, we constructed models which, in addition to previous experience, incorporated image attributes as predictors of identification accuracy: ground sample distance (GSD; or image resolution), Beaufort sea state (BSS), true species identity (ID), and the number of individual dolphins (NID). First, we assessed the effects of those variables on identification accuracy without considering among-reviewer variation (model 3). Then, we incorporated generic differences among survey participants by using reviewer number as a random effect (REV; model 4), which allowed for variability in the intercept term of the model. Similarly, to assess the level of dependency between images, we added encounter number as a second random variable (ENC; model 5). Again, model selection was based on AIC scores. In order to determine the relative importance of random and fixed effects, we computed the selected model's difference between the marginal R squared value (R2M) and conditional R squared (R2C) value, providing the variance explained by the random effects and both random and fixed effects combined. The analysis employed the 'glmer' function of the *lme4* library [38].

2.3.2. Majority-Based Identification

In the following analyses, we calculated for each image the proportion of correct selections (PCS_img) from a total of 13 answers by the reviewers. To assess the relative importance of image attributes affecting this proportion, we employed the random forest modelling approach, which represented two key advantages. First, it allowed for the assessment of non-linear effects; we hypothesised that image resolution, for instance, would only facilitate accurate identification at low GSD levels and that a drop in the proportion of correct selections would be observed for higher levels. Second, it enabled inference of hierarchy of significant effects while incorporating numerical *and* categorical variables [39].

The first analysis (model 6) included GSD, BSS, ID, and NID as explanatory variables and PCS_img as the response variable. We then reimplemented this model for the subset of high-confidence selections (model 7) to establish the conditions in which identification certainty enabled reliable prediction of accuracy. Finally, to determine whether higher identification certainty could be achieved through specific survey conditions, we explored the effects of the above attributes on the mean level of confidence per image (CNF_img, model 8) as a pseudo-numerical response variable (0–2; unidentifiable, guess, and definite). To implement the models, we employed the 'randomForest' function of the *randomForest* library [40]; we visualised the results using the 'rpart' function of the *rpart* library [41]. To assess the relative importance of predictors, we employed the percent increase in the mean squared error index (% IncMSE), using the 'measure_importance' function of the *randomForestExplainer* library [42].

2.3.3. Confusion Matrix

As mentioned above, the reviewers were requested to select one out of six cetacean species for every survey image. In the following analysis, we were interested in determining if some species were more likely to be mistaken for others listed on the legend and whether misidentification rates were symmetric, i.e., the probability of confusing one species for another was the same for the opposite scenario. Therefore, we constructed a confusion matrix with three rows and six columns for the species occurring in the survey images and legend, respectively, and with PCS_img as the response variable in the cells.

2.3.4. Multiple-Reviewer Frameworks

Finally, in the current study, we assessed the effects of covariates on species identification by the reviewers. Therefore, it was necessary to consider the entire dataset, inclusive of image qualities. However, given enough observations, researchers may choose to retain only the identifications agreed upon by all reviewers. To determine if this framework could be used to produce smaller error probabilities than the majority rule approach, we performed a series of Monte Carlo simulations, configuring all reviewer combinations of 1–13 individuals (n = 8191). For each combination, we calculated the proportion of images agreed upon by all members and the proportion of images identified correctly. Then, we generated the mean of each proportion across all groups comprising the same number of reviewers. Similarly, we used the majority rule approach to calculate the proportion of correct selections for all images and combinations and plotted the relationship between accuracy and group size in each of the alternative frameworks. Moreover, given that researchers will usually perform surveys in one predefined GSD, we generated a separate plot for each level of that variable. Finally, to determine the degree to which identification accuracy in each of those frameworks was also affected by sea state, we generated a different plot for every BSS category in each of the GSD levels examined.

3. Results

3.1. Reviewer-Related Variables and Image Attributes

Initially, we assessed the proportion of correct selections per reviewer (n = 13) and plotted the results by experience (Figure 2). None of the survey participants reported an experience level of training only; the novice reviewers had a higher rate of accurate identifications (77.05 ± 19.91, mean ± SD; n = 7; Supplementary S1, Figure S1) compared to the expert reviewers (76.06 ± 4.7, mean ± SD; n = 6). In line with this finding, including the reviewers' experience level as a predictor (EXP) did not significantly improve performance compared to the null model (AIC: 1084.37 and 1086.13, respectively; ANOVA, p = 0.095; Table 3). Finally, model 5, which included both reviewer and encounter as nested factors (REV and ENC, respectively), had a lower prediction error than models 3 and 4, i.e., the non-nested model and the one including reviewer as the only random effect (AIC: 1002.44, 1140.24, and 1028.6, respectively; Table 3). However, the R2M and R2C values of that model were 0.29 and 0.53, respectively. Hence, the proportion of variance explained by the fixed

effects, i.e., image attributes and experience, was larger than that of the random effects, i.e., among-reviewer variation and encounter (29% and 24%, respectively).

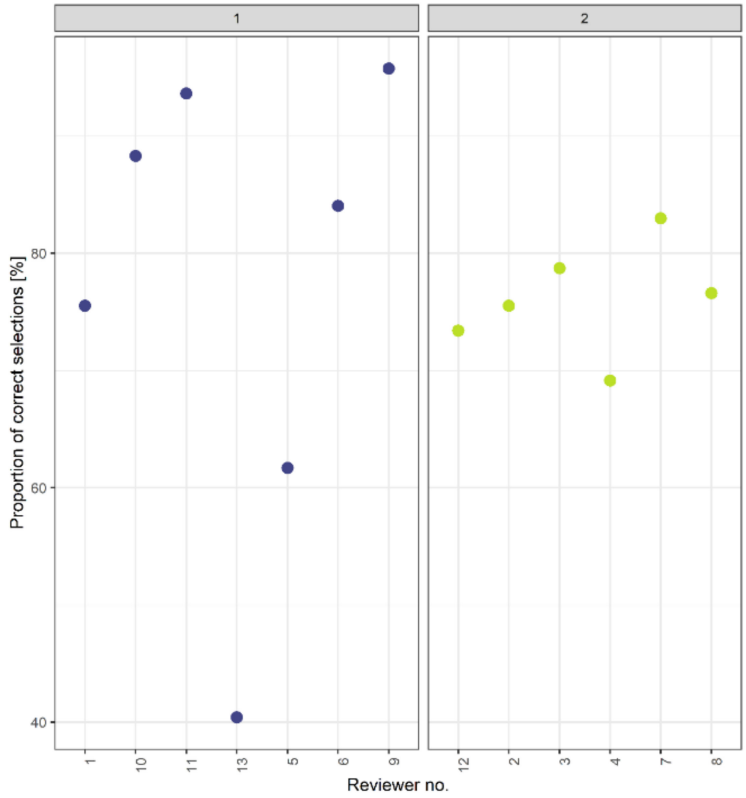

Figure 2. Proportion of correct selections per reviewer. Columns and colours represent the different levels of previous experience in conventional aerial surveys: (**1**, **left**) novice, 1–10 surveys; (**2**, **right**) expert, 10+ surveys.

Table 3. Akaike's information criterion (AIC) values of the general linear mixed models (GLMMS) assessing identification accuracy based on all reviewer answers (n = 1222).

Model	Effect of Interest	AIC
1	Null	1086.13
2	Previous experience	1084.37
3	Image attributes	1140.24
4	Among-reviewer variation	1028.6
5	Among-reviewer variation	1140.24

Notably, although the low AIC score associated with model 5 indicated a certain degree of dependence between images from the same encounter, incorporating this source of variance did not result in a substantial loss of significance to the predictors identified by model 4, namely GSD, BSS, and ID. Therefore, we focused the following analyses on the relative importance of their effects, i.e., regardless of the random variables.

3.2. Majority-Based Identification

When assessing the proportion of correct selections per image (PCS_img; model 6), we found that resolution was the most important predictor. The highest % IncMSE value was assigned to GSD, followed by BSS, ID, and, finally, NID (22.02%, 8.29%, 0.05%, −0.68%, respectively; Var_{ex} = 12.92%). The random forest decision tree (Figure 3) predicted an identification accuracy of 50% when including all images. However, considering only the GSD level of 0–1 cm/pixel improved identification accuracy to 67%, though this subset represented only 36% of the data. Additionally, the model predicted the same proportion

of correct selections for images in the BSS-0 category, i.e., regardless of resolution, and for all BSS-1 images with a pixel size below 2 cm (67% and 53%, respectively). In line with this observation, the variable interaction plot (Figure 4) indicated that the highest accuracy was obtained when both image resolution and sea state were optimal, i.e., GSD $\leq$ 0.5 cm/pixel and BSS-0, with a stronger effect associated with the former. Finally, the composition of high-resolution images by NID corresponded to a further increase of the PCS_img value, though this variable was assigned the lowest % IncMSE value.

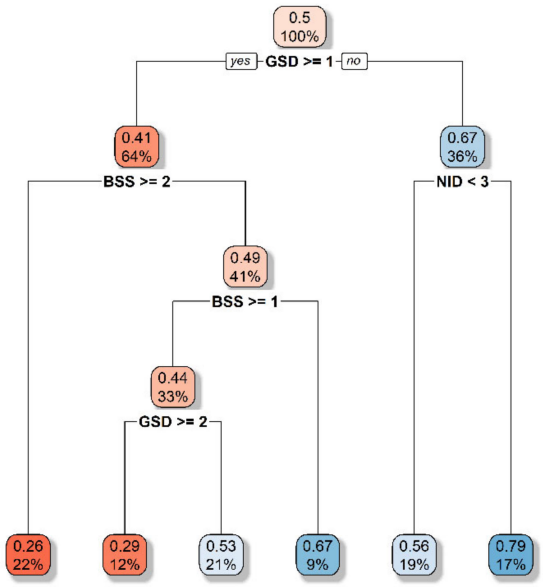

Figure 3. Hierarchy of image attributes affecting identification accuracy across all data. Each node shows the predicted proportion of correct selections (PCS_img) and percentage of observations in that dataset. Colour intensity is proportional to the predicted value. Other model trees, for the high-confidence selections and the reviewers' mean level of confidence, are available in Supplementary S2.

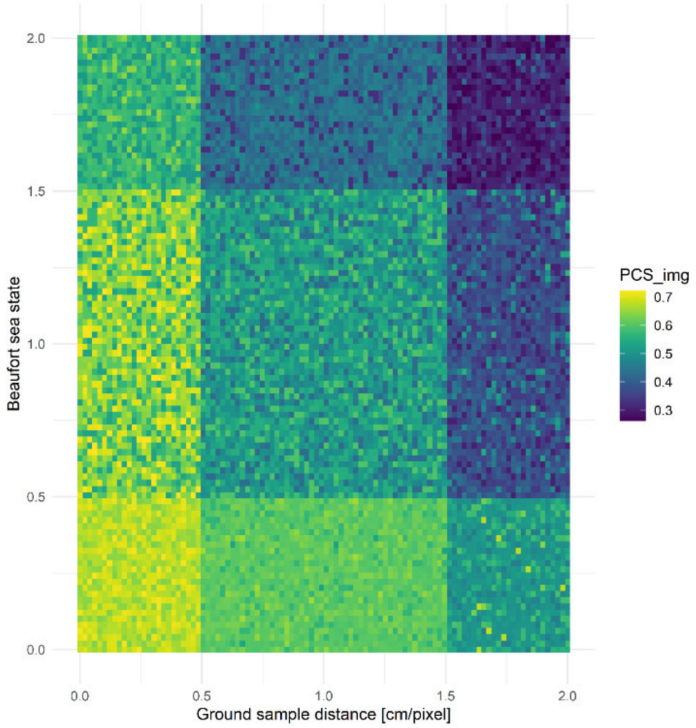

Figure 4. Interaction of ground sample distance (GSD) and Beaufort sea state (BSS) as predictors of the proportion of correct selections per image (PCS_img).

In the analysis of identification accuracy restricted to high-confidence selections (model 8), the highest relative importance was assigned to NID, followed by GSD, ID, and, lastly, BSS (2.36%, −0.42%, −3.10, −6.9%, respectively; Var_{ex} = −11.76%). This model predicted 70% accuracy when considering all images in that subset and, again, the tree displayed a primary split point at a GSD of 1 cm/pixel, with a PCS_img value of 78% (Supplementary S2, Figure S15). Finally, within this subset, the PCS_img value was markedly higher for images displaying three dolphins or more, with 89% of correct identifications.

Finally, we observed similar results in our analysis of the reviewers' mean level of confidence (CNF_img) as a pseudo-numerical variable (0–2; unidentifiable, guess, and definite). The relative importance of predictors in this model followed the same hierarchy observed for identification accuracy across all data: GSD, BSS, ID, and NID (24.7%, 17.53%, 9.5%, 3.14%, respectively; Var_{ex} = 29.54%). The model predicted a mean CNF_img of 1.4 for images captured at a GSD below 1 cm/pixel (Figure 5), compared to 1.1 across all data. Additionally, images captured in a GSD below 2 cm/pixel and a sea state below BSS-1 had a mean CNF_img of 1.2 (Supplementary S2, Figure S10). Other subsets of the data producing an increase in the level of confidence relative to the entire dataset (mean CNF_img = 1.1) included images of *S. coeruleoalba* captured at a GSD of 1 cm/pixel or higher (mean CNF_img = 1.5). Again, we observed a split point at a NID value of 3 within the high-resolution category; the level of confidence in images displaying three dolphins or more was 1.6 compared to 1.2 in images of fewer animals. All model outputs of the random forest analyses, including the decision trees, interaction, and violin plots (e.g., Figures 3–5, respectively), are available in the Supplementary Materials (Supplementary S2, Figures S2–S10 and S15).

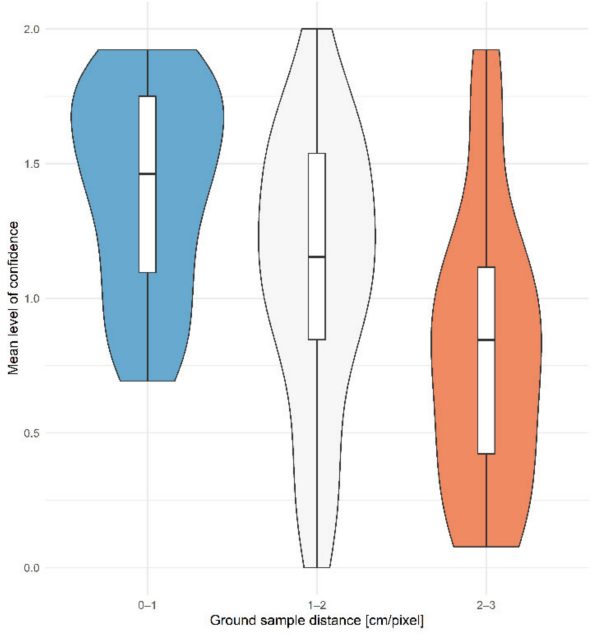

Figure 5. Effect of image resolution on the reviewers' mean level of confidence (CNF_img). Columns and colours represent ground sample distances (GSD; cm/pixel). Column widths indicate the kernel probability density at the corresponding Y axis values. The boxes contained within columns represent interquartile ranges with markers for median values.

3.3. Confusion Matrix

None of the survey species exhibited symmetric misidentification probabilities. *Stenella coeruleoalba* had a slightly lower proportion of correct selections than *D. delphis* and *T. truncatus* (0.5, 0.66, and 0.65, respectively; Table 4). This species was most confused with *D. delphis* and *T. truncatus* (0.19 ± 0.01, mean ± SD) and to a considerably lower extent with *G. griseus*, *P. phocoena*, and *S. bredanensis* (0.04 ± 0.04, mean ± SD). *Delphinus delphis* showed similar misidentification rates for *S. bredanensis*, *S. coeruleoalba*, and *T. truncatus* (0.11 ± 0.03, mean ± SD) but lower proportions for *G. griseus* and *P. phocoena*

(0.01 ± 0.01, mean ± SD). Images of *T. truncatus* produced relatively even distributions of misidentification rates across all other species (0.07 ± 0.02, mean ± SD).

Table 4. Confusion matrix for all images (**top**) and the high-confidence selections (**bottom**).

Observed Species	Dd	Gg	Pp	Sb	Sc	Tt
Dd	0.66	0	0.02	0.12	0.07	0.13
Sc	0.2	0.02	0.02	0.09	0.5	0.18
Tt	0.04	0.07	0.07	0.08	0.08	0.65
Observed Species	**Dd**	**Gg**	**Pp**	**Sb**	**Sc**	**Tt**
Dd	0.83	0	0	0.06	0.06	0.06
Sc	0.14	0	0	0.05	0.78	0.04
Tt	0.01	0.07	0.03	0.04	0.09	0.77

In the subset of high-confidence selections, the PCS_img values of *D. delphis*, *S. coeruleoalba*, and *T. truncatus* increased by 0.17, 0.28, and 0.12 (0.83, 0.78, and 0.77, respectively; Table 4). Here, too, *S. coeruleoalba* was most confused with *D. delphis* (0.14, mean), and less so with *G. griseus*, *P. phocoena*, *S. bredanensis*, or *T. truncatus* (0.02 ± 0.02, mean ± SD). *Delphinus delphis* was mistaken for *S. bredanensis*, *S. coeruleoalba*, and *T. truncatus* (0.06 ± 0, mean ± SD) but not for *G. griseus* or *P. phocoena*. Finally, the misidentification rates of *T. truncatus* were similar across all other species (0.05 ± 0.03, mean ± SD).

3.4. Multiple-Reviewer Frameworks

The agreement-based identification framework resulted in substantial data losses but utilised a broader range of image resolutions and sea states than the majority rule approach (Figure 6). When filtering the data by GSD only, we found that both frameworks displayed lower accuracies with decreasing image resolutions (a–c). In 0–1 cm/pixel-images (a), maximal accuracy was obtained in the agreement-based approach by all groups of more than four reviewers. However, this framework retained only 45% of the data, at best, whereas the majority-based identifications produced 65–75% accuracy and retained all images. In the GSD level of 1–2 cm/pixel (b), we observed a slight decrease in the proportion of images retained *and* the proportion of correct selections as more reviewers were consulted, indicating that agreement did not necessarily imply accuracy in this scenario. In contrast, the accuracy of majority-based identifications within that subset was below 0.5 in all group sizes. Finally, in low-resolution images, i.e., 2–3 cm/pixel (c), identification accuracy increased linearly with group size, whereas data retention remained between 25% in small groups and 10% in large groups. The majority-based identification accuracy was below 0.25 for all reviewer numbers in that GSD category.

To determine whether identification accuracy within each resolution level could be improved by further filtering the data to sea states, we generated a separate plot for each GSD and BSS combination (d–l). Images captured with a pixel size of 0–1 cm produced maximal accuracy in the agreement-based approach across all BSS levels (d–e), regardless of the number of reviewers. In contrast, the majority rule approach only produced maximal accuracy in the BSS-0-images (d) but remained 0.63–0.75 and 0.55–0.7 in the BSS-1 and BSS-2 categories (e–f), respectively. In GSDs of 1–2 cm/pixel (g–i), maximal accuracy in the agreement-based approach was achieved in low winds only, i.e., BSS-0 and BSS-1 (g,h), and depended on a group size of seven reviewers or more, yet data retention in those GSD and BSS scenarios remained below 30%. In contrast, the majority-based framework produced a proportion above 0.5 in the BSS-0 level only (g); none of the two frameworks produced reliable identifications in images of GSD 1–2 cm/pixel and BSS-2 (i). Finally, both frameworks performed well in 2–3 cm/pixel and BSS-0 (j), with maximal accuracy and a proportion of data retention above 0.3 in the agreement-based approach and over 60% accuracy in the majority-based identifications. In the image resolution and sea state combination of 2–3 cm/pixel and BSS-1 (k), the accuracy of identification in the agreement-based approach displayed a dramatic improvement with the increase in group size, which

corresponded to the linear trend observed in that resolution level when including all sea states (c). However, the proportion of data retention observed for the corresponding group size was approximately 10%, which, based on Table 1, represented two images only. Finally, identification accuracy in both frameworks was below 0.35 in images of 2–3 cm/pixel and BSS-2 (l). Available in Supplementary S2 are the above simulation outputs for all images, arranged by GSD levels (Figures S11–S13 and S16).

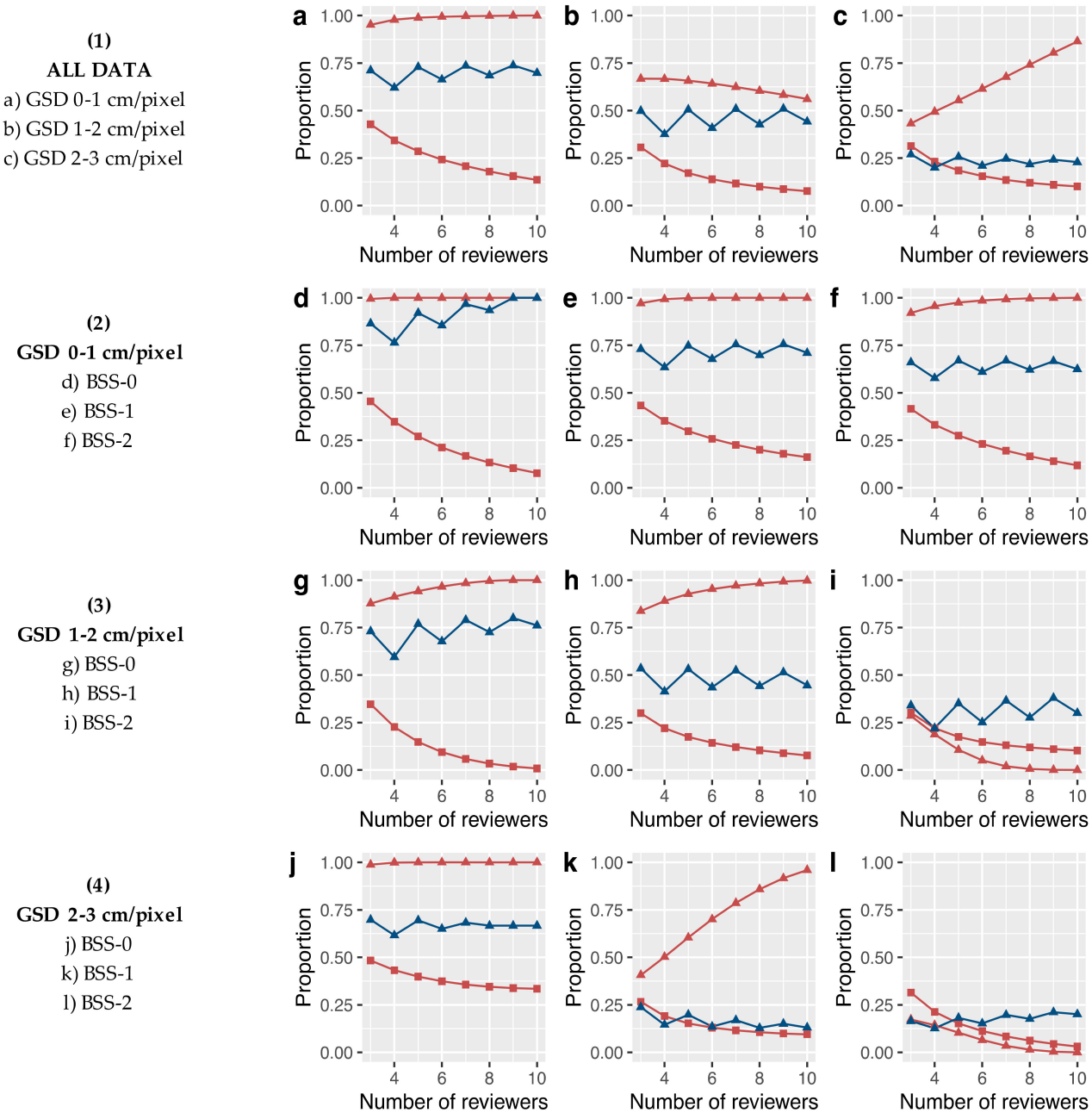

Figure 6. Effect of reviewer group size on the proportion of correct selections (triangles) and the proportion of retained data (rectangles) in the agreement-based framework (red) and the majority-rule approach (blue). Panel 1 displays the above relationship for each of the three ground sample distance (GSD) levels: (**a**) 0–1 cm/pixel; (**b**) 1–2 cm/pixel; (**c**) 2–3 cm/pixel. Panels 2, 3, and 4 depict the same relationship for each of the three Beaufort sea state (BSS) categories in the GSD levels of 0–1 cm/pixel, 1–2 cm/pixel, and 2–3 cm/pixel, respectively: (**d**,**g**,**j**) BSS-0; (**e**,**h**,**k**) BSS-1; (**f**,**i**,**l**) BSS-2.

4. Discussion

We conducted trial experiments to establish practical recommendations for reducing species misidentification probabilities when collecting, processing, and analysing UAV images during surveys of marine wildlife abundance. In the following sections, we discuss the results and limitations of our approach concerning each of those three survey phases and highlight relevant research directions for future applications of the proposed methodology.

4.1. Data Collection

We assessed the effects of image attributes, including resolution, sea state, true species identity, and the number of individual dolphins, on the accuracy of identifications by 13 trained reviewers and their mean level of confidence. We found that a GSD < 1 cm/pixel or BSS-0 was required to correctly identify 67% of the images (Figure 3). This finding suggests that if the aim of a UAV-based survey is to identify dolphins to species while also maintaining sufficient spatial coverage, researchers will require high-resolution cameras. Alternatively, reliable species identification in UAV images may be limited to surveys in low wind. We also found that identification accuracy was substantially higher if only considering the observations classed as 'definite' (PCS_img = 0.7), suggesting that certainty may serve as a useful indicator of accuracy. However, to facilitate a large proportion of high-confidence answers, a substantial amount of data might need to be collected, processed, and discarded, which may prove fiscally unviable. Conversely, given that even small probabilities of false positive errors may propagate into biased parameter estimates of population abundance [27,29], retaining the entire set of images would limit the survey's utility for ecological inference and conservation efforts. Our results further indicated larger PCS_img values for the subset of high-confidence selections obtained at a GSD < 1 cm/pixel (0.78). Again, the requirement for high-resolution images might come at the expense of a large area coverage. Therefore, decisions on flight parameters will ultimately depend on the proportion of sightings that need to be classed as confident for understanding the abundance of target species.

Concerning the effect of the true species identity, *S. coeruleoalba*-images with a GSD > 1 cm/pixel produced a higher level of confidence but did not improve accuracy. Similarly, images of *D. delphis* or *T. truncatus* captured at a GSD of 1–2 cm/pixel and a sea state of BSS-0 or BSS-1 produced a high level of confidence relative to the entire dataset (1.2 and 1.1, respectively) and were not associated with improved identification accuracy. However, we attribute the effect of NID to the absence of *S. coeruleoalba* from the 0–1 cm/pixel and BSS-2 categories (Table 1). Therefore, a high level of confidence could not be used as an indicator of identification accuracy when composing the images by species. Moreover, the results of our confusion matrices indicated similar proportions of correct identifications for all three species when considering the entire dataset or only the high-confidence selections. In line with our expectations, *D. delphis*, *S. coeruleoalba*, and *T. truncatus* were primarily mistaken for each other and not for the other optional species, which we considered of lower resemblance. However, we did not investigate the morphological traits, e.g., colour, shape, or size, affecting misidentification as demonstrated by previous authors, e.g., [30]. A design-based approach to dealing with species resemblance in large-scale abundance surveys would be to stratify the study area based on prior knowledge or expectations concerning spatial variation in underlying species densities. In order to maximise accuracy in areas of overlapping ranges, researchers may need to consider higher misidentification probabilities for similar species and plan a greater sampling effort to increase the proportion of accurate identifications.

Finally, the highest identification accuracy across all data and in the high-confidence selections belonged to the images displaying three dolphins or more, with PCS_img values of 0.79 and 0.89, respectively. An intuitive explanation is that comparing multiple animals could facilitate identification based on a broader range of viewing angles, swimming depths, or sizes. However, the NID variable was assigned the lowest relative importance as a predictor of identification certainty, suggesting that another mechanism was in effect. An alternative

explanation is that images displaying a small number of individuals may be more prone to misidentification. For example, in the case of partial availability in the frame, e.g., due to asynchronistic diving or swimming in a loose formation, a reviewer may falsely consider more candidate species that occur in smaller clusters. Thus, further research is warranted on how previous knowledge of species-specific group sizes is employed in the observation process.

4.2. Data Processing

Concerning the manual review of images, we first analysed the effect of previous experience and generic differences among participants on identification accuracy. Multiple authors have recognised experience as an essential variable that might explain the rates of false positive detections, e.g., [21,30]; we included it in our work to determine whether training schemes for marine wildlife identification in aerial images could be used to improve accuracy. Consistent with previous studies in the literature of occurrence sampling methods, e.g., [23,30], we did not find significant evidence that this variable impacted the proportion of correct identifications. However, our analysis referred to the reviewers' background in conventional aerial surveys rather than image-based identification. To better understand the potential of reducing error rates through training, future studies may do well to employ individuals trained for the manual review of images. However, eligible candidates may be difficult to find. That said, we showed that among-reviewer variation was a significant predictor of accuracy, suggesting that previous experience in conventional aerial surveys was not necessarily an inadequate predictor, but possibly the index we used to express it, i.e., the number of past surveys. Other predictors of ability may include self-assessment by reviewers or independent testing of their skills [34].

Finally, our analysis of multiple-reviewer frameworks indicated that the overall accuracy of agreement- and majority-based identifications depended primarily on image resolution and sea state, and to a lesser extent, on the number of reviewers consulted. However, the composition of images by sea state facilitated an understanding of specific subsets where a larger number of reviewers did improve the proportion of correct selections in the agreement-based approach. For example, for the GSD level of 1–2 cm/pixel, retaining only the images captured at a sea state of BSS-0 or BSS-1 improved identification accuracy from less than 70% across all group sizes (Figure 6b) to over 80% in small groups (g,h) and 100% in large ones (g,h). Additionally, the agreement-based framework retained a broader range of image resolutions and sea states compared to the majority rule approach. Hence, employing this framework may facilitate considerably higher accuracies in surveys conducted across various sea conditions. Moreover, given the potential of false positive errors to induce substantial bias in estimates, the data losses associated with the agreement-based approach may prove inevitable and should be considered when planning the sampling effort. Contrastingly, in surveys expected to produce a low number of images or high-confidence identifications, or in flights where either GSD or sea state is optimal throughout, the majority rule approach may prove feasible.

4.3. Data Analysis

Finally, we investigated the suitability of UAV-based data for statistical models accommodating false positives in post-survey analyses. More specifically, we were interested in the capacity of multiple-reviewer frameworks that rely on the degree of certainty or agreement to produce unambiguous records, which are essential for the model-based approach [29]. We found that, if only considering identifications classed as 'definite,' the highest proportion of correct selections in the majority rule framework was 0.78 for high-resolution images, i.e., GSD < 1 cm/pixel. Thus, the reliability of high-confidence identifications was insufficient to provide unambiguous records for modelling frameworks based on confidence matches. Furthermore, the pervasiveness of misidentification probabilities in this dataset raises concerns about the reliability of previous inferences of abundance in conventional *and* UAV-based surveys based on the above indices, e.g., [30]. Conversely, as discussed in the previous section, the analysis of multiple-reviewer frameworks revealed

maximal proportions of correct identifications in the agreement-based approach for a wide range of resolutions, i.e., 0–3 cm/pixel, depending on sea state. Our results suggest that, to accommodate false positive errors through post-survey data analysis, only the subset of identifications agreed upon by a sufficient number of reviewers, depending on the filtering scenario, should be considered unambiguous records.

Additionally, our study demonstrates the production of prior information that may optimise survey design, as discussed above, and aid in selecting one analytical approach over another in certain data circumstances. For example, Conn et al. [21] demonstrated the potential to produce reliable inferences using the double-observer framework when experimental data are not available, but a symmetry constraint is imposed; we showed that, for the species investigated in our study, misidentification probabilities were asymmetric and, therefore, an alternative framework might have been preferable for studies focused on those dolphins. Finally, the type of data we produced may explain the variation in misidentification rates in response to technical parameters and environmental conditions. This advantage becomes particularly relevant for surveys conducted across large temporal or spatial scales, where error probabilities may change during data collection, despite the researchers' attempts to avoid unfavourable field conditions.

5. Conclusions

This study presents a methodological approach to assess the probabilities of species identification errors in UAV-based surveys of marine wildlife abundance based on independent experiments. Our results indicate a limited integration potential for UAVs with conventional surveys across large areas where species are morphologically similar. We showed that the correct identification of *Stenella coeruleoalba* (striped dolphin), *Delphinus delphis* (short-beaked common dolphin), and *Tursiops truncatus* (common bottlenose dolphin) will depend on high-resolution cameras or the implementation of surveys in optimal sea conditions, i.e., GSD < 2 cm/pixel or BSS-0, respectively. Image reviewing should employ the agreement-based approach whereby only unanimous records are used, despite the substantial data losses associated with this framework. In identifications performed in the majority rule approach, using the subset of observations that are classed by the reviewers as 'definite' is likely to produce higher accuracies but not unambiguous records. Thus, statistical models relying on their availability should only employ the agreement-based approach and, specifically, images captured with a low GSD or processed by relatively large groups. The technique described above may be applied to other taxonomic groups of morphologically similar species with overlapping ranges. As UAVs evolve towards larger spatial scales and longer flight durations, trial experiments may become essential for reducing species identification errors in wildlife abundance surveys.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/rs14164118/s1, Supplementary S1: Fieldwork protocol, Supplementary S2: Analysis output, Data S1: Analysis script, Data S2: Metadata spreadsheets.

Author Contributions: Conceptualisation, E.B., C.C., A.H., A.S. and D.T.; formal analysis, E.B., O.G. and I.v.R.; funding acquisition, E.B., C.C., A.H., A.S. and D.T.; investigation, E.B., I.v.R. and A.H.; methodology, E.B., O.G., M.R., A.H. and A.S.; writing—original draft, E.B.; writing—review and editing, E.B., O.G., I.v.R., M.R., C.C., A.H., A.S. and D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Prince Albert II of Monaco Foundation (grant No. 2241) as part of the ACCOBAMS Survey Initiative (ASI), coordinated by the Agreement on the Conservation of Cetaceans of the Black Sea, Mediterranean Sea and Contiguous Atlantic Area (ACCOBAMS).

Data Availability Statement: The data presented in this study are openly available in Zenodo at doi:10.5281/zenodo.7013418.

Acknowledgments: We are grateful to the anonymous image reviewers that participated in the survey.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Burnham, K.P.; Anderson, D.R.; Laake, J.L. Estimation of Density from Line Transect Sampling of Biological Populations. *Wildl. Monogr.* **1980**, *72*, 3–202.
2. Hodgson, A.; Kelly, N.; Peel, D. Unmanned Aerial Vehicles (UAVs) for Surveying Marine Fauna: A Dugong Case Study. *PLoS ONE* **2013**, *8*, e79556. [CrossRef] [PubMed]
3. Anderson, K.; Gaston, K.J. Lightweight Unmanned Aerial Vehicles Will Revolutionize Spatial Ecology. *Front. Ecol. Environ.* **2013**, *11*, 138–146. [CrossRef]
4. Chabot, D.; Bird, D.M. Wildlife Research and Management Methods in the 21st Century: Where Do Unmanned Aircraft Fit in? *J. Unmanned Veh. Syst.* **2015**, *3*, 137–155. [CrossRef]
5. Christie, K.S.; Gilbert, S.L.; Brown, C.L.; Hatfield, M.; Hanson, L. Unmanned Aircraft Systems in Wildlife Research: Current and Future Applications of a Transformative Technology. *Front. Ecol. Environ.* **2016**, *14*, 241–251. [CrossRef]
6. Colefax, A.P.; Butcher, P.A.; Kelaher, B.P. The Potential for Unmanned Aerial Vehicles (UAVs) to Conduct Marine Fauna Surveys in Place of Manned Aircraft. *ICES J. Mar. Sci.* **2018**, *75*, 1–8. [CrossRef]
7. Vermeulen, C.; Lejeune, P.; Lisein, J.; Sawadogo, P.; Bouché, P. Unmanned Aerial Survey of Elephants. *PLoS ONE* **2013**, *8*, e54700. [CrossRef]
8. Hodgson, J.C.; Baylis, S.M.; Mott, R.; Herrod, A.; Clarke, R.H. Precision Wildlife Monitoring Using Unmanned Aerial Vehicles. *Sci. Rep.* **2016**, *6*, 22574. [CrossRef]
9. Kiszka, J.J.; Mourier, J.; Gastrich, K.; Heithaus, M.R. Using Unmanned Aerial Vehicles (UAVs) to Investigate Shark and Ray Densities in a Shallow Coral Lagoon. *Mar. Ecol. Prog. Ser.* **2016**, *560*, 237–242. [CrossRef]
10. Hodgson, A.; Peel, D.; Kelly, N. Unmanned Aerial Vehicles for Surveying Marine Fauna: Assessing Detection Probability. *Ecol. Appl.* **2017**, *27*, 1253–1267. [CrossRef]
11. Sykora-Bodie, S.T.; Bezy, V.; Johnston, D.W.; Newton, E.; Lohmann, K.J. Quantifying Nearshore Sea Turtle Densities: Applications of Unmanned Aerial Systems for Population Assessments. *Sci. Rep.* **2017**, *7*, 17690. [CrossRef] [PubMed]
12. Raoult, V.; Gaston, T.F. Rapid Biomass and Size-Frequency Estimates of Edible Jellyfish Populations Using Drones. *Fish. Res.* **2018**, *207*, 160–164. [CrossRef]
13. Cleguer, C.; Kelly, N.; Tyne, J.; Wieser, M.; Peel, D.; Hodgson, A. A Novel Method for Using Small Unoccupied Aerial Vehicles to Survey Wildlife Species and Model Their Density Distribution. *Front. Mar. Sci.* **2021**, *8*, 1–17. [CrossRef]
14. Brack, I.V.; Kindel, A.; Oliveira, L.F.B. Detection Errors in Wildlife Abundance Estimates from Unmanned Aerial Systems (UAS) Surveys: Synthesis, Solutions, and Challenges. *Methods Ecol. Evol.* **2018**, *9*, 1864–1873. [CrossRef]
15. Linchant, J.; Lisein, J.; Semeki, J.; Lejeune, P.; Vermeulen, C. Are Unmanned Aircraft Systems (UASs) the Future of Wildlife Monitoring? A Review of Accomplishments and Challenges. *Mamm. Rev.* **2015**, *45*, 239–252. [CrossRef]
16. Baxter, P.W.J.; Hamilton, G. Learning to Fly: Integrating Spatial Ecology with Unmanned Aerial Vehicle Surveys. *Ecosphere* **2018**, *9*, e02194. [CrossRef]
17. Marsh, H.; Sinclair, D.F. Correcting for Visibility Bias in Strip Transect Aerial Surveys of Aquatic Fauna. *J. Wildl. Manag.* **1989**, *53*, 1017–1024. [CrossRef]
18. Hagihara, R.; Jones, R.E.; Sobtzick, S.; Cleguer, C.; Garrigue, C.; Marsh, H. Compensating for Geographic Variation in Detection Probability with Water Depth Improves Abundance Estimates of Coastal Marine Megafauna. *PLoS ONE* **2018**, *13*, e0191476. [CrossRef]
19. Pollock, K.H.; Marsh, H.D.; Lawler, I.R.; Alldredge, M.W. Estimating Animal Abundance in Heterogeneous Environments: An Application to Aerial Surveys for Dugongs. *J. Wildl. Manag.* **2006**, *70*, 255–262. [CrossRef]
20. Miller, D.A.; Nichols, J.D.; McClintock, B.T.; Campbell Grant, E.H.; Bailey, L.L.; Weir, L.A. Improving Occupancy Estimation When Two Types of Observational Error Occur: Non-Detection and Species Misidentification. *Ecology* **2011**, *92*, 1422–1428. [CrossRef]
21. Conn, P.B.; McClintock, B.T.; Cameron, M.F.; Johnson, D.S.; Moreland, E.E.; Boveng, P.L. Accommodating Species Identification Errors in Transect Surveys. *Ecology* **2013**, *94*, 2607–2618. [CrossRef] [PubMed]
22. Dénes, F.V.; Silveira, L.F.; Beissinger, S.R. Estimating Abundance of Unmarked Animal Populations: Accounting for Imperfect Detection and Other Sources of Zero Inflation. *Methods Ecol. Evol.* **2015**, *6*, 543–556. [CrossRef]
23. Simons, T.R.; Alldredge, M.W.; Pollock, K.H.; Wettroth, J.M. Experimental Analysis of the Auditory Detection Process on Avian Point Counts. *Auk* **2007**, *124*, 986–999. [CrossRef]
24. Royle, J.A.; Link, W.A. Generalized Site Occupancy Models Allowing for False Positive and False Negative Errors. *Ecology* **2006**, *87*, 835–841. [CrossRef]
25. Miller, D.A.W.; Nichols, J.D.; Gude, J.A.; Rich, L.N.; Podruzny, K.M.; Hines, J.E.; Mitchell, M.S. Determining Occurrence Dynamics When False Positives Occur: Estimating the Range Dynamics of Wolves from Public Survey Data. *PLoS ONE* **2013**, *8*, e65808. [CrossRef]
26. Chambert, T.; Campbell Grant, E.H.; Miller, D.A.W.; Nichols, J.D.; Mulder, K.P.; Brand, A.B. Two-Species Occupancy Modelling Accounting for Species Misidentification and Non-Detection. *Methods Ecol. Evol.* **2018**, *9*, 1468–1477. [CrossRef]
27. McClintock, B.T.; Bailey, L.L.; Pollock, K.H.; Simons, T.R. Experimental Investigation of Observation Error in Anuran Call Surveys. *J. Wildl. Manag.* **2010**, *74*, 1882–1893. [CrossRef]
28. McClintock, B.T.; Bailey, L.L.; Pollock, K.H.; Simons, T.R. Unmodeled Observation Error Induces Bias When Inferring Patterns and Dynamics of Species Occurrence via Aural Detections. *Ecology* **2010**, *91*, 2446–2454.

29. Chambert, T.; Hossack, B.R.; Fishback, L.A.; Davenport, J.M. Estimating Abundance in the Presence of Species Uncertainty. *Methods Ecol. Evol.* **2016**, *7*, 1041–1049. [CrossRef]
30. McClintock, B.T.; Moreland, E.E.; London, J.M.; Dahle, S.P.; Brady, G.M.; Richmond, E.L.; Yano, K.M.; Boveng, P.L. Quantitative Assessment of Species Identification in Aerial Transect Surveys for Ice-Associated Seals. *Mar. Mammal Sci.* **2015**, *31*, 1057–1076. [CrossRef]
31. Chambert, T.; Miller, D.A.W.; Nichols, J.D. Modeling False Positive Detections in Species Occurrence Data under Different Study Designs. *Ecology* **2015**, *96*, 332–339. [CrossRef] [PubMed]
32. Miller, D.A.W.; Pacifici, K.; Sanderlin, J.S.; Reich, B.J. The Recent Past and Promising Future for Data Integration Methods to Estimate Species' Distributions. *Methods Ecol. Evol.* **2019**, *10*, 22–37. [CrossRef]
33. Dunshea, G.; Groom, R.; Griffiths, A.D. Observer Performance and the Effect of Ambiguous Taxon Identification for Fixed Strip-Width Dugong Aerial Surveys. *J. Exp. Mar. Bio. Ecol.* **2020**, *526*, 151338. [CrossRef]
34. Miller, D.A.W.; Weir, L.A.; McClintock, B.T.; Campbell Grant, E.H.; Bailey, L.L.; Simons, T.R. Experimental Investigation of False Positive Errors in Auditory Species Occurrence Surveys. *Ecol. Appl.* **2012**, *22*, 1665–1674. [CrossRef]
35. Choy, S.L.; O'Leary, R.; Mengersen, K. Elicitation by Design in Ecology: Using Expert Opinion to Inform Priors for Bayesian Statistical Models. *Ecology* **2009**, *90*, 265–277. [CrossRef]
36. Barnas, A.F.; Chabot, D.; Hodgson, A.J.; Johnston, D.W.; Bird, D.M.; Ellis-Felege, S.N. A Standardized Protocol for Reporting Methods When Using Drones for Wildlife Research. *J. Unmanned Veh. Syst.* **2020**, *8*, 89–98. [CrossRef]
37. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
38. Bates, D.; Maechler, M.; Bolker, B. Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2*, 18–22.
41. Therneau, T.; Atkinson, B.; Ripley, B. Rpart: Recursive Partitioning and Regression Trees. 2019. Available online: CRAN.R-project.org/package=rpart (accessed on 14 May 2022).
42. Paluszynska, A.; Biecek, P.; Jiang, Y. randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance. 2020. Available online: CRAN.R-peoject.org/package=randomForestExplainer (accessed on 14 May 2022).