






# Optimizing automated photo identification for population assessments

Philip T. Patton<sup>1,2</sup>  | Krishna Pacifici<sup>3</sup> | Robin W. Baird<sup>4</sup> | Erin M. Oleson<sup>2</sup> | Jason B. Allen<sup>5</sup> | Erin Ashe<sup>6</sup> | Aline Athayde<sup>7</sup> | Charla J. Basran<sup>8</sup> | Elsa Cabrera<sup>9</sup> | John Calambokidis<sup>4</sup> | Júlio Cardoso<sup>7</sup> | Emma L. Carroll<sup>10</sup>  | Amina Cesario<sup>11,12</sup> | Barbara J. Cheney<sup>13</sup> | Ted Cheeseman<sup>14,15</sup> | Enrico Corsi<sup>4</sup>  | Jens J. Currie<sup>1,16</sup> | John W. Durban<sup>17</sup> | Erin A. Falcone<sup>18</sup> | Holly Fearnbach<sup>17</sup> | Kiirsten Flynn<sup>4</sup> | Trish Franklin<sup>14,19</sup> | Wally Franklin<sup>14,19</sup> | Bárbara Galletti Vernazzani<sup>9,20</sup> | Tilen Genova<sup>21,22</sup> | Marie Hill<sup>2,23</sup> | David R. Johnston<sup>24</sup> | Erin L. Keene<sup>18</sup> | Claire Lacey<sup>1</sup>  | Sabre D. Mahaffy<sup>4</sup> | Tamara L. McGuire<sup>25</sup> | Liah McPherson<sup>1</sup> | Catherine Meyer<sup>26</sup> | Robert Michaud<sup>27</sup> | Anastasia Miliou<sup>28</sup> | Grace L. Olson<sup>16</sup> | Dara N. Orbach<sup>29</sup> | Heidi C. Pearson<sup>30</sup> | Marianne H. Rasmussen<sup>8</sup> | William J. Rayment<sup>24</sup> | Caroline Rinaldi<sup>31</sup> | Renato Rinaldi<sup>31</sup> | Salvatore Siciliano<sup>32</sup> | Stephanie H. Stack<sup>16,33</sup> | Beatriz Tintore<sup>28</sup> | Leigh G. Torres<sup>34</sup> | Jared R. Towers<sup>35</sup> | Reny B. Tyson Moore<sup>5</sup> | Caroline R. Weir<sup>36</sup> | Rebecca Wellard<sup>37,38</sup> | Randall S. Wells<sup>5</sup> | Kymberly M. Yano<sup>2,23</sup>  | Jochen R. Zaeschmar<sup>39</sup> | Lars Bejder<sup>1,40</sup>

## Correspondence

Philip T. Patton, Marine Mammal Research Program, Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, 46-007 Lilipuna Rd, Kāne'ohe, HI 96744, USA. Email: [pattonp@hawaii.edu](mailto:pattonp@hawaii.edu)

**Article impact statement:** Researchers can use high-performance identification algorithms to reduce the cost of population assessments without biasing abundance estimates.

## Funding information

Cooperative Ecosystem Studies Unit; NOAA Fisheries Quantitative Ecology and Socioeconomic Training Program

## Abstract

Several legal acts mandate that management agencies regularly assess biological populations. For species with distinct markings, these assessments can be conducted noninvasively via capture-recapture and photographic identification (photo-ID), which involves processing considerable quantities of photographic data. To ease this burden, agencies increasingly rely on automated identification (ID) algorithms. Identification algorithms present agencies with an opportunity—reducing the cost of population assessments—and a challenge—propagating misidentifications into abundance estimates at a large scale. We explored several strategies for generating capture histories with an ID algorithm, evaluating trade-offs between labor costs and estimation error in a hypothetical population assessment. To that end, we conducted a simulation study informed by 39 photo-ID datasets representing 24 cetacean species. We fed the results into a custom optimization tool to discern the optimal strategy for each dataset. Our strategies included choosing between truly and partially automated photo-ID and, in the case of the latter, choosing the number of suggested matches to inspect. True automation was optimal for datasets for which the algorithm identified individuals well. As identification performance declined, the optimization recommended that users inspect more suggested matches from the ID algorithm, particularly for small datasets. False negatives (i.e., individual was resighted but erroneously marked as a first capture) strongly predicted estimation error. A 2% increase in the false negative rate translated to a 5% increase in the relative bias in abundance estimates. Our framework can be used to estimate expected error of the abundance estimate, project labor effort, and find the optimal strategy for a dataset and algorithm. We recommend estimating

a strategy's false negative rate before implementing the strategy in a population assessment. Our framework provides organizations with insights into the conservation benefits and consequences of automation as conservation enters a new era of artificial intelligence for population assessments.

#### KEYWORDS

artificial intelligence, capture-recapture, cetacean, Jolly–Seber, misidentification, optimization, stock assessment

## INTRODUCTION

Several legal acts and international agreements mandate that wildlife management agencies regularly assess populations. For example, the US Marine Mammal Protection Act (MMPA) directs the National Marine Fisheries Service (NMFS) to assess every marine mammal population (or stock) in US waters every 3 years. This regular monitoring helps inform the conservation and management of each stock. For example, NMFS might list a stock as strategic if an assessment shows that it is declining or that the amount of fisheries bycatch exceeds its potential biological removal (i.e., the amount of human-caused mortality that it could annually sustain and recover) (Bettridge, 2023; Punt et al., 2020; Wade, 1998). Listing a stock as strategic has important management consequences, such as creating a take reduction plan for minimizing fisheries bycatch and any other harms. Of course, regular population assessments are also important for the conservation of species other than marine mammals. The US Endangered Species Act mandates that the Fish and Wildlife Service, or NOAA Fisheries, regularly assess species on the Endangered Species List. Outside the United States, the EU Birds Directive requires that member states regularly assess populations of protected bird species in their territory. Additionally, the EU Marine Strategy Framework Directive (MSFD) relies on population assessments and monitoring programs to show that member states have achieved or are maintaining good environment status, a primary goal of the MSFD, across the EU (Authier et al., 2017). Internationally, population assessments underpin the International Union for Conservation of Nature Red List, which complements the acts and agreements above by shedding light on the status of populations and the risks that they face globally (Braulik et al., 2023).

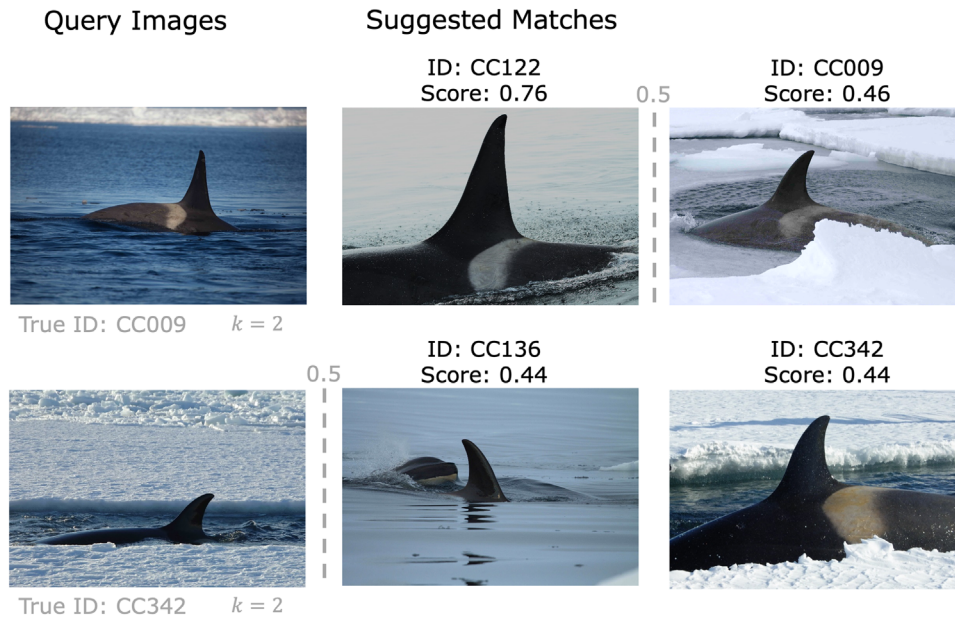
Population assessments typically include an abundance estimate, that is, an estimate of the total number of animals in the population. For example, the MMPA requires an estimate of the minimum population size ( $N_{\min}$ ) to, among other things, estimate potential biological removal (Bettridge, 2023). For species with individually identifying marks—such as several species of cats, whales, and dolphins—researchers can estimate abundance with capture-recapture and photographic identification (photo-ID). Capture-recapture with photo-ID has been applied in many contexts, including camera-trapping surveys of terrestrial mammals (Royle et al., 2009) and shipboard surveys of cetaceans (Hammond et al., 2021). The main advantage of photo-ID is that it allows researchers to apply capture-recapture techniques without having to physically mark animals. As such, agencies can

estimate abundance and thereby inform population assessments without invasively handling animals.

Capture-recapture via photo-ID, however, requires processing considerable quantities of photographic data, which can be time-consuming. For example, Tynce et al. (2016) estimated that data processing for a 2-year study of spinner dolphins (*Stenella longirostris*) cost over 4000 h of labor. This intensive labor partly stems from the need to compare images from each sampling occasion with the reference set, which contains every image of every known individual (Table 1). To facilitate this step, researchers have developed several individual identification algorithms (hereafter, ID algorithm), building off recent advancements in computer vision and artificial intelligence (AI) (e.g., Bergler et al., 2021; Cheeseman et al., 2021; Maglietta et al., 2020; Miele et al., 2021; Schneider et al., 2019). Most of these ID algorithms are partially automated, in that they estimate the similarity between the individual in the query image and every individual in the reference set (Figure 1). Then, a user

**TABLE 1** Terms used in automated identification of individual animals and their definitions.

| Term                | Definition   |
|---------------------|--|
| False negative      | designating the individual in the query image as a new individual when the individual has already been identified, effectively splitting its capture history |
| False positive      | mistaking the individual in the query image for another individual in the dataset, effectively blending 2 capture histories                                  |
| ID algorithm        | individual identification algorithm  |
| Misidentification   | incorrect assignment of an identity to an individual   |
| New individual      | determination that individual in the query image is distinct from all individuals that have already been identified  |
| Partially automated | ID algorithm estimates similarity, then human identifies individual  |
| Query image         | image containing an individual whose identity we wish to know  |
| Reference set       | every image of every known individual  |
| Similarity score    | ID algorithm's estimate of the similarity between 2 individuals  |
| Suggested matches   | most similar individuals in the dataset, listed in order, to the individual in the query image   |
| Truly automated     | ID algorithm estimates similarity and identifies individual  |



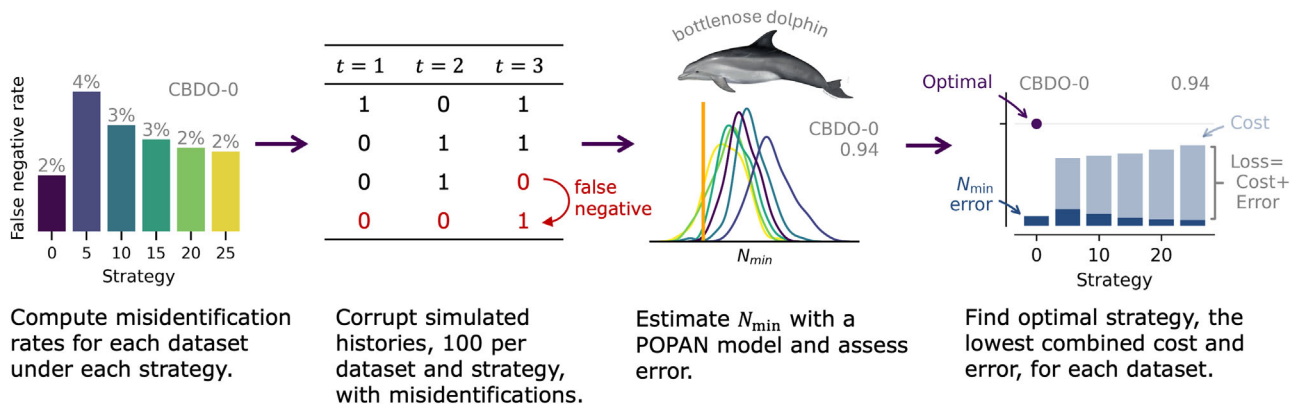
**FIGURE 1** Example output from an individual identification (ID) algorithm (in this case, AnyDorsal) with Antarctic killer whales (*Orcinus orca*). The identities of the individuals in the 2 query images is needed. *Suggested matches* is an ordered list of the closest individuals in the dataset based on similarity score. In both cases, the individual's true identity is the second suggested match; therefore,  $k = 2$ . The new-individual threshold for AnyDorsal is 0.5 (dotted line). Truly automated ID algorithms insert new individual in the suggested matches at this threshold. In this case, a truly automated ID algorithm would have produced a false positive with the first query image and a false negative with the second.

examines the most similar individuals from the reference set and decides whether one of them matches the individual in the query (we refer to this ordered list of most similar individuals as suggested matches). If not, the user adds the individual in question to the dataset as a first capture. Some ID algorithms have been designed for true automation, in that they can make this decision without human intervention (Cheeseman et al., 2021; Maglietta et al., 2020; Patton et al., 2023). The wide variety of ID algorithms and their different designs create several questions for practitioners. Are these algorithms effective enough that they can be truly automated? If not, how many suggested matches should practitioners inspect (Moore et al., 2022)? At one end of the spectrum, the user could only inspect the ID algorithm's first suggested match. At the other end of the spectrum, they could inspect every suggested match, that is, every individual in the dataset. Inspecting fewer suggested matches, or none in the case of true automation, would further reduce the labor associated with population assessments. Inspecting fewer, however, may lead to more misidentifications.

Misidentifications corrupt capture histories, potentially violating assumptions and biasing demographic estimates from capture-recapture models. For instance, capture-recapture models assume that "marks are neither lost nor overlooked, and are recorded correctly" (Williams et al., 2002). This is a challenging assumption for photo-ID because any ID method—be it a truly automated ID algorithm, a human identifying animals with a partially automated ID algorithm, or a human manually identifying animals—might misidentify individuals. Yoshizaki (2007) classified 2 main types of misidentifications with photo-ID, which we refer to as false negatives and false positives. False positives occur when the ID method mistakes one indi-

vidual for another, resulting in a recapture being moved from one capture history to another. False negatives occur when the individual in question has already been identified but the ID method fails to recognize it, thereby adding an erroneous new individual to the dataset. Both types of misidentifications can bias estimates of survival (Morrison et al., 2011; Rakhimberdiev et al., 2022; Tucker et al., 2019) and abundance (Ashe & Hammond, 2022; Bonner et al., 2016; Carlson et al., 1990; Johansson et al., 2020; Lukacs & Burnham, 2005; McClintock et al., 2014; Schofield & Bonner, 2015; Stevick et al., 2001; Urian et al., 2015; Yoshizaki, 2007; Yoshizaki et al., 2009). As such, ID algorithms present practitioners with a challenge—propagating misidentifications at a large scale—and an opportunity—reducing the cost of population assessments.

We explored several strategies for generating capture histories with an ID algorithm. We evaluated trade-offs between labor costs and estimation error—defined as the bias or variance or both of parameter estimates or assessment metrics—in a hypothetical population assessment. To that end, we conducted a simulation study parameterized with the misidentification rates from 39 photo-ID datasets representing 24 cetacean species. Although several of these datasets were not created for capture-recapture, the breadth of datasets provided a wide range of misidentification rates and dataset sizes that could be considered realistic and a wide range of species from which to draw inference. We fed the simulation results into a custom, multi-objective optimization tool to discern the optimal strategy for each dataset (Pease et al., 2021; Sanderlin et al., 2014). Our strategies included choosing between truly and partially automated photo-ID and, in the case of the latter, choosing the number of suggested matches to inspect. We did not consider



**FIGURE 2** Our approach to evaluating strategies for generating capture histories with an identification (ID) algorithm (dataset, 1 of 39 photo-ID datasets of cetaceans;  $N_{\min}$ , minimum population size; CBDO-0, one of the photo-ID datasets [in this case, a catalog of common bottlenose dolphins from Scotland]; POPAN, a parameterization of the Jolly–Seber model). False negatives are produced when an individual is resighted but erroneously logged as a first capture. Illustration courtesy of NOAA Fisheries.

the case of manual photo-ID, although this is also prone to misidentifications (Barlow et al., 2011; Johansson et al., 2020; Morrison et al., 2011). We estimated the misidentification rates from a multispecies ID algorithm (Patton et al., 2023) for each dataset under each strategy to demonstrate how practitioners can readily compute these rates for a generic ID algorithm. To explore these trade-offs, we narrowly defined the context of our hypothetical population assessment. This entailed building capture histories with an ID algorithm for a one-off, open population capture-recapture study in which we used a Jolly–Seber model (Jolly, 1965; Seber, 1965), specifically the parameterization developed by Schwarz and Arnason (1996) known as POPAN. Our goal was to provide insight into the consequences of different strategies for a range of species and to develop a framework for practitioners to evaluate trade-offs as we enter a new era of AI for population assessments.

## METHODS

The work flow for our analyses is in Figure 2; each plot represents a subsection of the methods. To start, we calculated misidentification rates for 39 photo-ID datasets under 6 strategies ( $a$ ) (notation defined in Table 2). These included true automation and partial automation, where the latter varied by the number of suggested matches inspected. We considered  $a \in [0, 5, 10, 15, 20, 25]$ , where  $a = 0$  was true automation and  $a > 0$  was partial automation. For example, with  $a = 5$ , the first 5 suggested matches were inspected. To evaluate the strategies, we simulated 100 capture histories for each strategy and dataset from an open population capture-recapture model, specifically, a Jolly–Seber model (POPAN parameterization). Then, we corrupted these capture histories according to the misidentification rates previously computed for each strategy and each dataset (Figure 2). We estimated apparent survival  $\phi$ , the superpopulation size  $N$ , and, subsequently,  $N_{\min}$  from each simulated history with a Bayesian Jolly–Seber model

and calculated each estimate's error by dataset and strategy (Figure 2). We estimated the cost of each strategy, in terms of labor effort, for each dataset. Finally, we used a custom optimization tool to explore trade-offs between cost and estimation error to discern the optimal strategy ( $a^*$ ) for each dataset (Figure 2).

We had to set, a priori, dozens of values (hyperparameters) for the simulation and optimization (see Appendix S1 for the values and thought process behind them). Whenever possible, we tried to use information from the 39 photo-ID datasets to set these values, with the intention to maximize the ecological realism and relevance to future population assessments. Readers can use the corresponding GitHub repository to replicate our analysis or customize the code to better match their situation (<https://github.com/philpatton/autocapture>).

## Estimating misidentification rates

To capture a breadth of plausible misclassification rates and dataset sizes and a diversity of species, we used the cetacean photo-ID datasets described in Patton et al. (2023). These datasets were curated to develop a multispecies ID algorithm, which we refer to as *AnyDorsal*, for dorsal images of cetaceans (Patton et al., 2023). AnyDorsal is a convolutional neural network that relies on ArcFace, a loss function designed for facial recognition, and transfer learning. Patton et al. (2023) describe the overall predictive performance of AnyDorsal, the truly automated version, for each of the 39 datasets in terms of mean average precision (MAP). The MAP indicates the performance of an ordered set of predictions, in our case, the suggested matches ( $d$ ) (Table 2). Precision is the reciprocal of the position ( $k$ ) of the true identity in  $d$ . In other words, if first prediction is correct, that is,  $k = 1$ , then the precision is  $1/1 = 1$ . If the second prediction is correct, that is,  $k = 2$ , then the precision is  $1/2 = 0.5$ . If the fifth prediction is correct, that is,  $k = 5$ , then the precision is  $1/5 = 0.2$  (Patton et al., 2023). Precision usu-

**TABLE 2** Definitions for notation used in capture-recapture, misidentification, and photo identification.

| Notation                | Definition  |
|-------------------------|---|
| General                 |   |
| $i$                     | index for the photo-ID dataset, $i \in [1, 2, \dots, 39]$                               |
| $a$                     | strategy (i.e., number of suggested matches inspected, $a \in [0, 5, 10, 15, 20, 25]$ ) |
| $j$                     | index for query image $j$   |
| $N_{\min}$              | minimum population size   |
| $R$                     | number of replicates in the simulation  |
| $r$                     | index for replicate in the simulation   |
| $d_{i,j}$               | ordered set of most similar individuals for query image $j$ of dataset $i$              |
| $k_{i,j}$               | position of the true identity in $d_{i,j}$ for query image $j$                          |
| Capture-recapture model |   |
| $Y_{m,t}$               | binary variable indicating capture of individual $m$ on occasion $t$                    |
| $t$                     | index for sampling occasion   |
| $T$                     | number of sampling occasions  |
| $m$                     | index for individual in capture history   |
| $n$                     | index for alternative individual in capture history                                     |
| $\phi$                  | apparent survival probability   |
| $b$                     | vector of true entry probabilities (also known as PENT)                                 |
| $N_i$                   | superpopulation size for dataset $i$  |
| $p_i$                   | capture probability for dataset $i$   |
| Misidentification       |   |
| $FP_{i,a}$              | empirical rate of false positives for dataset $i$ and strategy $a$                      |
| $FN_{i,a}$              | empirical rate of false negatives for dataset $i$ and strategy $a$                      |
| $\alpha_{i,a}$          | rate of misidentifications from evolving marks (type of false negative)                 |
| $\gamma_{i,a}$          | rate of misidentifications from ghosts (type of false negative)                         |
| $\delta_{i,a}$          | rate of false positive misidentifications   |
| $e_{m,t}$               | type of recapture: correct, false positive, ghost, or mark change                       |
| $q$                     | index for erroneous individual resulting from false negative                            |
| Optimization            |   |
| $RBIAS(\theta)$         | relative bias of an estimate $\theta$   |
| NE                      | $N_{\min}$ error  |
| $C_{i,a}$               | cost of strategy $a$ for dataset $i$  |
| $L_{i,a}$               | loss associated with each strategy $a$ and dataset $i$                                  |
| $a_i^*$                 | optimal strategy for dataset $i$ , that is, the strategy with the lowest loss           |
| $w_{\theta}, w_C$       | weights attributed to estimation error $\theta$ and cost, $w_{\theta} + w_C = 1$        |

ally has a cutoff. In this case, the cutoff was 5, meaning that the precision was 0 for any  $k > 5$ . The MAP was the mean of the precision across the dataset.

Our goal was to translate these precision scores into misidentification rates, namely, false positive and false negative rates, for each dataset under each strategy (Appendix S1). To do so,

we split each dataset into 2 parts, following the exact training and test splits from Patton et al. (2023). This left roughly two-thirds of each dataset as a reference set, which contained images of known individuals, and one-third of each dataset as a query set, which contained images of known and new individuals. For every image in the query set, we used AnyDorsal to produce 25 suggested matches from the reference set. We used the truly automated version of AnyDorsal, which inserts “new individual” into the suggested matches (Patton et al., 2023) (Figure 1). This gives the model the ability to predict new individual, which is important for evaluating the  $a = 0$  strategy. This step can be skipped if a partially automated ID algorithm is being evaluated. For each query image, we classified first  $a$  suggested matches,  $d_1 \dots d_a$ , as either a correct classification, false positive, or false negative (Appendix S1). Our classification scheme attempted to simulate the situation in which an experienced biologist is identifying animals with an ID algorithm. Appendix S1 describes this classification scheme and an alternative (also see “Discussion” and Appendix S2). Regardless of the method, computing misidentification rates allowed us to determine the downstream effects of different strategies for generating capture histories with an ID algorithm.

### Jolly–Seber data

We simulated  $R = 100$  capture histories from a Jolly–Seber model for each dataset ( $i$ ) and each strategy with Python. For the Jolly–Seber model, simulating a capture history requires setting the number of sampling occasions  $T$ , the superpopulation size  $N$ , the apparent survival probabilities  $\phi$ , the capture probabilities  $p$ , and the entry probabilities  $b$ . We used values that would be plausible for a Jolly–Seber study of a cetacean (see Van Cise et al., 2021 for a delphinid example), although capturing all 24 species’ life histories and all possible sampling designs would be intractable. For every dataset and strategy, we used 10 sampling occasions ( $T$ ) and an apparent survival probability of  $\phi = 0.9$ , which was constant across sampling occasions (Van Cise et al., 2021). We set the initial entry probability to  $b_0 = 0.35$ . Because the entry probabilities must sum to 1, we set the remaining entry probabilities to  $b_{i \neq 0} = (1 - b_0) / (T - 1)$ . We allowed  $N$  and  $p$  to vary by dataset in a scheme described in Appendix S1.

### Misidentification process

We randomly corrupted the true Jolly–Seber histories with misidentifications through a custom process, which synthesized several recent advances in misidentification. Specifically, we accounted for false positives (Bonner et al., 2016), a false negative model for mark changes (also known as evolving marks; Yoshizaki et al., 2009), and a false negative model for ghosts (Link et al., 2010). Ghosts and mark changes are 2 flavors of false negative that differ in their effect on subsequent recaptures after the misidentification. Our synthesis produced the following set of assumptions: misidentifications occur only on recaptures; individuals can only be involved

with 1 event—false positive, ghost, mark change, or correct identification—per occasion; recaptures are classified as mark changes with probability  $\alpha$ , ghosts with probability  $\gamma$ , false positives with probability  $\delta$ , or correct identifications with probability  $1 - (\alpha + \gamma + \delta)$ ; false positives result in allocating the recapture from individual  $m$  to individual  $m'$ ; mark changes result in adding an erroneous new individual to the capture history, and subsequent recaptures are allocated to the new, erroneous individual; ghosts result in adding an erroneous new individual to the capture history, and subsequent recaptures are allocated to the correct individual.

In other words, when individual  $m$  was recaptured at time  $t$ , we classified the recapture type with a categorical distribution,  $e_{m,t} \sim \text{Categorical}(\pi_{i,a})$ , where  $\pi_{i,a} = [\alpha_{i,a}, \gamma_{i,a}, \delta_{i,a}, 1 - (\alpha_{i,a} + \gamma_{i,a} + \delta_{i,a})]$ . Each cell probability in  $\pi$  corresponds to a recapture type: false negative (mark change), false negative (ghost), false positive, correct classification. If a recapture was classified as correct  $e_{m,t} = [0, 0, 0, 1]$ , that is, no misidentification was made, we left  $Y_{m,t}$  unchanged, where  $Y_{m,t}$  is a binary variable representing the capture of individual  $m$  at time  $t$ . If the recapture was a false positive,  $e_{m,t} = [0, 0, 1, 0]$ , we set  $Y_{m,t} = 0$  and  $Y_{m',t} = 1$ , where  $m'$  is a generic index for the mistaken-with individual. If the recapture was classified as one of the 2 types of false negative, we set  $Y_{m,t} = 0$ , created a new all zero history at index  $q$ , then set  $Y_{q,t} = 1$ . If the false negative was caused by a mark change  $e_{m,t} = [1, 0, 0, 0]$ , any  $Y_{m,t+} = 1$  were allocated to  $Y_{q,t+}$ . If the false negative resulted in a ghost  $e_{m,t} = [0, 1, 0, 0]$ ,  $Y_{m,t+}$  remained unchanged. The code for the misidentification process is at `miss_id.py` in this article's corresponding repository.

We set  $\delta_{i,a}$  for each dataset and strategy as the total number of false positive query images. These images were classified with the scheme described in Appendix S1, and  $\delta_{i,a}$  was divided by the total number of query images. To set  $\alpha_{i,a}$  and  $\gamma_{i,a}$ , we first calculated the false negative rate for each dataset and strategy. We defined the false negative rate as the number of false negative query images—which we classified with the scheme described in Appendix S1—divided by the total number of query images. Without prior knowledge of the cause of the false negatives, we set both  $\alpha_{i,a}$  and  $\gamma_{i,a}$  to one-half of the false negative rate.

## Estimating parameters and their error

We estimated  $N$ ,  $b_0$ ,  $p$ , and  $\phi$  from each set of simulated capture histories with a Bayesian formulation of the Jolly–Seber model in PyMC (Abril-Pla et al., 2023). For speed, we used a marginalized version of the model (McCrea & Morgan, 2015; Yackulic et al., 2020). We modeled the probabilistic parameters as constant, with the following prior distributions:  $p \sim \text{Uniform}(0, 1)$ ,  $\phi \sim \text{Uniform}(0, 1)$ ,  $b_0 \sim \text{Uniform}(0, 1)$ , where  $b_{t \neq 0} = (1 - b_0) / (T - 1)$  and  $\sum b = 1$ . This corresponds to model  $p(\cdot)$ ,  $\phi(\cdot)$ ,  $b(\cdot)$  in the dot notation of capture-recapture (Williams et al., 2002). To permit the use of the No-U-Turn Sampler (NUTS) version of Hamiltonian Monte Carlo, we used a flat prior for  $N$ , essentially,  $N \sim \text{Uniform}(-\text{inf}, \text{inf})$

(Hoffman & Gelman, 2014). This has the unfortunate side effect of treating  $N$  as a continuous variable. We believed this disadvantage, however, was outweighed by the greater sampling efficiency of the NUTS algorithm relative to the Metropolis–Hastings sampler (Monnahan et al., 2017). We fit each model with the version of the NUTS offered by PyMC. We simulated the model with 4 chains, 5000 tuning draws, and 10,000 post-tune draws. We checked for convergence with the Gelman–Rubin statistic (Gelman & Rubin, 1992) and by monitoring divergent transitions from the sampler. We calculated a derived quantity, the minimum population size, with the posterior statistics of  $N$ . The minimum population size ( $N_{\min}$ ) is a required component of stock assessment reports under the US MMPA (Bettridge, 2023; Wade, 1998). Following guidelines from the US Department of Commerce (Bettridge, 2023), we calculated  $N_{\min}$  as,

$$N_{\min,r} = \frac{\hat{N}_r}{\exp\left(0.842\sqrt{\log\left(1 + \text{CV}(\hat{N}_r)^2\right)}\right)}, \quad (1)$$

where  $\hat{N}$  is the posterior mean of  $N$  for the replicate  $r$  and  $\text{CV}(\hat{N})_r$  is the posterior coefficient of variation of  $\hat{N}_r$ .

We calculated estimation error with 2 functions: the relative bias (RBIAS) and a custom function we referred to as  $N_{\min}$  error (NE) (Sanderlin et al., 2014) (see below). Relative bias was used to compute the bias of an estimate, scaled by the estimate's true value, which helped with interpretability. For example, a relative bias value of 0.1, meant that we overestimated the parameter by 10%. We calculated RBIAS as,

$$\text{RBIAS}(\theta) = \frac{(1/R) \sum_{r=1}^R (\hat{\theta}_r - \theta)}{\theta}, \quad (2)$$

where  $R = 100$  is the number of replicates,  $\hat{\theta}_r$  is the posterior mean estimate of the parameter of interest  $\theta$  in the  $r^{\text{th}}$  replicate, and  $\theta$  is the true value of the parameter for the dataset (Sanderlin et al., 2014). We calculated estimation error for  $N_{\min}$  with the following custom function (akin to hinge loss):

$$\text{NE} = \frac{(1/R) \sum_{r=1}^R \max(0, N_{\min,r} - N)}{N}. \quad (3)$$

The NE does a better job of calculating estimation error for  $N_{\min}$  because it is unidirectional. That is, NE only penalizes  $N_{\min}$  when  $N_{\min}$  exceeds the true population size  $N$ . It is also relative, meaning that an NE of 0.05 suggests that, on average,  $\hat{N}_{\min}$  exceeded the true population size  $N$  by 5%. Overestimating  $N_{\min}$  like this could have grim conservation consequences because, in the context of US stock assessments,  $N_{\min}$  is used to calculate potential biological removal.

Finally, we were interested in which rate—that is, the false positive rate or the false negative rate—was driving the RBIAS of the superpopulation size  $N$  and the apparent survival  $\phi$

parameters. To that end, we fitted 2 linear models, one for each parameter, with the false negative rate and the false positive rate predicting the parameter's relative bias, that is,  $\text{RBIAS}(\theta) \sim \text{Normal}(\beta_0 + \beta_{\text{FN}}\text{FN}_{i,a} + \beta_{\text{FP}}\text{FP}_{i,a}, \sigma_\epsilon)$ . The  $\beta$  estimates indicated how each rate affects the relative bias of each estimate in terms of strength (magnitude of  $\beta$ ) and direction (the sign of  $\beta$ ).

## Optimization

We identified the optimal strategy for each dataset given the cost and estimation error associated with each strategy. We defined cost in terms of labor effort, specifically, the number of identities that would have to be inspected under each strategy for each dataset (Appendix S1). To find the optimal strategy, we used a loss function ( $L$ ) that mathematically incorporated both cost and estimation error. Then, we sought the optimal strategy for each dataset that minimized the loss. There are several ways to construct such a loss function (Conroy & Peterson, 2012; Pease et al., 2021; Sanderlin et al., 2014). For simplicity and generality, we used a weighted sum of cost and estimation error of  $N_{\min}$  (NE),

$$L_{i,a} = C_{i,a} w_C + \text{NE}_{i,a} w_\theta, \quad (4)$$

where  $C_{i,a}$  is the cost of each strategy for each dataset (see below),  $w_C$  is the weight attributed to cost,  $\text{NE}_{i,a}$  is the  $N_{\min}$  error (Equation 3), and  $w_\theta$  is the weight attributed to  $N_{\min}$  error. For the optimization, we focused on the estimation effort for  $N_{\min}$  because  $N_{\min}$  is important to stock assessments generally and the US MMPA specifically. We defined optimal strategy for each dataset as  $a_i^* = \arg \min_a (L_{i,a})$ , that is, the strategy with the lowest loss for each dataset.

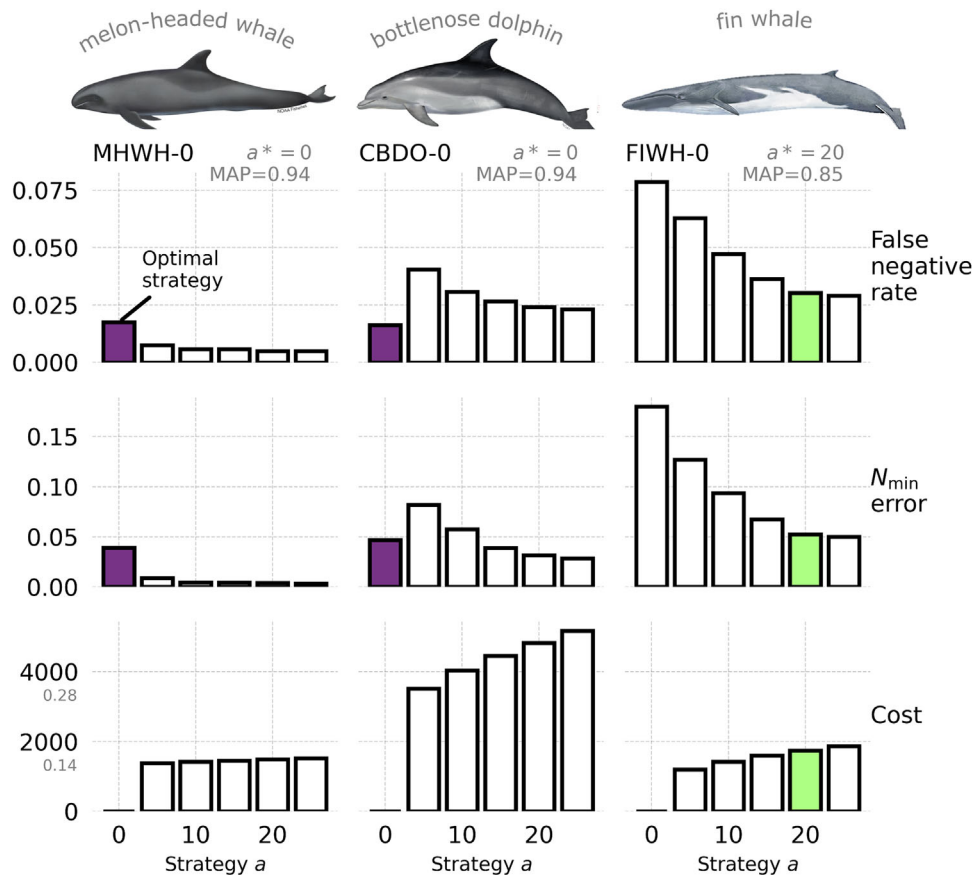
## RESULTS

In general, true automation was optimal,  $a^* = 0$ , for datasets where the algorithm matched images well, that is, the datasets with a high MAP. For example, for the melon-headed whale (*Peponocephala electra*) dataset (MHWH-0) (Figure 3), the  $N_{\min}$  error declined as the number of matches inspected increased ( $\text{NE}_0 = 0.037$  to  $\text{NE}_{25} = 0.004$ ) but these gains were outweighed by the increased cost ( $C_0 = 0$  to  $C_{25} = 0.11$ ). As such, the total loss increased from  $L_0 = 0.037$  to  $L_{25} = 0.114$ . This was also true for the common bottlenose dolphin (*Tursiops truncatus*) dataset from Scotland (CBDO-0), for which the decreased  $N_{\min}$  error ( $\text{NE}_0 = 0.047$  to  $\text{NE}_{25} = 0.028$ ) was not worth the increased cost ( $C_0 = 0$  to  $C_{25} = 0.37$ ) (Figure 3). This pattern held for 12 of the 15 highest MAP datasets (Figure 4). The exceptions were small datasets like that of Red Sea spinner dolphin (*Stenella longirostris*) (dataset SPDO-0), pygmy killer whale (*Feresa attenuata*) (PKWH-0), and Hawaiian short-finned pilot whale (*Globicephala macrorhynchus*) (SFPW-2), which had 110, 49, and 206 query images, respectively (see Appendix S3.1 for codes for each dataset).

As the overall MAP score decreased, the optimal number of suggested matches tended to increase. For example, the  $N_{\min}$  error for the fin whale (*Balaenoptera physalus*) dataset (FIWH-0) decreased from  $\text{NE}_0 = 0.18$  to  $\text{NE}_{20} = 0.052$  (Figure 3). Essentially, this corresponded to a 13-percentage-point drop, from  $N_{\min}$  exceeding  $N$  by 18% to 5%. This decline offset the added cost, from  $C_0 = 0$  to  $C_{20} = 0.12$ . As such, the optimal number of matches checked for this dataset was 20. Similar patterns held for the Bryde's whale (*Balaenoptera edeni*) dataset (BRWH-0), the Adriatic common bottlenose dolphin dataset (CBDO-1), the Commerson's dolphin (*Cephalorhynchus commersonii*) dataset (CMDO-0), and the goose-beaked whale (*Ziphius cavirostris*) dataset (CUBW-0), to name a few (Figure 4). Exceptions to this pattern included datasets with many query images, such as that of dusky dolphin (*Lagenorhynchus obscurus*) (DUDO-0) and the global humpback whale (*Megaptera novaeangliae*) dataset (HUWH-1). In general, the estimates of cost varied widely across datasets (Appendix S1). Two datasets, the Quebec beluga (*Delphinapterus leucas*) (BELU-1) and HUWH-1, were much more costly than the others. They both contained many more images than the others. As expected, the cost function was flatter between  $a = 5$  and  $a = 25$  for datasets with higher MAP.

The  $N_{\min}$  error tended to decrease as  $a$  increased (Figures 3, 4, & Appendix S3.4). For some datasets, such as MHWH-0 and FIWH-0, this decrease was monotonic (Figure 3). This monotonic decline was present for many datasets where some degree of partial automation was optimal,  $a^* > 0$ , for example, for FIWH-0, BRWH-0, CBDO-1, HUWH-1, CMDO-1, CUBW-0, both blue whale (*Balaenoptera musculus*) datasets (BLWH-0 and BLWH-1), and the southern right whale (*Eubalaena australis*) dataset (SRWH-0). For others, such as CBDO-0,  $N_{\min}$  error actually increased from  $a = 0$  to  $a = 5$ . True automation was optimal,  $a^* = 0$ , for several of these datasets, including Marianas short-finned pilot whale dataset (SFPW-1), both killer whale (*Orcinus orca*) datasets (KIWH-0 and KIWH-1), the gray whale (*Eschrichtius robustus*) dataset (GRWH-0), the sei whale (*Balaenoptera borealis*) dataset (SEWH-0), both beluga datasets (BELU-0 and BELU-1), and the Icelandic common minke whale (*Balaenoptera acutorostrata*) dataset (COMW-0). For some datasets, NE remained persistently high, even at  $a = 25$  (Figure 4 & Appendix S3.4).

The  $N_{\min}$  error closely tracked the false negative rate (Figure 3). At a more basic level, the false negative rate closely tracked RBIAS of the superpopulation size  $N$ . The linear model with the false negative and false positive rates predicting  $\text{RBIAS}(N)$  had an  $R^2 = 0.98$ , with  $\hat{\beta}_{\text{FN}} = 2.56$  and  $\hat{\beta}_{\text{FP}} = 0.236$  (Appendix S3.2). A  $\hat{\beta}_{\text{FN}}$  of 2.56 implied that increasing the false negative rate by 0.02 roughly corresponded to increasing the relative bias of  $N$  by 0.05. As such, the patterns observed for  $\text{NE}_{i,a}$  were mirrored by the false negative rate. For some datasets, the false negative rate decreased monotonically as  $a$  increased (Appendix S3.5). For others, the false negative rate increased from  $a = 0$  to  $a = 5$  and then decreased monotonically from  $a = 5$  to  $a = 25$ . For some datasets, the false negative rate remained persistently high, even at  $a = 25$ . The false positive rate fell to zero for most datasets at  $a = 5$ . At  $a = 0$ , the



**FIGURE 3** The false negative rate,  $N_{\min}$  error, and cost as a function of each strategy for generating capture histories with an individual identification algorithm. Three example datasets are represented: melon-headed whale (MHWH-0), common bottlenose dolphin (CBDO-0), and fin whale (FIWH-0) ( $a^*$ , optimal strategy [lowest combined  $N_{\min}$  error and cost  $C_{i,a}$ ]; MAP, measure of matching performance of the algorithm; gray numbers on  $y$ -axis, scaled cost; bar colors, optimal strategies as seen in Figure 4). Animal illustrations courtesy of NOAA Fisheries.

median false negative rate was 0.05 and the median false positive rate was 0.09 (Appendix S3.6).

Of the 4 Jolly–Seber parameters, superpopulation size  $N$  was most sensitive to misidentifications, whereas apparent survival  $\phi$  was the least sensitive (Figure 5 & Appendices S3.2 & S3.7). The linear model with the false negative and false positive rates predicting  $\text{RBIAS}(\phi)$  had an  $R^2 = 0.90$ , with  $\hat{\beta}_{\text{FN}} = -0.263$  and  $\hat{\beta}_{\text{FP}} = 0.194$  (Appendix S3.2). The  $\hat{\beta}_{\text{FN}}$  for this linear model was an order of magnitude smaller than that of the  $N$  model. Similarly, the signs of the coefficients indicate that the positive bias in  $\phi$  grows with an increasing false positive rate, whereas the negative bias in  $\phi$  grows with an increasing false negative rate. All told, the strategies had less of an influence, in absolute terms, on apparent survival  $\phi$  than the superpopulation size  $N$  (Figure 5 & Appendix S3.7).

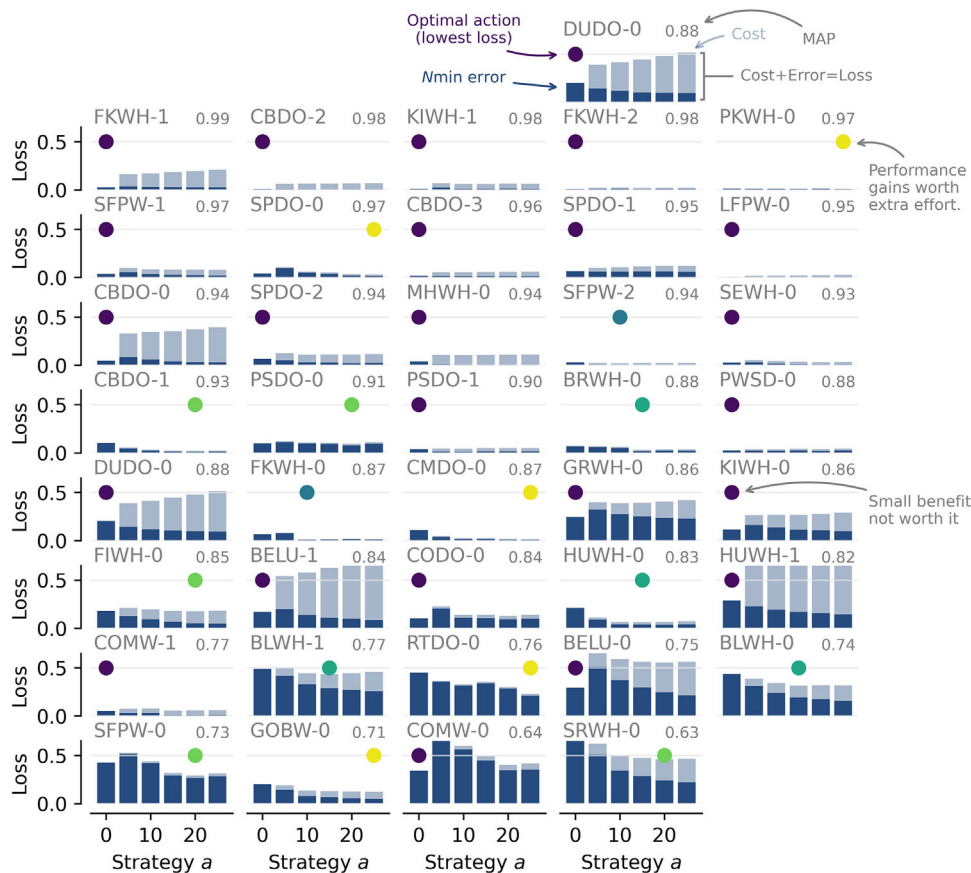
## DISCUSSION

In general, true automation was optimal ( $a^* = 0$ ) for datasets for which the ID algorithm matched images well in terms of MAP (Figure 4; also, MHWH-0 and CBDO-0 in Figure 3). Often, for these datasets,  $N_{\min}$  was estimated well under all strategies

(Appendix S3.4), meaning that cost dictated the decision process (Figure 4). Some form of partial automation was optimal (i.e.,  $a^* > 0$ ) for many lower performing datasets (Figure 4). In general, the estimate of the superpopulation size  $N$ , a component of  $N_{\min}$ , was more sensitive to misidentifications than the estimate of the apparent survival  $\phi$  (Figure 5). The relative bias in the superpopulation size  $N$  closely tracked the false negative rate (Figure 5), which tended to decline as the number of suggested matches inspected increased (Figure 3).

We recommend that agencies wishing to automate photo-ID for population assessments evaluate their matching algorithm beforehand. Our results suggest that a 2% increase in the false negative rate, all else being equal, translated to a 5% increase in the relative bias of the superpopulation size  $N$  (Appendix S3.8). Although the specific numbers associated with this increase, for example, the effect of the false negative rate FN on the relative bias of  $N$ ,  $\hat{\beta}_{\text{FN}}$ , were likely unique to our simulation setup and misidentification assumptions, the strong effect and strong correlation between FN and the relative bias in  $N$  reflected a well-known finding within the capture-recapture literature (Carlson et al., 1990; Hammond, 1986; Johansson et al., 2020; Lukacs & Burnham, 2005; Stevick et al., 2001; Yoshizaki, 2007). Many agencies may wish to cap the expected relative bias in





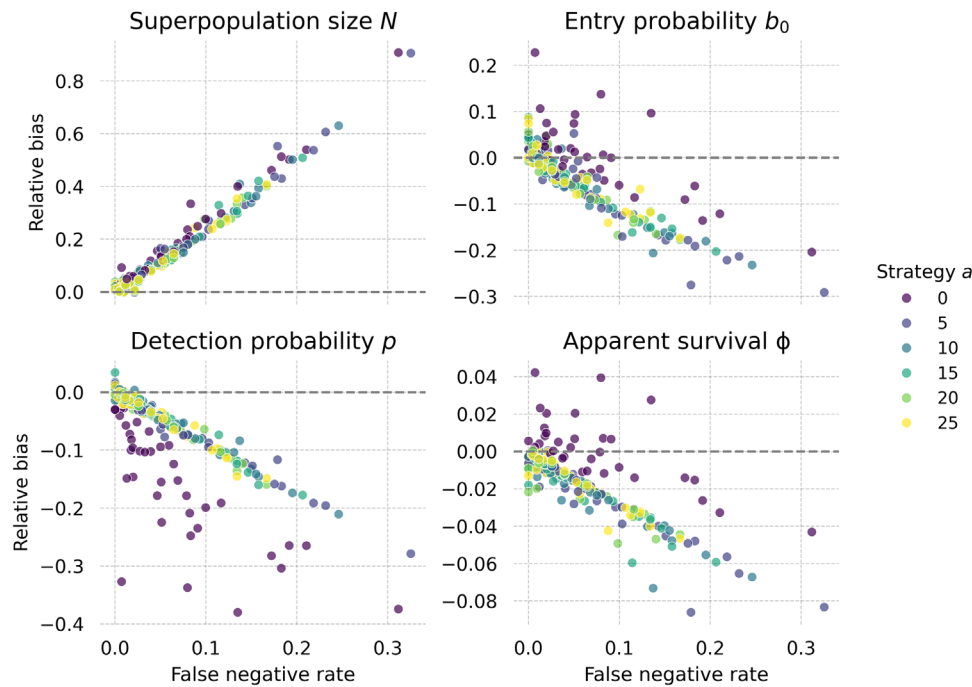
**FIGURE 4** Loss (i.e., combined estimation error and cost) for each cetacean photo identification dataset. Each bar represents a strategy for generating capture histories with an individual identification algorithm (dot near 0.5, optimal strategy [lowest combined cost  $C_{i,a}$  and  $N_{\min}$  error]; color gradient from dark to light corresponds to inspecting 0–25 suggested IDs from the algorithm; MAP, measure of matching performance for a dataset [top right on each graph]; graph upper limit, 0.65; datasets—BELU-1, HUWH-1, BELU-0, COMW-0, and SRWH-0 had strategies for which total loss exceeded 0.65; dataset abbreviations defined in Appendix S3.2).

$N$  to, say, 0.1, corresponding to a 10% overestimation of the true abundance  $N$ . As such, they would need to ensure that their matching procedure—be it a truly automated ID algorithm, a human matches images with a partially automated ID algorithm, or a human manually matching images—produces false negative matches no more than 4% of the time. To estimate this percentage, they could test their procedure on a representative sample of images and count the false negative matches. A representative sample might include the appropriate number of recaptured individuals to new individuals to the dataset or the appropriate spread of distinctiveness or image quality. The size of this sample would depend on the false negative rate and the desired precision of the estimate. For example, the sample would have to include around 250 images to reach a coefficient of variation (CV) of 0.3 for a false negative rate of 0.04, assuming the count of false negatives is a binomial random variable (Appendix S3.9).

Evaluating the matching algorithm by estimating the false negative and false positive rates would facilitate the use of our optimization tool without having to conduct a full simulation study. Appendix S2 shows 1 method for doing so for the Hawaiian pantropical spotted dolphin (*Stenella attenuata*) dataset (SPDO-1). The upshot is that, by evaluating the matching algo-

rithm, the user can estimate the false negative and false positive rates for each strategy and, in turn, estimate the relative bias of  $N$ . The accuracy of this estimate depends on the similarity of their case to our simulation setup and misidentification assumptions. From here, the user would be able to customize the optimization process to suit their research and policy goals. For example, say the user wishes to cap the relative bias in  $N$  to 10%. Further, assume that they have reasonable cost projections for a suite of strategies  $a$ , including true automation and a set of partially automated alternatives. Then, they simply need to find the cheapest strategy that produces a relative bias of  $N$  below 10%. That is, they would need to find the cheapest strategy that satisfies  $0.1 \leq -0.01 + 2.56 FN_a + 0.236 FP_a$ . Alternatively, they could cap the labor cost associated with matching at, say, 4000 h, and find the strategy with the lowest relative bias in  $N$  below budget. Finally, if they are worried that our simulation settings are too different from their proposed case, they could inform the optimization by simulating data and estimate the relative bias of  $N$ . Customizable simulation code for CJS and Jolly–Seber models and code for the optimization is available at <https://github.com/philpatton/autocapture>.

Although our analysis focused on cetacean datasets, our approach should apply to other photo-ID-based population



**FIGURE 5** The estimation error, expressed in terms of relative bias, for 4 parameters from a Jolly–Seber model as a function of the false negative rate (i.e., how often an individual was resighted but erroneously marked as a first capture) (dots, a cetacean photo identification dataset and a strategy for generating capture histories with an individual identification algorithm; strategy, inspecting 0, 5, 10, 15, 20, or 25 suggested identities from an individual identification algorithm). Entry probability, sometimes referred to as PENT, is a measure of recruitment from the POPAN formulation of the Jolly–Seber model. We held PENT constant across occasions in this paper; as such, only  $b_0$  is listed. A relative bias of  $-0.05$  suggests the parameter is underestimated by 5%.

assessments, such as camera trapping of terrestrial mammals. That said, we might expect the magnitude of  $\hat{\beta}_{\text{FN}}$  to differ, depending on the sampling design and the species being assessed. For example, Gardner et al. (2018) simulated open-population camera-trapping data with fewer sampling occasions,  $T = 5$ , and a lower apparent survival probability,  $\phi = 0.75$ , drawing inspiration from a tiger (*Panthera tigris*) camera-trapping dataset, which they also analyzed. Under these conditions, we would expect the magnitude of  $\hat{\beta}_{\text{FN}}$  to be lower because with lower survival and fewer occasions, there are fewer opportunities to misidentify an individual. As such, future users of our approach should carefully consider a realistic number of sampling occasions, apparent survival probability, and capture probability, all of which influence the magnitude of  $\hat{\beta}_{\text{FN}}$ . Patton et al. (2023) speculate that the same model structure as AnyDorsal could be adapted and retrained for multispecies ID of terrestrial mammals in camera-trapping studies. Should that effort be successful, researchers could use this framework to evaluate its performance in terms of population assessments.

Regardless of the taxa being studied, researchers should carefully consider the misidentification process for their system. Although our misidentification process should reasonably approximate many photo-ID studies, there are some limitations. For instance, we did not consider the situation where a researcher does not find a match in the suggested matches for the query image and does not mark the individual in the query image as a first capture. In a sense, the researcher is simply discarding the query image. Research organizations tend to do this when the query image is of poor quality (Rosel et al., 2011;

Urian et al., 2015) or there is only 1 image of the individual from the encounter (Morrison et al., 2011). As such, this strategy reduces misidentifications, mitigating the issue of ghost individuals, by adding a certainty threshold that must be crossed before adding a new individual to the dataset. Unfortunately, we did not have image quality scores or image counts per encounter for the 39 datasets. As a result, we may have overestimated the false negative rate for some datasets, particularly those not designed for capture-recapture (Rosel et al., 2011). Future users of our approach could tailor the misclassification process to explore the effect of discarding query images, especially if they have image quality scores for their images. Even so, Johansson et al. (2020) experimentally demonstrated that experienced biologists still misidentify individuals in high-quality images, creating ghost individuals.

Our simulation sheds light on 2 approaches—on opposite ends of the spectrum—to partially automated photo-ID. The first corresponds to  $a = 1$ , where the user only inspects the first suggested match. A user might take this approach with a high-performance ID algorithm that effectively separates individuals, for example, where the typical list of suggested matches  $[d_1, d_2, d_3]$  is something like  $[A, B, C]$  with corresponding similarity scores  $[0.81, 0.21, 0.19]$ , and the true identity in the query image is  $A$ . Occasionally, however, the ID algorithm might suggest  $d = [B, C, D]$  with the same similarity scores. In other words, the ID algorithm occasionally misjudges strong matches, strongly suggesting 2 distinct individuals are the same. The  $a = 1$  strategy seeks to efficiently catch these misidentifications, thereby minimizing false positives. This strategy, however, may

increase the false negative rate, which, in our simulation, was the more problematic misidentification rate. For several datasets, the false negative rate increased from  $a = 0$  to  $a = 5$  (Figure 3 & Appendix S2). Moreover, for every dataset, the false negative rate increased from  $a = 0$  to  $a = 1$  (this analysis required a slight modification to the classification scheme [Appendices S1 & S3.10]). The increase from  $a = 0$  to  $a = 1$  happened because some of the false positives that were prevented at  $a = 1$  became instead false negatives whenever the individual in the query image had already been identified. For the spotted dolphin dataset PSDO-1, which had an overall MAP of 0.90, the false negative rate at  $a = 0$  was 2.0%. At  $a = 1$ , the false negative rate was 6.5%, and at  $a = 5$ , it was 1.1%. This roughly translated to a drop in relative bias in  $N$  from 17% to 3% from  $a = 1$  to  $a = 5$  (Appendix S2). For the highest performing datasets,  $\text{MAP} > 0.95$ , the increase in the false negative rate from  $a = 0$  to  $a = 1$  was less pronounced, often around 2 percentage points (Appendix S3.10). As such, users may want to reserve the  $a = 1$  strategy for the highest performance ID algorithms and datasets. Nevertheless, any strategy should be vetted by first evaluating the matching algorithm on a sample of query images (Appendix S2). Further, any strategy should be designed to match the goals of the study.

The second strategy—at the other end of the spectrum—could be called  $a = \infty$ . The idea with this strategy is to continue inspecting suggested matches until a match is found or the end of the dataset is reached. This strategy, in effect, ensures that the false negative rate and false positive rate reach zero regardless of cost. As such, this may be a preferred strategy for resource-rich organizations, for datasets that lack a reliable ID algorithm, or for small datasets. Our results showed some evidence for this approach for certain datasets. For example,  $a = 25$ , the highest  $a$  for this study, was optimal for several of the smallest datasets. Continuing to the end of the dataset is essentially free, so why not? One can also evaluate this strategy in terms of the  $a$  value that would be necessary to reach a threshold value for the false negative rate, say, 4%. For 9 of the 13 lowest matching performance datasets,  $\text{MAP} < 0.85$ , the  $a$  value was over 50. For these datasets,  $a = \infty$  could be a worthy strategy. However, this approach might be overkill for better performing datasets. Twenty-one of the 26 highest performing datasets ( $\text{MAP} > 0.85$ ) reached the false negative rate of 4% by  $a = 10$ .

It is difficult to identify the species and dataset characteristics that would predict  $a^*$  for the ID algorithm we used in this study. Patton et al. (2023) identified several correlates with the overall MAP performance, which in turn correlated with our measures of estimation error, albeit imperfectly (Figures 3, 4, & Appendix S1). In fact, the third worst performing dataset in terms of MAP was the goose-beaked whale dataset CUBW-0, yet the abundance estimator for this dataset performed admirably under several scenarios, reaching a relative bias of 7.6% by  $a = 25$ . This was roughly the same relative bias as the PSDO-1 dataset at  $a = 25$  ( $RBLAS = 7.8\%$ ,  $\text{MAP} = 0.95$ ). These exceptions aside, the MAP value did roughly track estimation error. As such, the recommendations from Patton et al. (2023) would likely translate to our study. Nevertheless, we recommend that users evaluate their matching procedure beforehand.

One limitation of our study is that we did not explore the use of capture-recapture models that explicitly account for misidentification to estimate abundance (e.g., Link et al., 2010) or survival (Morrison et al., 2011). For some datasets, such as DUDO-0, GRWH-0, and KIWH-0, this led to highly biased estimates of  $N_{\min}$  with little variance, ensuring that the credible intervals did not contain the true value of  $N$  (Appendix S3.4). These strongly erroneous estimates may cause concern, because  $N_{\min}$  is a critical component of US stock assessments. We attribute this highly biased estimates to the persistently high rate of false negatives in the data, even for strategies with high  $a$ . In practice, many stock assessment scientists would either attempt to minimize the misidentification rates, for example, by increasing  $a$  or explicitly modeling the misidentifications. There are several of these misidentification capture-recapture models available to practitioners, depending on the analysis objective. Appendix S3.3 lists several available misidentification models for open or closed models or for false positives or false negatives. In many cases, it is only possible to deal with 1 misidentification type or objective—abundance or survival—at a time. For example, we are unaware of any Jolly–Seber models, where the goal is estimating both survival and abundance, that incorporate misidentifications. That said, we would be curious to see whether the conditional model of Morrison et al. (2011), where the first capture of each individual is set to 0, would work in a Jolly–Seber context. As far as we know, only models for ghosts are available in Program Mark.

AI presents scientific and management agencies with both a challenge and an opportunity. Our study highlights the opportunity, namely, that when the algorithm matches images very well, agencies may be able to use AI to reduce labor effort while minimally increasing estimation error. However, our study also highlights the challenges. Marginal increases in the false negative match rate upwardly bias abundance estimates. Such overestimation could have grim conservation consequences because stock assessments can be the primary resource for decision-makers. As such, it is critical that agencies estimate the false negative rate for their algorithm under different strategies. This information, along with the framework presented here, should help managers decide on the best way to implement AI for population

## AFFILIATIONS

<sup>1</sup>Marine Mammal Research Program, Hawai'i Institute of Marine Biology, University of Hawai'i at Mānoa, Kāne'ohe, Hawai'i, USA

<sup>2</sup>NOAA Fisheries Pacific Islands Fisheries Science Center, Honolulu, Hawai'i, USA

<sup>3</sup>Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, North Carolina, USA

<sup>4</sup>Cascadia Research Collective, Olympia, Washington, USA

<sup>5</sup>Sarasota Dolphin Research Program, Brookfield Zoo Chicago, c/o Mote Marine Laboratory, Sarasota, Florida, USA

<sup>6</sup>Oceans Initiative, Seattle, Washington, USA

<sup>7</sup>Projeto Baleia à Vista (ProBaV), Ilhabela, Brazil

<sup>8</sup>Research Center in Húsavík, University of Iceland, Húsavík, Iceland

<sup>9</sup>Centro de Conservación Cetacea (CCC), Santiago, Chile

<sup>10</sup>School of Biological Sciences, University of Auckland - Waipapa Taumata Rau, Auckland, New Zealand

- <sup>11</sup>Tethys Research Institute, Milano, Italy
- <sup>12</sup>The Swire Institute of Marine Science, The University of Hong Kong, Hong Kong, China
- <sup>13</sup>School of Biological Sciences, University of Aberdeen, Cromarty, UK
- <sup>14</sup>Marine Ecological Research Centre, Southern Cross University, Lismore, New South Wales, Australia
- <sup>15</sup>Happywhale.com, Santa Cruz, California, USA
- <sup>16</sup>Pacific Whale Foundation, Wailuku, Hawai'i, USA
- <sup>17</sup>SR3, SeaLife Response, Rehabilitation and Research, Des Moines, Iowa, USA
- <sup>18</sup>Marine Ecology and Telemetry Research, Seabeck, Washington, USA
- <sup>19</sup>The Oceania Project, Hervey Bay, Queensland, Australia
- <sup>20</sup>Pacific Whale Foundation, Santiago, Chile
- <sup>21</sup>Morigenos - Slovenian Marine Mammal Society, Piran, Slovenia
- <sup>22</sup>Sea Mammal Research Unit, Scottish Oceans Institute, University of St Andrews, St Andrews, UK
- <sup>23</sup>Cooperative Institute for Marine and Atmospheric Research, Research Corporation of the University of Hawai'i, Honolulu, Hawai'i, USA
- <sup>24</sup>Marine Science Department, Te Tari Putaiao Taimoana, University of Otago, Otago, New Zealand
- <sup>25</sup>The Cook Inlet Beluga Whale Photo-ID Project, Anchorage, Alaska, USA
- <sup>26</sup>School of Biological Sciences, Te Kura Mātauranga Koiora, University of Auckland, Auckland, New Zealand
- <sup>27</sup>Groupe de Recherche et D'éducation sur les Mammifères Marins (GREMM), Tadoussac, Quebec, Canada
- <sup>28</sup>Archipelagos Institute of Marine Conservation, Samos Island, Greece
- <sup>29</sup>Department of Life Sciences, Texas A&M University-Corpus Christi, Corpus Christi, Texas, USA
- <sup>30</sup>Department of Natural Sciences, University of Alaska Southeast, Juneau, Alaska, USA
- <sup>31</sup>L'association Evasion Tropicale, Bouillante, France
- <sup>32</sup>Departamento de Ciências Biológicas, Escola Nacional de Saúde Pública/Fiocruz, Rio de Janeiro, Brazil
- <sup>33</sup>Pacific Whale Foundation Australia, Urangan, Queensland, Australia
- <sup>34</sup>Marine Mammal Institute, Oregon State University, Newport, Oregon, USA
- <sup>35</sup>Bay Cetology, Alert Bay, British Columbia, Canada
- <sup>36</sup>Falklands Conservation, Stanley, Falkland Islands
- <sup>37</sup>Centre for Marine Science and Technology, Curtin University, Bentley, Western Australia, Australia
- <sup>38</sup>Project ORCA, Perth, Western Australia, Australia
- <sup>39</sup>Far Out Ocean Research Collective, Paihia, New Zealand
- <sup>40</sup>Zoophysiology, Department of Bioscience, Aarhus University, Aarhus, Denmark assessments.

## ACKNOWLEDGMENTS

We thank everyone who contributed to collecting and processing the photo-ID data that was instrumental to parameterizing this study. The graduate assistantship for P.T.P. was funded by the NOAA Fisheries Quantitate Ecology and Socioeconomics Training (QUEST) Fellowship via the Cooperative Ecosystem Studies Unit (CESU) award NA19NMF4720181. We also thank the University of Hawai'i Information Technology Services—Cyberinfrastructure, funded in part by the National Science Foundation CC\* awards 2201428 and 2232862, for computing resources. This paper represents SOEST contribution number 11879 and HIMB contribution number 1981.

## ORCID

Philipp T. Patton  <https://orcid.org/0000-0003-2059-4355>  
 Emma L. Carroll  <https://orcid.org/0000-0003-3193-7288>  
 Enrico Corsi  <https://orcid.org/0000-0001-7655-8754>  
 Claire Lacey  <https://orcid.org/0000-0003-0541-8193>  
 Kymberly M. Yano  <https://orcid.org/0000-0003-2054-3080>

## REFERENCES

- Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesebeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., & Zinkov, R. (2023). PyMC: A modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, *9*, e1516.
- Ashe, E., & Hammond, P. S. (2022). Effect of matching uncertainty on population parameter estimation in mark-recapture analysis of photo-identification data. *Mammalian Biology*, *102*(3), 781–792.
- Authier, M., Commanducci, F. D., Genov, T., Holcer, D., Ridoux, V., Salivas, M., Santos, M. B., & Spitz, J. (2017). Cetacean conservation in the Mediterranean and Black Seas: Fostering transboundary collaboration through the European Marine Strategy Framework Directive. *Marine Policy*, *82*, 98–103.
- Barlow, J., Calambokidis, J., Falcone, E. A., Baker, C. S., Burdin, A. M., Clapham, P. J., Ford, J. K. B., Gabriele, C. M., LeDuc, R., Mattila, D. K., Quinn II, T. J., Rojas-Bracho, L., Straley, J. M., Taylor, B. L., Urbán R., J., Wade, P., Weller, D., Witteveen, B. H., & Yamaguchi, M. (2011). Humpback whale abundance in the North Pacific estimated by photographic capture-recapture with bias correction from simulation studies. *Marine Mammal Science*, *27*(4), 793–818.
- Bergler, C., Gebhard, A., Towers, J. R., Butyrev, L., Sutton, G. J., Shaw, T. J. H., Maier, A., & Nöth, E. (2021). FIN-PRINT: A fully-automated multi-stage deep-learning-based framework for the individual recognition of killer whales. *Scientific Reports*, *11*(1), 23480.
- Bettridge, S. (2023). *National Marine Fisheries Service Instruction 02-204-01*.
- Bonner, S. J., Schofield, M. R., Noren, P., & Price, S. J. (2016). Extending the latent multinomial model with complex error processes and dynamic Markov bases. *Annals of Applied Statistics*, *10*(1), 246–263.
- Braulik, G. T., Taylor, B. L., Minton, G., Notarbartolo di Sciara, G., Collins, T., Rojas-Bracho, L., Crespo, E. A., Ponnampalam, L. S., Double, M. C., & Reeves, R. R. (2023). Red-list status and extinction risk of the world's whales, dolphins, and porpoises. *Conservation Biology*, *37*(5), e14090.
- Carlson, C. A., Mayo, C. A., & Whitehead, H. (1990). Changes in the ventral fluke pattern of the humpback whale (*Megaptera novaeangliae*), and its effect on matching: Evaluation of its significance to photo-identification research. *Reports of the International Whaling Commission (Special Issue)*, *12*, 105–111.
- Cheeseman, T., Southerland, K., Park, J., Olio, M., Flynn, K., Calambokidis, J., Jones, L., Garrigue, C., Frisch Jordán, A., Howard, A., Reade, W., Neilson, J., Gabriele, C., & Clapham, P. (2021). Advanced image recognition: A fully automated, high-accuracy photo-identification matching system for humpback whales. *Mammalian Biology*, *102*(3), 1618–1476.
- Conroy, M. J., & Peterson, J. T. (2012). *Decision making in natural resource management: A structured, adaptive approach*. Wiley-Blackwell.
- Gardner, B., Sollmann, R., Kumar, N. S., Jathanna, D., & Karanth, K. U. (2018). State space and movement specification in open population spatial capture-recapture models. *Ecology and Evolution*, *8*(20), 10336–10344.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.
- Hammond, P. S. (1986). Estimating the size of naturally marked whale populations using capture-recapture techniques. *Reports of the International Whaling Commission*, *8*(Special Issue), 253–282.
- Hammond, P. S., Francis, T. B., Heinemann, D., Long, K. J., Moore, J. E., Punt, A. E., Reeves, R. R., Sepúlveda, M., Sigurðsson, G. M., Siple, M. C., Víkingsson, G., Wade, P. R., Williams, R., & Zerbini, A. N. (2021). Estimating the abundance of marine mammal populations. *Frontiers in Marine Science*, *8*, 735770.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

- Johansson, Ö., Samelius, G., Wikberg, E., Chapron, G., Mishra, C., & Low, M. (2020). Identification errors in camera-trap studies result in systematic population overestimation. *Scientific Reports*, *10*(1), 6393.
- Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, *52*(1/2), 225–247.
- Link, W. A., Yoshizaki, J., Bailey, L. L., & Pollock, K. H. (2010). Uncovering a latent multinomial: Analysis of mark–recapture data with misidentification. *Biometrics*, *66*(1), 178–185.
- Lukacs, P. M., & Burnham, K. P. (2005). Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. *Journal of Wildlife Management*, *69*(1), 396–403.
- Maglietta, R., Renò, V., Caccioppoli, R., Seller, E., Bellomo, S., Santacesaria, F. C., Colella, R., Cipriano, G., Stella, E., Hartman, K., Fanizza, C., Dimauro, G., & Carlucci, R. (2020). Convolutional neural networks for Risso's dolphins identification. *IEEE Access*, *8*, 80195–80206.
- McClintock, B. T., Bailey, L. L., Dreher, B. P., & Link, W. A. (2014). Probit models for capture–recapture data subject to imperfect detection, individual heterogeneity and misidentification. *Annals of Applied Statistics*, *8*(4), 2461–2484.
- McCrea, R. S., & Morgan, B. J. T. (2015). *Analysis of capture-recapture data*. Chapman & Hall.
- Miele, V., Dussert, G., Spataro, B., Chamailé-Jammes, S., Allainé, D., & Bonenfant, C. (2021). Revisiting animal photo-identification using deep metric learning and network analysis. *Methods in Ecology and Evolution*, *12*(5), 863–873.
- Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348.
- Moore, R. B. T., Urian, K. W., Allen, J. B., Cush, C., Parham, J. R., Blount, D., Holmberg, J., Thompson, J. W., & Wells, R. S. (2022). Rise of the machines: Best practices and experimental evaluation of computer-assisted dorsal fin image matching systems for bottlenose dolphins. *Frontiers in Marine Science*, *9*, 849813.
- Morrison, T. A., Yoshizaki, J., Nichols, J. D., & Bolger, D. T. (2011). Estimating survival in photographic capture–recapture studies: Overcoming misidentification error. *Methods in Ecology and Evolution*, *2*(5), 454–463.
- Patton, P. T., Cheeseman, T., Abe, K., Yamaguchi, T., Reade, W., Southerland, K., Howard, A., Oleson, E. M., Allen, J. B., Ashe, E., Athayde, A., Baird, R. W., Basran, C., Cabrera, E., Calambokidis, J., Cardoso, J., Carroll, E. L., Cesario, A., Cheney, B. J., ... Bejder, L. (2023). A deep learning approach to photo-identification demonstrates high performance on two dozen cetacean species. *Methods in Ecology and Evolution*, *14*(10), 2611–2625.
- Pease, B., Pacifici, K., & Collazo, J. A. (2021). Survey design optimization for monitoring wildlife communities in areas managed for federally endangered species. *Animal Conservation*, *24*(5), 756–769.
- Punt, A. E., Siple, M., Francis, T. B., Hammond, P. S., Heinemann, D., Long, K. J., Moore, J. E., Sepúlveda, M., Reeves, R. R., Sigurðsson, G. M., Víkingsson, G., Wade, P. R., Williams, R., & Zerbini, A. N. (2020). Robustness of potential biological removal to monitoring, environmental, and management uncertainties. *ICES Journal of Marine Science*, *77*(7–8), 2491–2507.
- Rakhimberdiev, E., Karagicheva, J., Saveliev, A., Loonstra, A. J., Verhoeven, M. A., Hooijmeijer, J. C., Schaub, M., & Piersma, T. (2022). Misidentification errors in reencounters result in biased estimates of survival probability from CJS models: Evidence and a solution using the robust design. *Methods in Ecology and Evolution*, *13*(5), 1106–1118.
- Rosel, P., Mullin, K., Garrison, L., Schwacke, L., Adams, J., Balmer, B., Conn, P., Conroy, M., Eguchi, T., Gorgone, A., Hohn, A., Mazzoil, M., Schwartz, C., Sinclair, C., Speakman, T., Urian, K., Vollmer, N., Wade, P., Wells, R., & Zolman, E. (2011). Photo-identification capture-mark-recapture techniques for estimating abundance of bay, sound and estuary populations of bottlenose dolphins along the US East coast and Gulf of Mexico: A workshop report.
- Royle, J. A., Nichols, J. D., Karanth, K. U., & Gopalaswamy, A. M. (2009). A hierarchical model for estimating density in camera-trap studies. *Journal of Applied Ecology*, *46*(1), 118–127.
- Sanderlin, J. S., Block, W. M., & Ganey, J. L. (2014). Optimizing study design for multi-species avian monitoring programmes. *Journal of Applied Ecology*, *51*(4), 860–870.
- Schneider, S., Taylor, G. W., Linquist, S., & Kremer, S. C. (2019). Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, *10*(4), 461–470.
- Schofield, M. R., & Bonner, S. J. (2015). Connecting the latent multinomial. *Biometrics*, *71*(4), 1070–1080.
- Schwarz, C. J., & Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, *52*(3), 860–873.
- Seber, G. A. (1965). A note on the multiple-recapture census. *Biometrika*, *52*(1/2), 249–259.
- Stevick, P. T., Palsboll, P. J., Smith, T. D., Bravington, M. V., & Hammond, P. S. (2001). Errors in identification using natural markings: Rates, sources, and effects on capture recapture estimates of abundance. *Canadian Journal of Fisheries and Aquatic Sciences*, *58*(9), 1861–1870.
- Tucker, A. M., McGowan, C. P., Robinson, R. A., Clark, J. A., Lyons, J. E., DeRose-Wilson, A., Du Feu, R., Austin, G. E., Atkinson, P. W., & Clark, N. A. (2019). Effects of individual misidentification on estimates of survival in long-term mark–resight studies. *Condor: Ornithological Applications*, *121*(1), 1–13.
- Tyne, J. A., Loneragan, N. R., Johnston, D. W., Pollock, K. H., Williams, R., & Bejder, L. (2016). Evaluating monitoring methods for cetaceans. *Biological Conservation*, *201*, 252–260.
- Urian, K., Gorgone, A., Read, A., Balmer, B., Wells, R. S., Berggren, P., Durban, J., Eguchi, T., Rayment, W., & Hammond, P. S. (2015). Recommendations for photo-identification methods used in capture-recapture models with cetaceans. *Marine Mammal Science*, *31*(1), 298–321.
- Van Cise, A. M., Baird, R. W., Harnish, A. E., Currie, J. J., Stack, S. H., Cullins, T., & Gorgone, A. M. (2021). Mark-recapture estimates suggest declines in abundance of common bottlenose dolphin stocks in the main Hawaiian Islands. *Endangered Species Research*, *45*, 37–53.
- Wade, P. R. (1998). Calculating limits to the allowable human-caused mortality of cetaceans and pinnipeds. *Marine Mammal Science*, *14*(1), 1–37.
- Williams, B. K., Nichols, J. D., & Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic Press.
- Yackulic, C. B., Dodrill, M., Dzul, M., Sanderlin, J. S., & Reid, J. A. (2020). A need for speed in Bayesian population models: A practical guide to marginalizing and recovering discrete latent states. *Ecological Applications*, *30*(5), e02112.
- Yoshizaki, J. (2007). Use of natural tags in closed population capture-recapture studies: Modeling misidentification. [PhD thesis]. North Carolina State University.
- Yoshizaki, J., Pollock, K. H., Brownie, C., & Webster, R. A. (2009). Modeling misidentification errors in capture–recapture studies using photographic identification of evolving marks. *Ecology*, *90*(1), 3–9.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Patton, P. T., Pacifici, K., Baird, R. W., Oleson, E. M., Allen, J. B., Ashe, E., Athayde, A., Basran, C. J., Cabrera, E., Calambokidis, J., Cardoso, J., Carroll, E. L., Cesario, A., Cheney, B. J., Cheeseman, T., Corsi, E., Currie, J. J., Durban, J. W., Falcone, E. A., ... Bejder, L. (2025). Optimizing automated photo identification for population assessments. *Conservation Biology*, e14436. <https://doi.org/10.1111/cobi.14436>