The effects of acoustic misclassification on cetacean species abundance estimation

Marjolaine Caillat[a)]
Sea Mammal Research Unit, Scottish Oceans Institute, St Andrews University, St Andrews KY16 8LB, United Kingdom

Len Thomas
Centre for Research into Ecological and Environmental Modelling, The Observatory Buchanan Gardens, University of St Andrews, St Andrews KY16 9LZ, United Kingdom

Douglas Gillespie
Sea Mammal Research Unit, Scottish Oceans Institute, St Andrews University, St Andrews KY16 8LB, United Kingdom

(Received 4 July 2012; revised 29 January 2013; accepted 11 February 2013)

To estimate the density or abundance of a cetacean species using acoustic detection data, it is necessary to correctly identify the species that are detected. Developing an automated species classifier with 100% correct classification rate for any species is likely to stay out of reach. It is therefore necessary to consider the effect of misidentified detections on the number of observed data and consequently on abundance or density estimation, and develop methods to cope with these misidentifications. If misclassification rates are known, it is possible to estimate the true numbers of detected calls without bias. However, misclassification and uncertainties in the level of misclassification increase the variance of the estimates. If the true numbers of calls from different species are similar, then a small amount of misclassification between species and a small amount of uncertainty around the classification probabilities does not have an overly detrimental effect on the overall variance. However, if there is a difference in the encounter rate between species calls and/or a large amount of uncertainty in misclassification rates, then the variance of the estimates becomes very large and this dramatically increases the variance of the final abundance estimate.
 [http://dx.doi.org/10.1121/1.4816569]

PACS number(s): 43.60.Bf [ZHM] Pages: 2469–2476

I. INTRODUCTION

Over the last two decades, researchers and managers have become increasingly aware of the advantages of using passive acoustic monitoring over visual cues to detect marine mammals. Many studies, in particular those processing large datasets from long-term fixed hydrophone deployments, rely on automatic detectors and species classifiers to decrease the time and cost of analysis.

The repertoire of vocalizations by marine mammals is large and highly variable across species. Some species, such as large whales, produce calls that are easily recognized by an experienced observer or by an automatic classifier. However, many of the delphinid species produce highly variable calls where the frequency range of the different species' vocalizations overlaps to a large degree. These sounds are more challenging to classify. Classification algorithms have been developed by a number of researchers to identify delphinid sounds (e.g., Datta and Sturtivant, 2002; Gillespie *et al.*, 2013; Oswald *et al.*, 2007). The rate of misclassification in these examples was determined by testing the classifiers on recordings of species whose identity had been determined visually. However, none of these classifiers are perfect, and there remains considerable misclassification between species.

In any management strategy, accurate and precise quantification of population size ("abundance") is crucial to develop appropriate management actions. A standard method for estimating abundance based on acoustic detections is cue counting, where the cues are the vocalizations detected (Marques *et al.*, 2009, 2011). The general formula to estimate a species' abundance from cues is given by

$$\hat{\mathrm{N}} = \frac{n(1-\hat{\mathrm{c}})}{a\hat{T}\hat{\mathrm{P}}\hat{\mathrm{r}}}A, \tag{1}$$

where n is the number of detected cues, $\hat{\mathrm{c}}$ is the estimated proportion of false positives detected (calls classified as the species of interest which originated from other species or other sources of noise), a is the area in which cues can be detected, $\hat{\mathrm{P}}$ is the estimated average probability of a cue being detected within this area during recording time T, $\hat{\mathrm{r}}$ is the estimated cue production rate and A is the total study area (Marques *et al.*, 2009). Apart from the fact that this formula requires knowledge of the cue production (i.e., vocalization) rate, which is unknown for many species, the abundance estimate in Eq. (1) only considers the presence of one species at a time in the area of interest.

In this paper, we only address the issue of determining the true number of calls $\hat{v}$, which in Eq. (1) is the term

a)Author to whom correspondence should be addressed. Electronic mail: mc326@st-andrews.ac.uk

$\hat{v} = n(1 - \hat{c})$. Marques *et al.* (2009) estimated the proportion of false positive detections, $\hat{c}$, by visually examining 30 periods of 10 min from 6 days of recordings, a process which relied heavily on a human operator being able to distinguish between the sounds of interest and a range of other sound sources.

If the main source of false positive detections is the presence of other species with similar vocalizations in the study area, then the rate of false positive detections will be strongly related to the relative call densities from the different species. For example, if we know that species A and B are often confused by the classifier, and that species B is much more common or more vocal than species A, then a high percentage of the detections attributed by the classifier to species A will in fact be false positives detections resulting from the presence of species B. If on the other hand, species B were extremely rare or very silent, then there would be few misclassifications assigned to species A from species B.

Since we are interested in estimating the density of multiple species within a given study area, it becomes necessary to replace the $(1 - \hat{c})$ term with the more general equation

$$\hat{\mathbf{v}} = M(\mathbf{n}), \tag{2}$$

where $\hat{\mathbf{v}}$ and $\mathbf{n}$ are now vectors representing the true numbers of calls and the numbers of calls counted for each species after misclassification, respectively, and M is a more general misclassification operator.

The level of misclassification between species can generally be described in terms of a confusion matrix (e.g., Oswald *et al.*, 2007; Gillespie *et al.*, 2013), which summarizes the probabilities for correct, false positive and false negative classifications of all species considered. The confusion matrix [Eq. (3)] is a square matrix of dimension $m \times m$ in which each element of the matrix p_{ij} is the probability of classifying species j (column) as species i (rows). In particular, the entries for $i = j$ represent the probabilities of correctly classifying a species (success) and the off-diagonals ($i \neq j$) are probabilities of incorrectly classifying species j as species i (failure). A small p_{ij}, $\forall\, I \neq j$, means a low misclassification rate of species j as species i while a large p_{ij}, $\forall\, i \neq j$, means a high misclassification rate. On the other hand, a small p_{ij}, $\forall\, i = j$, means a low correct classification rate of species j and vice versa for a high p_{ij}, $\forall\, i = j$. Hence, the confusion matrix is given as

$$\mathrm{C} = (p_{i,j})_{1 \le i,j \le m} \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mj} & \cdots & p_{mm} \end{pmatrix}, \tag{3}$$

where $\Sigma_{j\mathrm{p}_{ij}} = 1\ \forall\ 1 \le j \le m$.

The expected number of detected calls $E(\mathbf{n})$ for each species following misclassification is therefore given by

$$E(\mathbf{n}) = C\mathbf{v} \tag{4}$$

and it follows that the true number of detections for each species can be estimated using

$$\hat{\mathbf{v}} = C^{-1}\mathbf{n}, \tag{5}$$

where C^{-1} is the inverse of the confusion matrix C.

Species classification is a stochastic process where each classification may be considered as an independent random event. In addition, we cannot assume that the confusion matrix is known precisely since it is typically derived from a finite sample of real data. Gillespie *et al.* (2013) show uncertainties, expressed as a measure of standard deviation, ranging from 0.04 to 0.48 for the probabilities of a typical confusion matrix. The stochastic nature of the classification process combined with our imperfect knowledge of the confusion matrix add to the uncertainty of any estimate of the true number of detected cues ($\hat{\mathbf{v}}$) and consequently, to the uncertainty of estimated species abundance if misclassification is taken into account.

With this in mind, this paper presents the first statistical analysis of the effects of species misclassification in acoustic surveys. In particular, it examines the bias and precision of the estimates of the true number of detected calls from multiple species which arise from the stochastic nature of the confusion process, as well as the uncertainty within the confusion matrix. We achieved this by looking at hypothetical confusion matrices and simulated data.

After a brief description of the classification process in mathematical terms, which also serves as an introduction of notation, we begin by looking at a simple model containing only the stochasticity within the classification process. We

TABLE I. The five different confusion matrixes (a–e) used during the simulation studies. Confusion matrix a is the identity matrix (no misclassification), b and c both have a high correct classification rate, but differ in that the misclassification rates of b are equal between species, whereas they are different in c. Confusion matrices d and e both have low rates of correct classification and again differ in that misclassification is equal between species in d, but varies in e.

		(a) True species				(b) True species				(c) True species				(d) True species				(e) True species			
		SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Predicted species	SpA	1	0	0	0	0.85	0.05	0.05	0.05	0.85	0.08	0.02	0.01	0.52	0.16	0.16	0.16	0.52	0.04	0.20	0.20
	SpB	0	1	0	0	0.05	0.85	0.05	0.05	0.10	0.85	0.03	0.09	0.16	0.52	0.16	0.16	0.15	0.52	0.13	0.05
	SpC	0	0	1	0	0.05	0.05	0.85	0.05	0.03	0.05	0.85	0.05	0.16	0.16	0.52	0.16	0.10	0.14	0.52	0.23
	SpD	0	0	0	1	0.05	0.05	0.05	0.85	0.02	0.02	0.10	0.85	0.16	0.16	0.16	0.52	0.23	0.30	0.15	0.52
		Scenario x.a				Scenario x.b				Scenario x.c				Scenario x.d				Scenario x.e			

TABLE II. Summary of the scenarios tested in the simulation study: similar misclassification rates means that elements of the confusion matrix outside the diagonal are the same between species (scenarios x.b and scenarios x.d), whereas for different misclassifications rates, they are different between species (scenarios x.b and scenarios x.e).

		Balanced data	Unbalanced data
No misclassification		Scenario 1.a	Scenario 2.a
Low misclassification rates	Similar misclassification rates	Scenario 1.b	Scenario 2.b
	Different misclassification rates	Scenario 1.c	Scenario 2.c
High misclassification rates	Similar misclassification rates	Scenario 1.d	Scenario 2.d
	Different misclassification rates	Scenario 1.e	Scenario 2.e

then extend this analysis by incorporating uncertainty in the rates of misclassification.

II. THE CLASSIFICATION PROCESS

We assume that classification events are independent of each other. Thus the classification for each species j can be described as the outcome of a multinomial process, where the vector of probabilities of the corresponding multinomial distribution is given by the jth column of the confusion matrix.

The numbers of trials in these multinomial distributions are the true number of detections $\mathbf{v}$, i.e., v_j is the number of trials, or the true number of detections for species j.

The expected observed number of vocalizations of species i (n_i) is equal to the number of vocalizations of species i correctly classified as species i plus the false positive classifications when vocalizations of another species $j \neq i$ have been misclassified as species i,

$$E[n_i] = \overbrace{p_{ii}v_i}^{\text{Correct Classified}} + \sum_{j \neq i} \overbrace{p_{ij}v_j}^{\text{Misclassified species}} . \quad (6)$$

The following interpretation will be useful when simulations are considered later on: Since we have identified each column with the probability vector of a multinomial distribution, it follows from Eq. (6) that the observed data for species i (n_i) is the sum of the output values of the ith components of m multinomial distributions, i.e.,

$$n_i = \sum_{j=1}^{m} \text{Multi}(v_j, \mathbf{p}_{.j})[i] \quad (7)$$

with the number of trials being the true number of detections v_j and the multinomial probability for species j being the jth column $\mathbf{p}_{.j}$ of the confusion matrix, e.g., n_1 is the sum of the first realized values of m multinomial distributions.

III. METHODS

For this study, we have not considered the effects of animal encounter rate, which can be an important source of uncertainty on animal abundance estimates, but would detract from the primary purpose of this paper which is to examine the effects of misclassification. We therefore consider only the following two sources of uncertainty:

(1) The stochastic nature of the classification process.
(2) Uncertainty in our knowledge of the classifier performance (i.e., uncertainty on the values of the elements of the confusion matrix).

First, we only consider the stochastic nature of the classification process, by assuming that the confusion matrix is known (i.e., no uncertainty). In a second step, we include additional uncertainty in the values of the confusion matrix itself.

The bias and variance on our estimates of the true number of detected calls was assessed using five different confusion matrixes (Table I) with increasing levels of misclassification. These include the identity matrix (i.e., no misclassification) and four others containing both low and high rates of misclassification with the misclassification being either the same (scenarios b or d) or differing for each species (scenarios c or e).

For each confusion matrix we evaluated the bias and variance using both balanced data (i.e., same number of calls for each species, scenario 1) and unbalanced data (i.e., differing numbers of calls per species, scenario 2). All models were developed with four species. For balanced data, we assumed that the true number of calls was exactly 3000 for each species. For unbalanced data, we selected values of 8000, 3000, 950, and 50 calls, respectively. Thus the total number of calls is the same as the balanced data, but with a 160-fold difference in the number of vocalizations between the most and the least abundant species.

The ten different scenarios (five confusion matrixes with balanced and unbalanced data) are summarized in Table II.

TABLE III. Examples of Dirichlet α parameters used for species A for each scenario. For the remaining species α parameters were the same but in different order to match the confusion matrices.

α for:	Scx.a	Scx.b	Scx.c	Scx.d	Scx.e
Low uncertainty	100,0,0,0	85,5,5,5	85,10,3,2	52,16,16,16	52,15,10,23
High uncertainty	0.1,0,0,0	0.85,5,5,5	0.85,0.1,0.03,0.02	0.52,0.16,0.16,0.16	0.52,0.15,0.1,0.23

TABLE IV. Analytically derived mean expected values for the true number of calls, E[$\hat{\mathbf{v}}$], and coefficient of variation (CV, expressed as a percentage).

	Scenario 1 (balanced data)				Scenario 2 (unbalanced data)			
Confusion matrix	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
a	3000 (0%)	3000 (0%)	3000 (0%)	3000 (0%)	8000 (0%)	3000 (0%)	950 (0%)	50 (0%)
b	3000 (1.19%)	3000 (1.19%)	3000 (1.19%)	3000 (1.19%)	8000 (0.54%)	3000 (1.19%)	950 (3.34%)	50 (59.9%)
c	3000 (1.12%)	3000 (1.36%)	3000 (1.14%)	3000 (1.17%)	8000 (0.57%)	3000 (1.48%)	950 (2.91%)	50 (43.85%)
d	3000 (4.10%)	3000 (4.10%)	3000 (4.10%)	3000 (4.10%)	8000 (1.75%)	3000 (4.10%)	950 (12.13%)	50 (223.51%)
e	3000 (3.98%)	3000 (3.00%)	3000 (4.07%)	3000 (4.96%)	8000 (1.59%)	3000 (3.29%)	950 (10.66%)	50 (299.92%)

For the simple case, in which the variance within the values of the confusion matrix is assumed zero, we have derived an analytical solution for the bias and variance on the true number of detected calls (Appendix). However, when uncertainty is added to the confusion matrix, the analytical approach becomes more complex, so we also explore bias and variance through simulation. When variability in the values of the confusion matrix is added to the model, bias and precision are measured from simulation only.

For each simulation (b), the numbers of misclassified, or observed, calls $\mathbf{n}_b$ were generated from the sum of four multinomial distributions with parameters v_b representing the true number of calls and p's being the confusion matrix probabilities [Eq. (7)]. The estimated true number of calls $\hat{v}_b$ was then estimated by multiplying the inverse of the confusion matrix by the number of misclassified (observed) calls Eq. (8),

$$\hat{v}_b = C^{-1}\mathbf{n}_b. \tag{8}$$

For each scenario, this process was repeated 10 000 times and the mean [Eq. (A3) in Appendix] and variance [Eq. (A10) in Appendix] of the estimated $\hat{\mathbf{v}}$ calculated.

When uncertainty in the confusion matrix was considered, the columns p_j of the confusion matrix are considered to be realizations of a probability distribution. To meet the requirement that columns have to sum to 1, this distribution was chosen to be a Dirichlet. The Dirichlet distribution is a multivariate probability distribution parameterized by a vector α of positives reals, $p \sim \mathrm{Dir}(\alpha)$ where $\sum_{i=1}^{k} p_{ij} = 1$ (Gelman, 2004).

For each of the 10 000 simulation trials, new values for the confusion matrix probabilities p_{ij} were generated from a Dirichlet distribution; these were then used in the same multinomial misclassification process as for the simpler situation. The true number of calls $\hat{\mathbf{v}}$ was again estimated using the inverse of the mean of the confusion matrix used to simulate the observed data [Eq. (5)].

Simulations were run with two levels (low and high) of uncertainty on the confusion matrix. In both situations, the alpha parameters of the Dirichlet distribution were selected such that the means of the parameters were equal to the confusion matrix probabilities of the different scenarios (Table III). However to generate low uncertainty in the confusion matrix, the parameters were selected to have a variance equal to 0.01 on average. The parameters for the high uncertainty were selected to match a variance of 0.1 observed with real data of Gillespie *et al.* (2013).

IV. RESULTS

Through this study the variance was represented by the coefficient of variation (CV), which is the standard deviation of the estimate divided by the estimate, generally reported in percent. When uncertainties in the probabilities of the confusion matrix were not taken into account, the analytical approach (Appendix) demonstrated that the means of $\hat{\mathbf{v}}$ were an unbiased estimate of the truth ($\mathbf{n}$), (Table IV). The simulations verified this result (Table V); no significant difference between means and variances calculated analytically and estimated through simulation was observed.

As expected, without misclassification, the estimates were unbiased and precise (CV = 0). A decrease in the rate of correct classifications (scenarios b and c versus d and e) did not affect the $\hat{\mathbf{v}}$ estimate's means, but it did significantly increase the variance and so the CV of these estimates (Fig. 1).

Where there were different numbers of calls from the four species, we again obtained unbiased estimates of the true numbers of calls [Fig. 1(B)]. The CV on the estimates of numbers of the more common species dropped (due to lower variance coming from misclassifications of the rarer species) but the CV of the estimates of the numbers of rare species

TABLE V. Simulation result, without uncertainty in the confusion matrix, of mean expected values for the true number of calls E[$\hat{\mathbf{v}}$], and coefficient of variation (CV, expressed as a percentage).

	Scenario 1				Scenario 2			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Scenario x.a	3000 (0%)	3000 (0%)	3000 (0%)	3000 (0%)	8000 (0%)	3000 (0%)	950 (0%)	50 (0%)
Scenario x.b	2999.93 (1.18%)	3000.12 (1.18%)	3000.01 (1.19%)	2999.94 (1.18%)	8000.37 (0.55%)	2999.47 (1.19%)	950.14 (3.67%)	50.02 (59.89%)
Scenario x.c	3000.69 (1.12%)	2998.99 (1.36%)	3000.14 (1.15%)	3000.18 (1.17%)	7999.46 (0.56%)	3000.40 (1.49%)	949.95 (2.94%)	50.19 (43.7%)
Scenario x.d	2999.87 (4.09%)	3001.49 (4.14%)	2998.55 (4.08%)	3000.09 (4.12%)	8000.74 (1.75%)	3000.72 (4.08%)	949.64 (12.14%)	48.90 (229.82%)
Scenario x.e	2997.28 (4.03%)	3002.00 (2.98%)	3000.30 (4.07%)	3000.41 (4.92%)	7999.63 (1.59%)	3000.88 (3.27%)	948.58 (10.69%)	50.92 (295.94%)

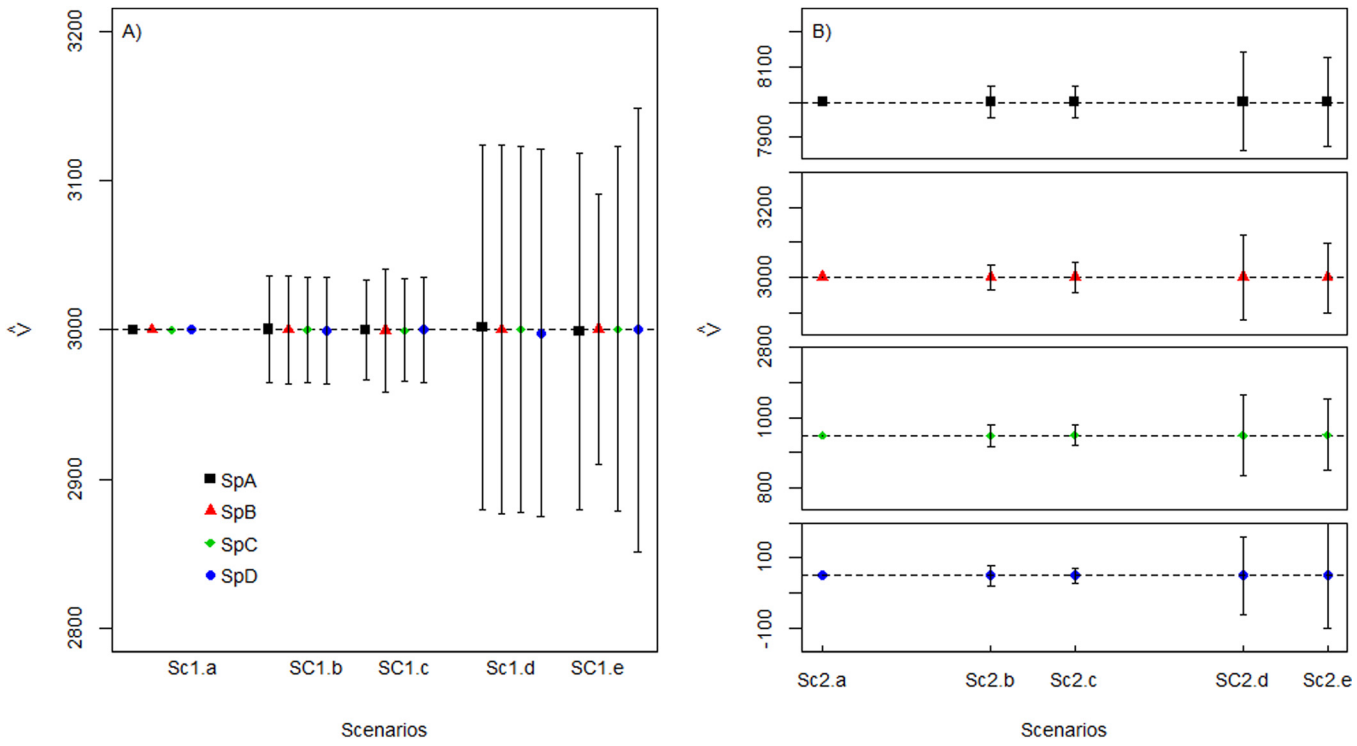

FIG. 1. (Color online) Expected true number of detections for each species, from simulation without uncertainty within the confusion matrix: (A) for balanced data scenarios Sc1a to Sc1e. (B) For unbalanced data scenarios Sc2a to Sc2b. Solid bars show the standard deviation and the dotted line the true number of detections.

calls rose significantly, reaching over 200% with confusion matrixes c and d (Fig. 2 and Table V).

When uncertainty in the confusion matrix was included, the simulations again showed unbiased estimation of $\hat{\boldsymbol{v}}$ for all the misclassification scenarios (Table VI and Table VII). However, adding uncertainty to the confusion matrix generated a large increase in the CV due to an increase of the variance (Fig. 3). With balanced data the CV, across all scenarios, increased on average from 2% without uncertainty to 11.7% with low uncertainty and to 87.7% with high uncertainty [Fig. 3(A)].

With the unbalanced data the average CV across all scenarios for the common species (species A and B) increased on average from 1.4% without uncertainty to 9% with low uncertainty to 68.6% with high uncertainty in the confusion matrix. For the rare species (species D) the average CV across the five scenarios was at 124.9% without uncertainty rising to 1009.3% with a low level of uncertainty and 7030.3% with a high level of uncertainty [Fig. 3(B)]. With the high variability in the confusion matrix some individual simulation results gave some negative estimates of $\hat{\boldsymbol{v}}$, which is clearly not possible with real data.

The presence of uncertainties in the confusion matrix did not alter the fact that a confusion matrix with low misclassification will give a more precise estimation of $\hat{\mathbf{v}}$ than a confusion matrix with a high misclassification rates (Tables VI and VII).

V. DISCUSSION

Our results show that it is possible to derive unbiased estimates the true number of detections of each species from data containing misclassified acoustic detections. However the precision of the estimates is strongly related to the degree of misclassification (Fig. 1) and the degree of uncertainty within the confusion matrix (Fig. 3).

A low CV (<10%) on the estimated numbers of calls can be achieved in some situations, such as when there are

TABLE VI. Simulation result, with a low level of uncertainty in the confusion matrix, of mean expected values for the true number of calls E[$\hat{\mathbf{v}}$], and coefficient of variation (CV, expressed as a percentage).

	Scenario 1				Scenario 2			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Sc x.a	3000 (0%)	3000 (0%)	3000 (0%)	3000 (0%)	8000 (0%)	3000 (0%)	950 (0%)	50 (0%)
Sc x.b	3000.11 (6.51%)	3000.58 (6.58%)	2999.38 (6.61%)	2999.92 (6.54%)	8000.48 (4.60%)	2999.24 (8.57%)	949.98 (24.85%)	50.30 (467.87%)
Sc x.c	2999.72 (6.68%)	2999.89 (6.54%)	3000.13 (6.57%)	3000.25 (6.61%)	7999.70 (4.60%)	3000.05 (8.58%)	950.17 (25.19%)	50.07 (471.00%)
Sc x.d	3002.12 (22.90%)	2996.36 (22.77%)	3001.90 (22.25%)	2999.35 (22.81%)	7998.41 (14.47%)	3000.28 (30.81%)	950.71 (92.89%)	50.60 (1722.71%)
Sc x.e	2999.25 (21.00%)	2999.79 (17.48%)	2999.06 (21.97%)	3001.90 (28.79%)	8001.65 (13.42%)	2999.24 (19.90%)	950.78 (105.79%)	48.3 2 (2578.82%)

TABLE VII. Simulation result, with a high level of uncertainty in the confusion matrix, of mean expected values for the true number of calls E[$\hat{\mathbf{v}}$], and coefficient of variation (CV, expressed as a percentage).

	Scenario 1				Scenario 2			
	SpA	SpB	SpC	SpD	SpA	SpB	SpC	SpD
Sc x.a	3000 (0%)	3000 (0%)	3000 (0%)	3000 (0%)	8000 (0%)	3000 (0%)	950 (0%)	50 (0%)
Sc x.b	2999.94 (61.89%)	3000.14 (61.28%)	3000.02 (62.65%)	2999.90 (61.70%)	8000.04 (42.64%)	3000.03 (80.85%)	949.97 (226.46%)	49.94 (4485.69%)
Sc x.c	2999.96 (62.53%)	3000.00 (60.69%)	3000.06 (65.51%)	2999.98 (62.27%)	7999.79 (44.42%)	3000.11 (83.19%)	950.01 (236.21%)	50.08 (4490.55%)
Sc x.d	3000.26 (214.59%)	2999.97 (217.66%)	2999.84 (212.69%)	2999.94 (218.69%)	8000.43 (101.44%)	2999.42 (214.96%)	949.61 (646.61%)	50.53 (12 788.65%)
Sc x.e	2999.66 (195.02%)	2999.97 (164.58%)	3000.15 (200.83%)	3000.22 (274.79%)	8000.13 (93.18%)	2999.67 (138.27%)	949.83 (751.36%)	50.37 (16 944.28%)

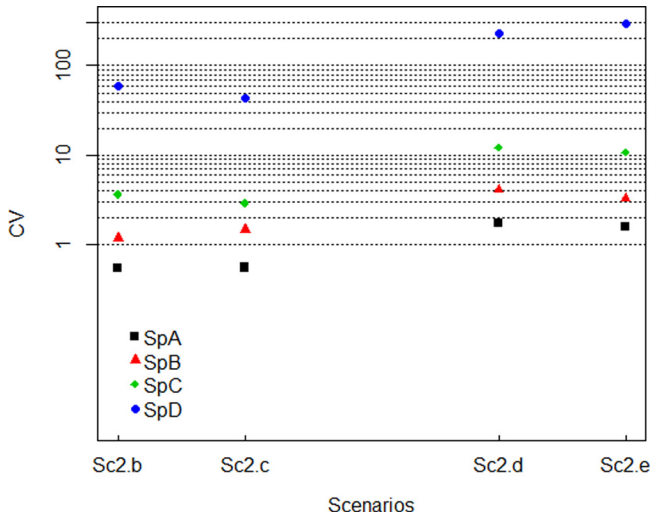

FIG. 2. (Color online) CV for unbalanced data for each scenario (Sc2b to Sc2e), with different misclassification rates. The *y* axis is on the log10 scale.

similar numbers of calls between species, a low misclassification rate, and low uncertainty within the confusion matrix. In cases where there are large differences in the numbers of detected calls between species (scenarios 2.x), the uncertainty is much higher on the estimates of the number of calls from the rarer species. In the more optimistic scenarios (low misclassification rate and low uncertainty within the confusion matrix), the CV for the common species A and B varied between 0.55% to almost 9%. However, the CV rises close to 100% for less common species (species C) in scenarios with a high rate of misclassification and low uncertainty for the values of the confusion matrix. For species with a very low encounter rate (Species D), even with a small level of uncertainty and low misclassification rate, the CV is higher than 400%, reaching the value of 2500% with a high misclassification rate. With uncertainties in the confusion matrix similar to those observed in real data (Gillespie *et al.*, 2013), the CV is higher than 50%, even for common species, and the estimate becomes totally uninformative for the rare species (CV $>$ 10 000%).

From our results it appears that uncertainty in the confusion matrix is the parameter responsible of most of the variance of the estimates. Indeed the average CV, across all species and all misclassification rates, is 70 times higher when a high level of uncertainty (average CV across 4 species = 1885) is assumed for the confusion matrix than where there is no uncertainty in the confusion matrix (average CV across 4 species = 27). Whereas the average variance, across all species and all levels of uncertainty within the confusion matrix, is only 29 times higher for models with a high misclassification rate (mean CV = 13 211) than for models with a low misclassification rate (mean CV = 450). A CV of 10% on a density estimate is considered as very good, a CV of 20% as reasonable and a CV of 100% near useless (Thomas and Marques, 2012). Particularly for rare species, CV's are often high, generally due to a low encounter rate. For example, Hammond *et al.* (2002) used visual line transect distance sampling methods to estimate the abundance of the relatively common European harbor porpoise, *Phocoena phocoena*,

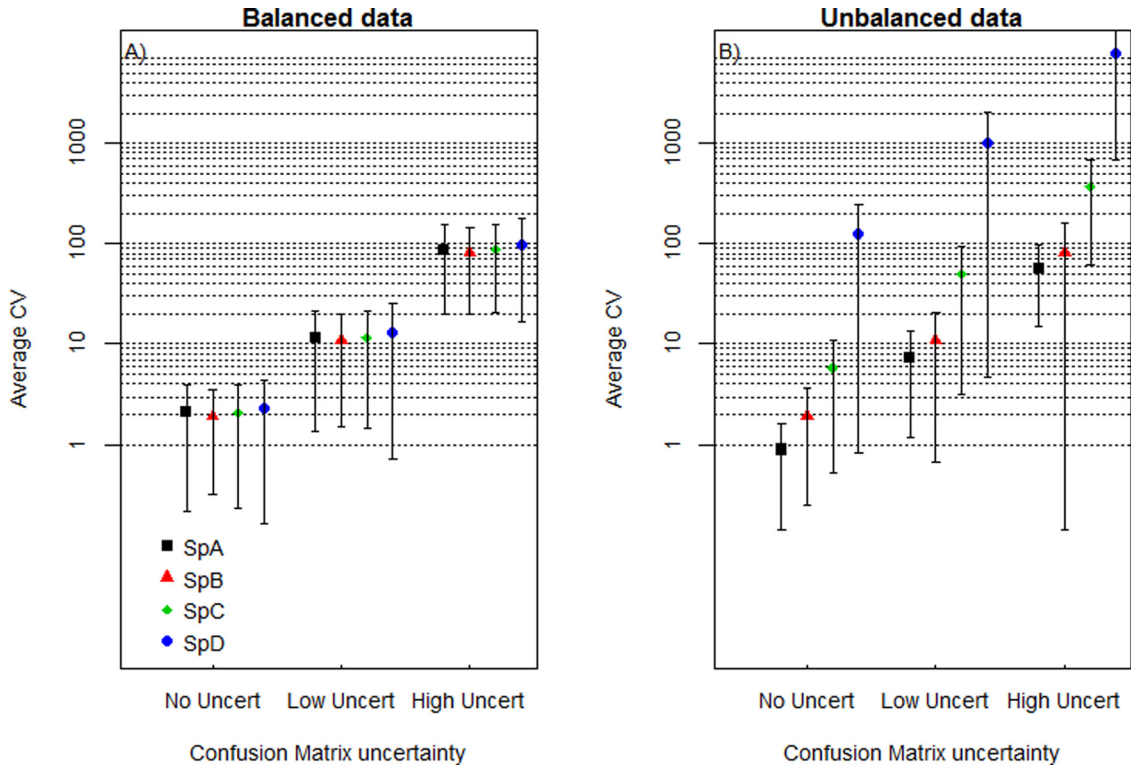

FIG. 3. (Color online) Mean CV across the five scenarios (A) Sc1a to Sc1e and (B) Sc2a to Sc2e) for each species and each level of uncertainty of the confusion matrix values: no uncertainty, low uncertainty and high uncertainty. The *y* axis is on the log10 scale.

with a CV of 14%, but the abundance of the rarer common dolphin *Delphinus delphis* from the same survey, had a CV of 67%. Gerrodette *et al*. (2011) estimated the abundance of the extremely rare Vaquita *Phocoena sinus* in the Gulf of California with a CV of 73%.

In this paper, we have only considered uncertainty in estimates of the true number of detections due to misclassification. In practice, however, significant contributions to the overall CV can be expected from the estimate of detection range, the encounter rate, and the estimate of vocalization (cue) rate which is unknown for many species. Thomas and Marques (2012) outline a number of methods for estimating both detection range and cue rate and the method chosen will be dependent on both the species and the study area.

Clearly the additional contributions to the overall CV of an acoustic abundance estimate from both misclassification and from uncertainty of the vocalization rate are important. However, acoustic survey methods using fixed sensors can often collect significantly more data than visual surveys, which will reduce the contribution to the CV from the encounter rate.

If we consider the species for which the true number of detection is estimated with a CV lower than 50% (for example, common species A and B), we can hope that, despite unavoidable misclassifications, acoustic detections provide useful information. However for the rare species, a small amount of misclassification from the more common species can render the acoustic data useless for all practical purposes.

Uncertainty on the values of the confusion matrix depends heavily on the amount and the quality of the available training data. The more data available to train the classifier, the more accurate are the statistics of the classified sounds used in the whistle classifier and the uncertainty on the values of the confusion matrix decreases. However, whistle classifiers should ideally be trained using visually confirmed data from the same study area since it is know that different sub-populations of a species may produce significantly different vocalizations (e.g., Rendell *et al*., 2006; Riesch *et al*., 2006; May-Collado and Wartzok, 2008; Janik, 2009). When developing a classifier for use in a particular study, there may therefore be a trade-off between the desire to acquire as much data as possible from multiple studies, possibly in different geographic areas and the desire to use a smaller amount of locally acquired data.

Being able to know the true number of detections from misidentified observed data is not a problem specific to the cue counting method discussed in this paper. In the case of estimating abundance of cetacean population using unidentified acoustic cues, the first question will always be about the true number of detections of each species, irrespective of the specific survey method applied. Thus, at its root, the problem considered here arises equally in any situation where it is known that there is misclassification between multiple species.

Since the uncertainty on the estimate of each species as highly dependent on the presence of other species, incorporating information on the likely abundance of calls from other species will hopefully lead to more robust estimates. We are therefore developing a Bayesian model which incorporates prior information on the relative abundance of calls from different species (based on previous survey work and

information on call rates) as well as the uncertainty on the values in the confusion matrix.

ACKNOWLEDGMENTS

We particularly thank Professor Peter Jupp for assistance deriving the analytical approach given in Appendix. This work was funded through the Natural Environment Research Council and SMRU Ltd.

APPENDIX: ANALYTIC ESTIMATE OF THE BIAS AND VARIANCE OF THE TRUE NUMBER OF DETECTED CALLS WHEN THERE IS NO UNCERTAINTY IN THE VALUES OF THE CONFUSION MATRIX

The notations used in this appendix are the same as the notations defined in the main body of the text.

The mean of a multinomially distributed random variable $y \sim \text{Multinom}(v,p)$ is (Royle and Dorazio, 2008).

$$E[y_j] = vp_j \tag{A1}$$

with v being the numbers of trials and p the event probabilities.

The expected value of a sum is equal to the sum of the expected values

$$E\left[\sum_{j=1} Y_j\right] = \sum_{j=1} E(Y_j). \tag{A2}$$

In the following, these two expressions [Eqs. (A1) and (A2)] are used to derive the expected values of $\hat{\boldsymbol{v}}$.

Our model can be described as

$$\begin{aligned} E[\hat{\boldsymbol{v}}] &= E[C^{-1}\,\mathbf{n}] \\ &= C^{-1}E[\mathbf{n}] \end{aligned} \tag{A3}$$

with $\mathbf{v}$ being the true number of detections, $\mathbf{C}$ being a constant confusion matrix and $\mathbf{n}$ the observed detections.

Since $\mathbf{n}$ is a sum of several multinomial elements [Eq. (7)] the latter is given by

$$\begin{aligned} &n_i = y_{i1} + y_{i2} + y_{i3} + y_{i4} \\ &\text{with } y_{\cdot i} \sim \text{Multinom}_j(v_j, \boldsymbol{p_j}), \\ &E[n_i] = \sum_{j=1}^{m} E(y_{ij}) = \sum_{j=1}^{m} v_j p_{ij}. \end{aligned} \tag{A4}$$

The variance and covariance of a multinomial distribution are (Royle and Dorazio 2008)

$$\text{Var}(y_j) = vp_j(1-p_j), \tag{A5}$$

$$\text{cov}(y_i, y_j) = -vp_i p_j. \tag{A6}$$

In general, the variance/covariance of a matrix multiplying an uncorrelated random variable $\mathbf{Z}$ is

$$\text{cov}(C\mathbf{Z}) = C\text{cov}(\mathbf{Z})C^T. \tag{A7}$$

With our model from Eq. (A7)

$$\text{cov}(\hat{\mathbf{v}}) = \text{cov}(C^{-1}\,\mathbf{n}) = c^{-1}\text{cov}(\mathbf{n})C^{-1^T}. \tag{A8}$$

Again identifying $\mathbf{n}$ as the sum of multinomial random variables, we have

$$\begin{aligned} &\text{cov}(\mathbf{n}) \\ &= \begin{bmatrix} \text{var}(n_i) & \cdots & \text{cov}(n_m, n_m) & \cdots & \text{cov}(n_1, n_m) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{cov}(n_i, n_1) & \cdots & \text{var}(n_j) & \cdots & \text{cov}(n_i, n_j) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{cov}(n_m, n_1) & \cdots & \text{cov}(n_m, n_j) & \cdots & \text{var}(n_m) \end{bmatrix} \end{aligned} \tag{A9}$$

with

$$\text{var}(n_i) = \sum_{j=1}^{m} \text{var}(y_{ij}) = \sum_{j=1}^{m} v_j p_{ij}(1-p_{ij}) \tag{A10}$$

and

$$\text{cov}(\mathbf{n_i n_k}) = \sum_{j} \text{cov}(y_{ij}, y_{kj}) = -\sum_{j} v_j p_{ij} p_{kj}. \tag{A11}$$

Datta, S., and Sturtivant, C. (**2002**). "Dolphin whistle classification for determining group identities," Signal Process. **82**, 251–258.

Gelman, A. (**2004**). *Bayesian Data Analysis* (Chapman and Hall/CRC, London), pp. 576–577.

Gerrodette, T., Taylor, B. L., Swift, R., Rankin, S., Jaramillo-Legorreta, A. M., and Rojas-Bracho, L. (**2011**). "A combined visual and acoustic estimate of 2008 abundance, and change in abundance since 1997, for the vaquita, Phocoena sinus," Mar. Mamm. Sci. **27**, E79–E100.

Gillespie, D., Caillat, M., Gordon, J., and White, P. R. (**2013**). "Automatic detection and classification of odontocete whistles," J. Acoust. Soc. Am. **134**, xxx–xxx.

Hammond, P. S., Berggren, P., Benke, H., Borchers, D. l., Collet, A., Heide-Jørgensen, M. p., Heimlich, S., Hiby, A. R., Leopold, M. F., and Øien, N. (**2002**). "Abundance of harbour porpoise and other cetaceans in the North Sea and adjacent waters," J. Appl. Ecol. **39**, 361–376.

Janik, V. M. (**2009**). "Acoustic communication in delphinids," Adv. Study Behav. **40**, 123–157.

Marques, T., Munger, L., Thomas, L., Wiggins, S., and Hildebrand, J. (**2011**). "Estimating North Pacific right whale *Eubalaena japonica* density using passive acoustic cue counting," Endangered Species Res. **13**, 163–172.

Marques, T. A., Thomas, L., Ward, J., DiMarzio, N., and Tyack, P. L. (**2009**). "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales," J. Acoust. Soc. Am. **125**, 1982–1994.

May-Collado, L. J., and Wartzok, D. (**2008**). "A comparison of bottlenose dolphin whistles in the Atlantic Ocean: Factors promoting whistle variation," J. Mammal. **89**, 1229–1240.

Oswald, J. N., Rankin, S., Barlow, J., and Lammers, M. O. (**2007**). "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **122**, 587–595.

Rendell, L. E., Matthews, J. N., Gill, A., Gordon, J. C. D., and Macdonald, D. W. (**2006**). "Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation," J. Zool. **249**, 403–410.

Riesch, R., Ford, J. K. B., and Thomsen, F. (**2006**). "Stability and group specificity of stereotyped whistles in resident killer whales, *Orcinus orca*, off British Columbia," Anim. Behav. **71**, 79–91.

Royle, J. A., and Dorazio, R. M. (**2008**). *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities* (Academic Press, Oxford, UK), pp 31.

Thomas, L., and Marques, T. A. (**2012**). "Passive acoustic monitoring for estimating animal density," Acoust. Today **8**, 35–44.