A sperm whale cautionary tale about estimating acoustic cue rates for deep divers

Tiago A. Marques,[1,2,a] Carolina S. Marques,[2] and Kalliopi C. Gkikopoulou[1,3]

[1] *Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews, KY16 9LZ, Scotland, +44 1334 461842, +44 1334 461800*

[2] *Centro de Estatística e Aplicações, Departamento de Biologia Animal, Faculdade de Ciências da Universidade de Lisboa, Portugal*

[3] *Sea Mammal Research Unit, Scottish Ocean Institute, University of St Andrews, St Andrews, Fife, KY17 9LZ, UK*

Passive acoustic density estimation has been gaining traction in recent years. Cue counting uses detected acoustic cues to estimate animal abundance. A cue rate, the number of acoustic cues produced per animal per unit time, is required to convert cue density into animal density. Cue rate information can be obtained from animal borne acoustic tags. For deep divers, like beaked whales, data have been analyzed considering deep dive cycles as a natural sampling unit, based on either weighted averages or generalized estimating equations. Using a sperm whale DTAG (sound-and-orientation recording tag) example we compare different approaches of estimating cue rate from acoustic tags, illustrating that both approaches used before, might introduce biases and suggest that the natural unit of analysis should be the whole duration of the tag itself.

[a] tiago.marques@st-andrews.ac.uk

Keywords: cue rate, sperm whale, foraging deep dive, density estimation, passive acoustics, DTAG

1. INTRODUCTION

Using the sounds produced by animals to estimate their abundance via passive acoustic monitoring (PAM) is becoming increasingly popular for species that are otherwise hard to detect visually. A prime example of such species are deep diving cetaceans which spend prolonged periods at depth, making them hard to survey visually. Therefore, it is not surprising that beaked whales Ziphiidae and sperm whales *Physeter macrocephalus* densities have been estimated using PAM methods (e.g. Barlow & Taylor, 2006, Lewis et al. 2007, Marques et al. 2009, Ward et al. 2012).

Cue counting is an indirect PAM approach to estimate animal abundance, where instead of counting the animals themselves, we count cues produced by the animals. For PAM the cues are sounds of the species of interest, in the sperm whale case usually those will be regular echolocation clicks. Cue counting was originally developed in the 1980s within the realms of the IWC (International Whaling Commission) for estimating baleen whale abundance from whale blows (e.g. Hiby & Ward, 1996). If a cue is effectively instantaneous - as for whale blows, a short duration sound or the onset of a long sound - then the only required multiplier to convert an estimate of the density of cues into a density of animals is the cue production rate. We define the cue production rate, or cue rate for short, as the mean number of cues produced per animal per unit time. Naturally, one would like to obtain a cue rate that is valid for the time when and place where the main survey occurs (e.g. Marques et al. 2013). Otherwise, mismatches might potentially induce a bias in cue rate, which will translate into a corresponding bias in the estimated animal density.

Cue production rates reported in the literature for deep divers, namely beaked whales, have considered deep dive cycles (DDCs) as a natural unit for analysis. A DDC can be defined as the period corresponding from the time an animal starts the descent for a deep foraging dive until the time it does the same for the next DDC. DDCs are highlighted in an example sperm whale sound-and-movement tag data from DTAG deployments (Johnson & Tyack, 2003) in Fig. 1. DDCs might be more or less clearly defined units depending on a given species' behaviour. Marques et al. (2009), considering Blainville's beaked whales (*Mesoplodon densirostris*), estimated the cue rate from a weighted average of the number of cues per unit time per DDC, where the weights were the durations of the DDCs. Warren et al. (2017), working with DTAG data from both Blainville's beaked whales and Cuvier's beaked whales (*Ziphius cavirostris*), considered a generalized estimation equation (GEE) approach, using as response variable the number of clicks per DDC, with DDC duration as an offset. Note this offset is equivalent to weighting by DDC duration.

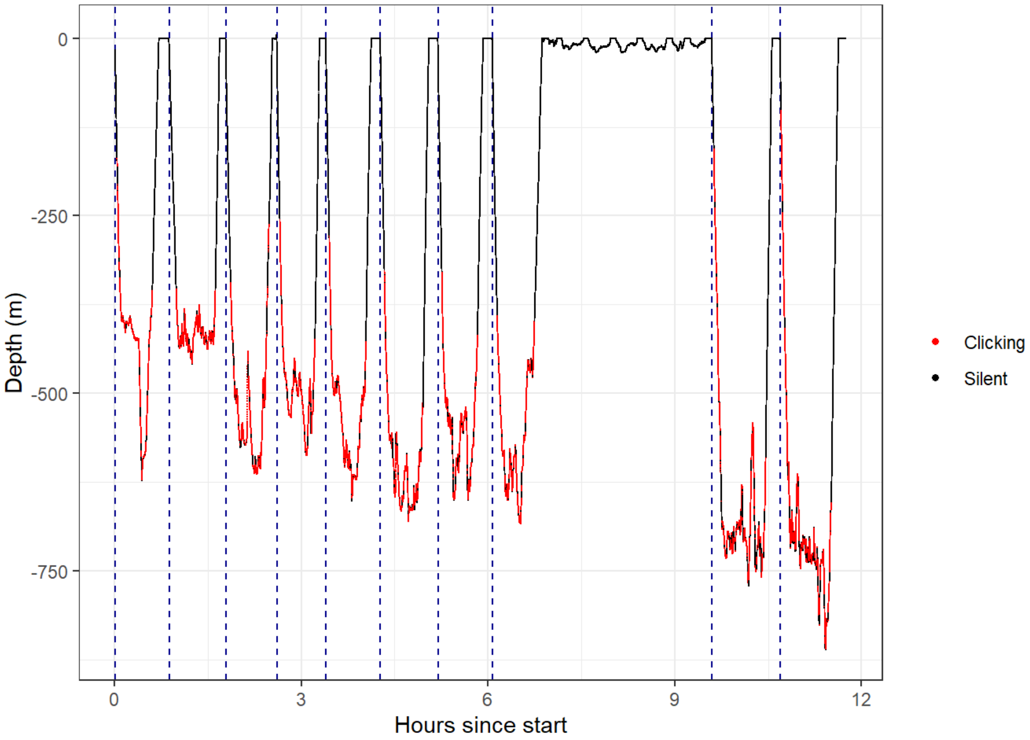

Fig. 1. An example sperm whale depth profile with the corresponding echolocation clicks overlaid. The 10 full DDCs available from this tag are highlighted. The eighth deep dive cycle is longer than the others, with the whale taking its time at the surface before submerging into the nineth deep dive.

We hypothesize that using the DDC as the unit for analysis when estimating cue rates might lead to biased inferences in the case where the DDC duration is correlated to the cue rate during a deep dive, which is likely the case by the very definition of a DDC. We compare different ways of calculating cue rates, and their associated precisions, from tag data, including averages and weighted averages, per DDC and per tag, and contrast these with regression modelling approaches to evaluate if previously used GEEs perform as expected. We illustrate the potential issues with a unique dataset of 104 sperm whale DTAGs and provide practical guidance for estimating acoustic cue rates from animal borne tags.

2. METHODS

We first describe the DTAG dataset used, then the methods used to extract the individual echolocation regular foraging clicks, considered the acoustic cue of interest, and finally the different analysis options to estimate cue rates from the acoustic data.

A. Data collection

We consider a sample of 104 DTAGs deployed on sperm whales, at 8 different sites and covering 13 different years. For additional details including counts of tags per year-site combination see the Supplementary Material. We focus on estimating a pooled cue rate for the species assuming the sample of tags would be representative for that purpose. We address potential issues in doing so in the discussion.

The DDCs were defined as periods starting at the moment the whale submerges into a deep foraging dive all the way till the next time it submerges for the subsequent deep foraging dive. For simplicity we considered deep foraging dives to be dives deeper than 100 meters. We note explicitly that a deep dive foraging cycle might include portions of time where the whale is at the surface not engaged in any deep foraging behaviour, say while resting or socializing at the surface (cf. eighth DDC in Fig. 1). This will be one of the reasons for why care must be taken when using such DDCs as sampling units.

B. Data processing

For each tag, the sound files were processed to extract the times of emission for each regular echolocation click emitted by the tagged whale. Depending on the tags, custom built MATLAB functions to do so included either findclicks or findallclicks and findmissedclicks functions. These tools have been developed by Mark Johnson (freely available at: https://soundtags.wp.st-andrews.ac.uk/dtags/dtag-3/). For easier reference the functions are hosted also at https://github.com/TiagoAMarques/DeepDiverCueRates (folder: *click_extraction_matlab*).

To identify the timing of regular clicks from the tagged whale in the on-animal DTAG sound recording, a semi-supervised click detector was used, as described in Johnson et al. (2006).. The sound files were processed sequentially in 15 second chunks through a supervised click detector to identify likely clicks from the tagged animal, using a 4-pole Butterworth band-pass filter (with cut-off frequencies at 3 and 20 kHz) and a level threshold based on the envelope of the click. An experienced analysist went through manual inspection of spectrograms (512 sample fast Fourier transform (FFT) with a Hamming window and 50% overlap) formed by 15s intervals of the sound recording (as described in Warren et al., 2017), accepting/rejecting the click identifications or, if needed, manually adding potential missed clicks. Clicks from the tagged animal were distinguished from those of other nearby whales in two ways (Johnson et al., 2006). Clicks from the tagged whale have both low-frequency energy that is absent in clicks recorded from non-tagged whales (Johnson et al., 2009) as well as a consistent angle of arrival on the tag, θ , computed from $\theta=\sin-1(\tau c/d)$, where c is the speed of sound in seawater, d is the hydrophone separation (0.025m) and τ is the time delay between the two hydrophone signals, measured by cross-correlation. The arrival angle of clicks from the tagged whale, when corrected for the tag orientation on the whale, will be consistently close to zero, as the sound source from the tagged animal is directly in front of the tag, while those from other whales will vary widely as the focal and non-focal whales maneuver. The DTAG click extraction process, which is a semi-supervised click identification ensures that all clicks produced by the focal whale are identified.

For further analysis of our sperm whale dataset, we consider two datasets derived from the above procedure: (1) The dataset with all the DDCs and (2) a dataset corresponding to removing DDCs lasting over one hour. As an example, in Figure1 this would correspond to exclude the 8th DDC. We refer to these as the complete and the reduced datasets, respectively. These allow us to illustrate different aspects of the analysis, as if these were two different datasets with slightly different characteristics, the latter representing species with deep dive behavior similar to a sperm whale, but without the long resting periods at the surface.

C. Data analysis

We focus on the estimation of a pooled (pooling across both years and locations) cue production rate *r* for sperm whales, defined as the number of cues produced per time unit, per animal:

$$r = \frac{\#\, sounds}{time\ animal}. \qquad (1)$$

Note that in general we ignore, in wording and notation alike, the fact that this is a measure per animal. Here we consider sounds to be regular echolocation clicks and the time unit seconds, for convenience, but the above expression could be used for any arbitrary time period, meaning in particular it could be calculated by DDC, by tag or by any arbitrary time period (e.g. per 5 minutes). Then one can use those sampling units to average across a sample to get a mean cue rate for the population of interest, accounting for possible non-independence, as required. The population average cue rate, using a standard mean based on DDCs, is then estimated by

$$\hat{r}^{ds} = \frac{\sum_{j=1}^{n_d} r_j^d}{n_d} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n_{di}} \frac{c_{ik}}{t_{ik}}}{n_d}, \qquad (2)$$

where the superscripts *d* are used for deep *D*ive and *s* for *S*tandard mean, r_j^d represents the cue rate for the jth dive cycle (j=1,2,...,n$_d$), c_{ik} and t_{ik} represent respectively the number of clicks in, and the duration of, the kth DDC of the ith whale (k=1,2,...,n$_{di}$), and n$_{di}$ is the number of deep dives recorded for whale i. On the other hand, one could *W*eight (note superscript *w* below) for the DDC duration, as was done in Marques et al. (2009), leading to

$$\hat{r}^{dw} = \frac{\sum_{i=1}^{n_d}\sum_{k=1}^{n_{di}} \frac{c_{ik}}{t_{ik}} t_{ik}}{\sum_{i=1}^{n_d} t_{ik}} = \frac{\sum_{i=1}^{n_d} c_{ik}}{\sum_{i=1}^{n_d} t_{ik}}. \qquad (3)$$

Note that this otherwise apparently more complex estimator (than the standard average, given the weights) actually reduces to a simpler expression, the total number of detected cues, across all tags, divided by the total recording time, again across all tags. Both of the above consider the DDCs as the sampling unit, as has been done before in the literature. We note that, strictly speaking, that is the definition of pseudoreplication (Hurlbert, 1984), where the independence came in as a stated assumption, for a proof of concept of PAM density estimation (DE) in Marques et al. (2009). On the other hand, if we consider the *n* tags as the sampling units (i.e., the animals, superscript *a*), we have a standard average (superscript *s*) estimator as

$$\hat{r}^{as} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} r_i^a}{n} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{c_i}{t_i}}{n} \qquad (4)$$

where r_i^a represents the cue rate for the i^th^ whale (i=1,2,...,n), c_i and t_i represent respectively the number of clicks in, and the duration of, the i^th^ tag. The corresponding weighted average (superscript *w*) version, now weighting by tag duration, is

$$\hat{r}^{aw} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{c_i}{t_i} t_i}{\sum_{\mathrm{i=1}}^{\mathrm{n}} t_i} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} c_i}{\sum_{\mathrm{i=1}}^{\mathrm{n}} t_i}. \qquad (5)$$

For each of the approaches we also estimate the corresponding precision and 95% confidence intervals. The variance of a standard mean is straightforward and present in any introductory statistics book, and the variance for the weighted mean was calculated considering the methods proposed by Gatz & Smith (1995).

For a direct comparison with the methods from Warren et al. (2017), we also estimate the cue rate based on regression models. For these, the unit of analysis considered was always the DDC. We consider the same GEE inspired approach: modelling the number of clicks per DDC, using the DDC duration as an offset, considering tag as the grouping variable (akin to a random effect), with a log link function, an "independence" correlation matrix and robust standard errors. The GEE treats the deep dives as the sampling unit but inflates standard errors on the estimated cue rate via the correlation structure induced by the grouping variable. Therefore, in terms of how the correlation structure is dealt with, it represents a half-way house between treating deep dives or tags as the sampling unit, while considering a regression model instead of an analytical formula for the average. The intercept of this model corresponds to an estimate of the cue rate. To obtain the standard error on the response scale we used a delta method approximation (Powell, 2007). To obtain 95% confidence intervals we assumed normality on the link scale and back-transformed to the response scale. For comparison we also implement the same GEE analysis without the offset and without both offset and grouping variable, and use corresponding generalized linear mixed models (GLMM) with whale as a random effect, instead of the GEEs. When not considering the offset we modelled the rate directly and considered a Gamma distribution for the response. For 9 DDCs (corresponding to 2% of the DDCs) there were 0 clicks. To avoid issues with the Gamma not coping with the response variable being exactly zero we replaced these observations by 0.5 (or 1/20th of the observed minimum positive count of 10 clicks per DDC). The practical impact of this tweak is negligible, but fitting with the Gamma family becomes possible. Good reviews on GLMMs and GEEs in Ecology are Bolker et al. (2009) and Pekár & Brabec (2017), respectively.

Analysis was implemented using R (R Core Team, 2022). The GEE model was implemented using the geeglm function in the geepack package (Højsgaard et al. 2006) and the GLMM model via glmer in the R package lme4 (Bates et al. 2015). All of the code to reproduce the statistics and figures in the paper is provided as supplementary pdf file. This pdf is generated via an RMarkdown dynamic report. All the data and original .Rmd file that allows one to reproduce or update the analysis are shared as a github repository: https://github.com/TiagoAMarques/DeepDiverCueRates.

3. RESULTS

We had 826 DDCs across a total of 104 tag deployments from sperm whales, with a median number of 7 DDC per tag, ranging from 1 to 31 DDC per tag. Tag durations ranged from 0.3 to 25.8 hours. The observed cue rates per tag varied between 0.13 and 1.61 clicks per second, with a median value of 0.86, while in the case of DDC these ranged between 0 and 1.93, with a median value of 0.93. (Figure 2). In the case of our dataset, the cue rate per DDC tended to increase with DDC duration for the reduced dataset, but in the full dataset, the longer DDC presented long periods without vocalizations, and hence cue rates tended to decrease with DDC duration (Figure 3).

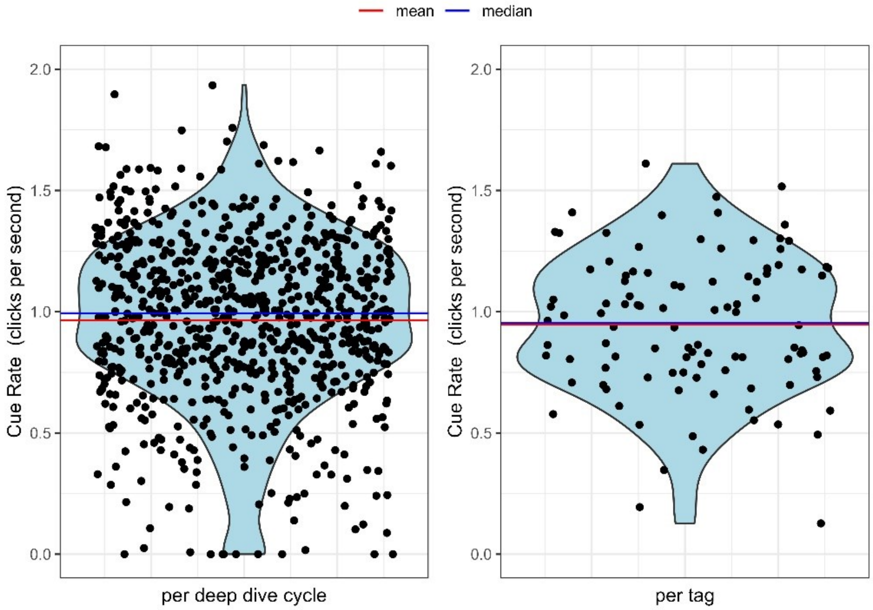

Fig. 2. Violin plots of the observed cue rates per deep dive cycle (left) and per tag (right). The x-coordinate is a non-interpretable jitter for improved data visualization. There are 826 points in the left plot corresponding to the cue rates for

all the deep dive cycles available for the 104 tag deployments on whales in the right plot.

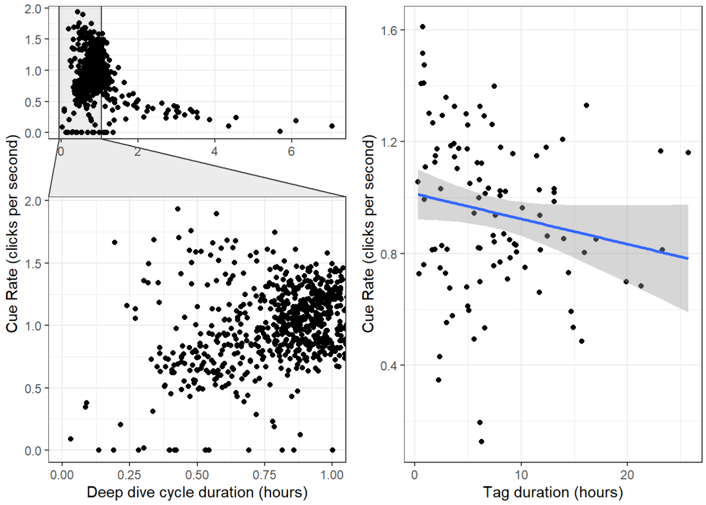

Fig. 3. Observed cue rates as a function of duration, both for deep dive cycles (DDCs, left) and tags (right). The left bottom panel zooms in on DDCs shorter than 1 hour. On the right plot a regression line is represented, with the grey lines representing the corresponding 95% confidence interval on the regression line.

The results obtained for each of the analytic estimators and the regression models are compared in Fig. 4. As expected, for any given method, the estimates based on the reduced dataset were always higher than those obtained for the complete dataset. This is a rather obvious consequence of the fact that, by construction, the longer DDCs tend to include periods when the animal spent considerable time at the surface, when they are mostly silent. The 8th DDC in Fig. 1 is a good example.

The sperm whale cue rate, considering an unweighted average based on the full dataset at the tag level, was 0.947 clicks per second, with a 95% confidence interval (CI) of 0.892,1.003. The same tag-level analysis, but weighted by tag duration, estimated a lower cue rate: 0.911 clicks per second (95% CI 0.854, 0.968). Corresponding unweighted estimates at the DDC level were lower in the reduced dataset (the difference in green values from the 2nd analysis to the 3rd analysis in Fig. 4), but higher for the complete dataset (the difference in blue values from the 2nd analysis to the 3rd analysis in Fig. 4). When weighed by DDC, the effect is opposite in the reduced and the complete datasets, with the cue rate increasing in the former, but decreasing in the latter. This is the same pattern observed when adding an offset to the GEE with whale as grouping variable.

The analyses weighted by duration, either at the tag or DDC level, lead necessarily to the same cue rate point estimate, with a marked difference in precision, naturally higher for the DDC analysis.

Similar estimates are obtained for the GEE and the corresponding GLMM counterparts.

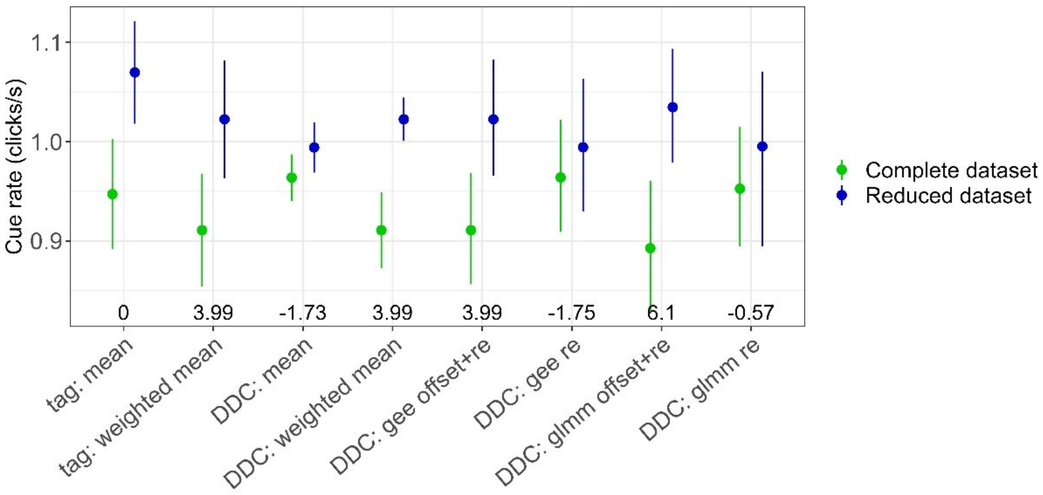

Fig. 4 Estimated cue rates, and corresponding 95% confidence intervals, considering the different approaches. Left to right, the first two analyses are at the tag level, the remaining ones at the deep dive cycle (DDC) level. For the regression analysis, offset means the duration of the DDC was included as an offset and "re" denotes if the model included whale as a grouping variable (GEE) or as a random effect (GLMM). The number shown for each approach corresponds to the bias in animal density that a biased cue rate would induce, assuming as the truth for cue rate the unweighted average at the tag level, unbiased by design.

4. DISCUSSION

We presented pooled cue rate estimates for sperm whales across all the tag recordings available to us, for the purpose of comparing several methods to estimate cue rates, including averages and weighted averages, and a variety of regression models. Variability over time and space means that such overall mean cue rates might not be representative of any one place at any one time, and therefore we recommend these values are not used to inform any sperm whale PAM density estimation exercises. This is the topic of a separate research thread we are currently pursuing.

To inform a PAM density survey, we are interested in estimating a cue rate for a population of whales. Therefore, the true variability that we are interested in is the variability across whales. For that reason alone, one should expect that approaches that consider the whale (i.e. tag record) as the sampling unit would be preferable. Perhaps surprisingly, that was not considered by earlier attempts (e.g. Marques et al. 2009, Warren et al. 2017). Here we consider the standard average at the tag level as a gold standard, since that should be unbiased by design.

A. To weight or not to weight: that is (and might remain) the question

The weighted average at the tag level was lower than the unweighted average, a reflection of the fact that longer tag records had slightly lower cue rates (cf. Fig. 3, right panel). There is considerable overlap in the 95% CI between the weighted and unweighted average at the tag level, and there is no obvious reason for why tags with longer duration (hence, everything else being the same, also on average with a larger number of DDCs) might have lower cue rates. We put forward a couple of possible explanations. The first is that deep diving itself could represent a behavior that promotes tag displacement and/or release. In such a case, animals spending longer times at the surface and hence for which the tag would be more likely to stay on longer would be oversampled, and hence the true cue rate would be biased low. However, this is unlikely, and in fact one might even argue the bias would be the other way around, with high pressure experienced by the tag during deep dives meaning the suction cups would be less prone to displace. The second is that animals available for tagging might be preferentially in a vocal mode (say because they are found acoustically), and hence they might be, more often than not, engaged in behaviors that have higher cue rates than the average animal in the population. On the other hand, Warren et al. (2020) noted, using a subset of the tags used here, an opposite pattern: if anything a time decaying effect detectable up to the 5th deep dive with less buzz sounds (but no results are reported for echolocation clicks, the focus here). There is no strong reason to prefer one estimate over the other; weighted averages can have higher precision if the weights are sensible. But if they are not, bias could creep in. Unless there is an alternative explanation that would not induce bias for why longer tag deployments might have lower cue rates, we suggest that an unweighted average at the tag level might be preferred. The weighted by DDC approach might introduce bias, or in this case perhaps differences in the estimates which might be driven by randomness rather than signal.

The weighted averages lead to identical point estimates, irrespective of whether we consider an analysis at the tag level or at the DDC level (cf. equations 3 and 5), but the precision is higher at the DDC level. This will always be the case, the analysis at the DDC level is strictly pseudoreplication, but naturally only the analysis at the tag level is sensible. When Marques et al. (2009) estimated a cue rate to inform a passive acoustic density estimation exercise for the first time, only 5 tags were available, and hence considering the DDC as the independent sampling unit was an attractive choice. That choice, and implicit assumption about independence across DDCs, meant that instead of a 5-tag sample size, the authors considered a sample size of 21 DDCs. But as with most assumptions - there are no free lunches in statistics - that assumption comes at a cost. Here illustrated for the case of the sperm whales, that cost is underestimating the true precision. In other words, we underestimate the variance of the cue rate estimate, and therefore we would underestimate the true variance on the corresponding estimated density.

For our sperm whale example, considering the unweighted average at the DDC overestimates the cue rate. This is a consequence of most DDCs being below 1 hour, and therefore, having higher cue rates than all the DDCs together. This is clearly evidenced when one looks at the contrast in behavior between the reduced and complete datasets at the DDC level. Cue rate increases in the complete dataset, but it decreases in the reduced dataset. When you compare just between DDC averages, naturally you obtain a higher value for the weighted average in the reduced dataset. Again, this happens because longer DDCs, with lower cue rates, were removed. On the other hand, you obtain a lower value for the DDC weighted average in the full dataset, since in that dataset there are a considerable large number of longer DDCs, typically with lower cue rates, being then given larger weights. There is a correlation between the cue rate by DDC and the DDC duration, with the largest DDCs being associated with the lowest cue rates. This happens because these correspond to DDCs where whales might spend a considerable amount of time at the surface. In fact, for these instances, the definition we considered for a DDC can be misleading: for long DDCs the whale might spend much more time doing something else, like resting or socializing, than actually performing the deep foraging dive than coins the DDC term used to define the period. Hence, when we weight by DDC duration, compared to an unweighted tag analysis, we underestimate the cue rate by about 100*(0.947-0.911)/ 0.947 =3.84%, which would correspond to, all else being equal, an upward bias in density of 100*(1/0.911-1/0.947)/(1/0.947)=3.99%.

While 4% might be a relatively small bias given the usual coefficient of variation of abundance/density estimates, any bias that can be removed necessarily leads to improved inferences. Additionally, this correlation between cue rate and the weights (DDC duration) means that properly calculating the variance of the corresponding weighted mean would require incorporating the covariance between the observations and the weights, which is far from straightforward.

The decision of weighing or not by recording duration is unfortunately more nuanced than one might hope. One can easily construct a scenario where such weighting would be desirable. An example is when shorter duration tags do not provide a reliable cue rate for the tagged animal. On the other hand, when all tags are long enough to obtain a reliable individual cue rate per tag, weighting could induce bias towards some animals with unusually long records. The decision will depend on how variable animals are over time compared to the variability between animals. Weighting becomes more relevant as variability within animals increases and across animals decreases, but where to change from a standard average to a weighted average given said ratio of variabilities and average tag duration might be a hard question to answer.

B. About regression models

The analysis considering (1) the GEE regression model and (2) the weighted average, considering DDC as the sampling unit, estimate the same quantity, and hence we get the same point estimates for either dataset. We do note that the GEE confidence intervals are wider than those for the weighted mean, reflecting lower precision in the estimated means. This is a more sensible precision, since tag deployments are the independent sampling units, not DDCs.

The pattern found in Fig. 3 illustrates that, while required given the definition of what a cue rate is (cues per unit time), from a regression modelling perspective, DDC duration might not be a sensible offset, since the relationship between it and the cue rate is far from being proportional. In other words, an estimated coefficient for the offset included in the model as a variable would not be 1, which is strictly what an offset corresponds to. This is a reminder that use of offsets in regression corresponds to an implicit, often unstated, assumption. Since this is an assumption that is easy to test, by plotting the data as we did, we suggest in general should be tested when using an offset.

GEEs and GLMMs model different conceptual quantities. GEEs are often referred to as marginal models, and provide population level averages, while GLMMs, also known as conditional models, will provide individual level averages (Fieberg et al. 2009). This is often ignored, especially if the interest of inferences is on how covariates influence a response. However, in the case of cue rates from tag data, the distinction might be crucial. We are actually interested on the mean estimated by the GLMM, a mean across whales, not by that of the GEE as implemented, a mean across DDCs. At least for our example, differences in point estimates across the two approaches were minimal, with slightly higher variances estimated via the GLMM, but that might not be the case in general for other species that might have a different diving and sound production behaviour.

One might wonder why bother with regression models if analytical expressions of averages provide such similar results for mean cue rates. The power of regression approaches truly emerges when additional covariates that cue rates might depend on are available. That opens the door to model-based estimates of cue rate, that can be predicted for the actual survey conditions. As an hypothetical example, one could imagine cue rate being dependent on survey level covariates, like study area bottom depth, because animals spend less or more time silent travelling to the bottom to feed depending on bottom depth. Then a regression model allows one to estimate the cue rate for the depths at which the survey sensors were placed. One might additionally be able to model cue rates as a function of animal level covariates (e.g. sex), or even covariates that change over time within animal (e.g. animal depth). While this might provide interesting biological information, for a PAM density estimate we will require an average cue rate, which will have to be averaged over the survey conditions. Hence, knowing that cue rate differs by sex or by the depth at which an animal is not enough and might be of little use in practice: To use that information we would need to know, for all the animals within the survey area, the sex of the animals or the depths at which they were diving, to average across those distributions to obtain the correct multiplier. A pragmatic approach in such cases, which comes at the cost of a strong untestable assumption, is to assume that the sample of animals we have tagged provides an unbiased sample over which, once averaged across, we can estimate the corresponding average cue rate.

C. Cue rates to inform PAM surveys

As a conclusion, we highlight what we knew from the start. We want a cue rate estimate to convert a cue density per time into an animal density. Therefore, we need the average cue rate that applies for the time and place the survey took place.

Fundamentally, the variability in this cue rate estimate that we are interested is that across animals, hopefully obtained from a random sample of animals. Therefore, an analysis that focuses on the DDCs, where the duration of the DDCs is correlated with the variable of interest, here cue rate, might be biased. We recommend that researchers calculating cue rates from similar tag data take due care to consider analysis that reflect the variability at the level of the individual whales. We encourage researchers to be careful when considering inferences where DDCs might be tempting to use as natural sampling units.

Cue rates are a required multiplier for cue counting approaches to estimate animal density and abundance. Reliable methods to estimate the cue rate and its corresponding variability are needed. For the sake of this paper, with the objective of evaluating potential bias induced by the methods used, we pooled all the data and focused on a pooled cue rate across space and time, as if the samples were a suitable random sample for that purpose. Nonetheless, the sample was not balanced in either time or space, so it might be biased for any given time or place. Even for a given time and place we have tags for, one should consider carefully whether a sample of tagged animals is, in general, a representative sample of the animals available. With a small sample might be likely to get a few STRANGE animals (a la Webster and Rutz, 2020) that could compromise inferences. As an example, if animals to be tagged are found acoustically, implying they would be in a "vocally active" mode, cue rates of these animals could be potentially higher than the cue rates of other animals, biasing cue rate up and correspondingly density estimates low. Obtaining a cue rate for a new location or time period should require a good understanding about a species' cue production and implications of potential spatio-temporal differences in cue rates.

While we considered here a pooled mean across all tags, collected at different times and different places, we know ultimately a cue rate might be affected by a multitude of factors, be it season, region, demography, etc. That is the objective of current research and is unlikely to lead to an answer that fits all questions.

We did not consider any additional covariates in the regression models. A model-based approach where cue rate can be predicted for the time and place one's survey was conducted, conditional on observed covariates, might be desirable. For cues like echolocation clicks the cue rate variability and factors affecting it might be tamed, and a reasonable value obtained in such a way. At the other extreme, say for social sounds, said variability might preclude obtaining estimates with acceptable precision. In such cases, or when cue rate is density dependent, the only option to use a cue counting approach for density estimation might be to use a sample of tagged animals to estimate the cue rate for the place and time the survey took place. This remains a considerable drawback in estimating animal abundance from passive acoustic data via cue counting, and further work is required to understand drivers of cue rate variability and to identify for which species and cues these might be stable enough to lead to reliable estimates of abundance.

SUPPLEMENTARY MATERIAL

See supplementary material at [URL will be inserted by AIP] for a dynamic report that reproduces all the analysis and outputs statistics and figures on the paper via RMarkdown. The data and the dynamic report are hosted at a github repository:

https://github.com/TiagoAMarques/DeepDiverCueRates/. We also host there the custom-built MATLAB functions used to identify the clicks of the tagged whales.

ACKNOWLEDGMENTS

This research was conducted under the ACCURATE project, funded by the US Navy Living Marine Resources program (contract no. N3943019C2176). TAM and CSM thank partial support by CEAUL (funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UIDB/00006/2020). We would like to acknowledge the hard work of all the researchers that collected the tag data and that facilitated our access to said data, either in raw format or especially as processed files with times of click emission by the tagged whale: Patrick Miller, Mark Johnson, Peter Tyack, Peter Teglberg Madsen, Shane Gero, Pernille Helene Tønnesen, Mónica Silva, Claudia Inês Botelho de Oliveira, Rui Prieto, Leigh Hickmott and Michael Moore. Chloe Malinka provided very useful comments on a draft of the manuscript. We additionally thank Mark Johnson for having developed the DTAGs and the code for tools that support their use which we leveraged in this paper.

AUTHOR DECLARATIONS

The authors have no conflicts of interest to declare. The data collection was carried out by the researchers contributing with their tags to this joint effort, all of them operating under strict ethical permits. This manuscript was produced within

ACCURATE, and the project got a secondary data Ethics Approval from the School of Biology Ethics Committee at the University of St Andrews.

REFERENCES

Barlow, J. & Taylor, B. L. 2006 Estimates of sperm whale abundance in the northeastern temperate Pacific from a combined acoustic and visual survey Marine Mammal Science 21, 429-445

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67: 1-48

Bolker, B. M.; Brooks, M. E.; Clark, C. J.; Geange, S. W.; Poulsen, J. R.; Stevens, M. H. H. & White, J.-S. S. 2009 Generalized linear mixed models: a practical guide for ecology and evolution Trends in Ecology & Evolution 24, 127 – 135

Fieberg, J.; Rieger, R. H.; Zicus, M.C. & Schildcrout, J. S. 2009 Regression modelling of correlated data in ecology: subject-specific and population averaged response patterns Journal of Applied Ecology 46, 1018-1025

Gatz, D. F. & Smith, L. 1995 The standard error of a weighted mean concentration--I. Bootstrapping vs other methods Atmospheric Environment, 29, 1185-1193

Hiby, A. R. & Ward, A. J. 1986 Analysis of cue-counting and blow rate estimation experiments carried out during the 1984/85 IDCR minke whale assessment cruise Report of the International Whaling Commission 36, 473-476

Højsgaard, S., Halekoh, U. & Yan J. (2006) The R Package geepack for Generalized Estimating Equations Journal of Statistical Software, 15: 1--11

Hurlbert, Stuart H. 1984 Pseudoreplication and the design of ecological field experiments Ecological Monographs 54, 187-211

Johnson, M. P. and Tyack, P. L. (2003) A digital acoustic recording tag for measuring the response of wild marine mammals to sound IEEE Journal of Oceanic Engineering 28, 3-12

Johnson, M., Madsen, P. T., Zimmer, W. M. X., De Soto, N. A., & Tyack, P. L. (2006). Foraging Blainville's beaked whales (*Mesoplodon densirostris*) produce distinct click types matched to different phases of echolocation. Journal of Experimental Biology, 209(24), 5038-5050

Lewis, T., Gillespie, D., Lacey, C., Matthews, J., Danbolt, M., Leaper, R., McLanaghan, R. & Moscrop, A. 2007 Sperm whale abundance estimates from acoustic surveys of the Ionian Sea and Straits of Sicily in 2003 Journal of the Marine Biological Association of the United Kingdom 87, 353-357

Marques, T. A.; Thomas, L.; Ward, J.; DiMarzio, N. & Tyack, P. L. (2009) Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales The Journal of the Acoustical Society of America, 125, 1982-1994

Marques, T. A.; Thomas, L.; Martin, S. W.; Mellinger, D. K.; Ward, J. A.; Moretti, D. J.; Harris, D. & Tyack, P. L. (2013) Estimating animal population density using passive acoustics Biological Reviews, 88, 287-309

Pekár, S. & Brabec, M. 2017 Generalized estimating equations: A pragmatic and flexible approach to the marginal GLM modelling of correlated data in the behavioural sciences Ethology 124, 86-93

Powell, L. A. 2007 Approximating variance of demographic parameters using the delta method: a reference for avian biologists. The Condor, 109, 949-954

Ward, J. A.; Thomas, L.; Jarvis, S.; DiMarzio, N.; Moretti, D.; Marques, T. A.; Dunn, C.; Claridge, D.; Hartvig, E. & Tyack, P. (2012) Passive acoustic density estimation of sperm whales in the Tongue of the Ocean, Bahamas Marine Mammal Science, 28, E444-E455

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. (2017) Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America, 141, 1962-1974

Warren, V.E., Miller, P.J. and Tyack, P.L. 2020. Short-term responses of sperm whales *Physeter macrocephalus* to the attachment of suction cup tags. Marine Ecology Progress Series 645, 219-234.

Webster, Michael M., and Christian Rutz. How STRANGE are your study animals? Nature 582, 337-340