Hindawi Publishing Corporation
International Journal of Zoology
Volume 2012, Article ID 428752, 5 pages
doi:10.1155/2012/428752

Research Article

An Evaluation of Ad Hoc Presence-Only Data in Explaining Patterns of Distribution: Cetacean Sightings from Whale-Watching Vessels

Louisa K. Higby,[1,2] Richard Stafford,[3] and Chiara G. Bertulli[2,4]

[1] *School of Ocean Sciences, Bangor University, Menai Bridge, Anglesey LL59 5AB, UK*
[2] *School of Engineering and Natural Sciences, Faculty of Life and Enviromental Sciences, Elding Whale-Watching, Ægisgata 7, 101 Reykjavik, Iceland*
[3] *Division of Science, Institute of Biomedical and Environmental Science and Technology, University of Bedfordshire, Luton LU1 3JU, UK*
[4] *School of Engineering and Natural Sciences, Faculty of Life and Enviromental Sciences, University of Iceland, Sturlugata 7, 101 Reykjavik, Iceland*

Correspondence should be addressed to Louisa K. Higby, louisahigby@gmail.com

Received 1 March 2012; Accepted 10 April 2012

Academic Editor: Anne Goodenough

The analysis of presence-only data is a problem in determining species distributions and accurately determining population sizes. The collection of such data is common from unequal or nonrandomised effort surveys, such as those surveys conducted by citizen scientists. However, causative regression-based methods have been less well examined using presence-only data. In this study, we examine a range of predictive factors which might influence Cetacean sightings (specifically minke whale sightings) from whale-watching vessels in Faxaflói Bay in Iceland. In this case, environmental variables were collected regularly regardless of whether sightings were recorded. Including absences as well as presence in the analysis resulted in a multiple-generalised linear regression model with significantly more explanatory power than when data were presence only. However, by including extra information on the sightings of the whales, in this case, their observed behaviour when the sighting occurred resulted in a significantly improved model over the presence-only data model. While there are limitations of conducting nonrandomised surveys for the use of predictive models such as regression, presence-only data should not be considered as worthless, and the scope of collection of these data by citizen scientists using modern technology should not be underestimated.

1. Introduction

Presence-only data, data where presence of a species or individual is recorded, but where absences are not, are frequent in many *ad hoc* scientific surveys, such as those datasets collected by volunteers or citizen scientists [1–3]. While presence-only data have been shown to produce good maps of species ranges in some occasions (e.g., [1]), using such data to infer changes in distribution, population sizes, and other ecological parameters can be difficult [2]. Reasons for this largely relate to unequal sampling effort [1–4]. Low- or zero-sampling effort could easily miss the presence of a low-density species in certain areas, but the amount of effort applied is largely unknown, and hence lack of presence of a species could relate to a real absence, or simply a lack of effort. Misidentification of species, or misreporting of locations, can confound such studies, although such issues can also occur in any volunteer programme, regardless of sampling strategy employed [1, 2].

Volunteers, and citizen scientists recruited through "crowd-sourcing" events, however, are a cheap method of collecting data over a wide spatial or temporal scale [1, 5]. As such, presence-only data are becoming common, and an evaluation of their use in scientific research is timely. While this study does not strictly use citizen science data, the sampling regime used is, by necessity, not of equal effort in space or time, and some aspects of the dataset collected are, again by necessity, presence only. However,

since absence data were also recorded, the dataset provides an ideal opportunity to test the use of presence-only data in causative regression models.

Multiple linear regression and associated linear model reduction methods are a common tool in addressing habitat or environmental variables related to habitat selection by organisms [6–8]. For example, in a stepwise model reduction approach, a large number of explanatory factors can be used to predict presence (or number) of individuals in each location, and through examination of each factor's relative importance, the most important factors involved in explaining the majority of the variation in the dependent variable can be found [9].

Such a multiple regression approach to predicting sightings of minke whales (*Balaenoptera acutorostrata*) was used in the current study. Data on sightings of these species were collected from whale-watching vessels, and a range of environmental factors, such as weather conditions, sea state, and temperature, were also recorded or obtained from existing public data sources. These factors were used in conjunction with sightings of Cetaceans and observations on their behaviour to identify the role of environmental factors in the occurrence of whale sightings. Since data were collected regularly throughout the cruise, many factors were known when sightings did not occur, and removing these "absence" data points in subsequent analyses allowed for an evaluation of presence-only data in causative regression models.

2. Methods

Boat-based surveys were carried out from Faxaflói Bay (64.20871° N, 22.19869° W), on the south west coast of Iceland during whale-watching trips. Each survey session was broken down into 15 min observation intervals where environmental variables were recorded (Table 1). In addition, minke whale (*Balaenoptera acutorostrata*) encounters were recorded, along with the numbers seen at each sighting, and this figure was used as a dependent variable in the analysis (with zero indicating absence).

A total of 634 data points were used in the analysis, of which 133 recorded sightings of minke whales (15-minute periods when 1 or more minke whales were observed). These surveys were carried out over 104 days, between April and July 2010 (days when the sea state exceeded Beaufort scale 3 were discarded from the analysis—following prior recommendations, e.g., [10, 11]).

A multiple linear model was constructed with all environmental factors listed in Table 1 as possible explanatory factors for the number of minke whales seen in any 15-minute time period (with the exception of behaviour—which could not be included in models where absence data was used). A stepwise model reduction process was then undertaken (using forward and backward processes) as described in [9], using AIC as a model reduction method. Next, any 15-minute period where no minke whales were seen was removed to create a presence-only dataset, and the stepwise model process ran again. Next, the presence-only dataset was analysed again using all the previous environmental factors along with behaviour of the cetacean when it was seen.

Given that data are count data and residuals from standard linear models were not normally distributed, generalised linear models were used for analysis (as per [12, 13]). The full dataset showed variance far exceeded the mean for the counts of both minke whales and white-beaked dolphins, and GLMs based on the negative binomial distribution were used [13]. When zero counts were excluded, data were still not normally distributed, but followed the assumptions of Poisson distributions (mean~variance), and these GLMs were based on this distribution [13]. Given GLM does not provide a goodness-of-fit statistic (such as R^2), we use (1) an evaluation of the sum of the residuals as a measure of goodness of fit to compare presence and absence data models, where lower values indicate better fit, (2) manually calculated R^2 as a model comparison mechanism, but not as a true measure of the predictive power of the model, by subtracting the quotient of the residual and model deviance from one, and (3) ran the analysis using traditional linear models, despite limitations applied to count data, to provide a more comparative study of the proportion of variability explained. In this case, all data for dependent variables (minke whale sighting number) were $\log_{10} + 1$ transformed, since this normalised the residuals of the presence-only dataset.

3. Results

Analysis of the full dataset, excluding behaviour but including absences of minke whales, produced a reduced model with five significant ($P < 0.05$) explanatory factors (Table 2), a mean squared residual value of 0.528, and an estimated $R^2 = 0.394$. In comparison, traditional linear modelling (despite nonnormal residuals) resulted in a highly significant final model ($F_{9,\,624} = 16.22$; $P < 0.001$) and an adjusted $R^2 = 0.178$.

The reduced dataset—only using data when minke whales were present—produced a model with considerably the worst explanatory power (although significance testing was not possible due to the different sizes of datasets), with two significant explanatory factors (Table 2), with a mean-squared residual value of 0.907 and an estimated $R^2 = 0.140$. Similar decreases in fit were obtained by traditional linear models ($F_{5,\,127} = 2.91$; $P = 0.005$; adjusted $R^2 = 0.086$).

Inclusion of the behaviour of the minke whale as an observation using the presence-only dataset gave an improved fit regression with two significant explanatory factors (Table 2), and mean squared residual value of 0.804 and an R^2 of 0.240. In this case, feeding behaviour occurred during significantly more sightings than the other behaviours. Again, similar trends were found with traditional linear models ($F_{7,\,125} = 4.67$; $P < 0.001$; adjusted $R^2 = 0.163$); this model was significantly better than the presence-only model not including behaviour (ANOVA test on two fitted models, $P = 0.0015$).

From an analysis of explanatory variables in the reduced models (Table 2), it can be seen that cetacean sightings were affected by similar factors in most models—sea and

Table 1: Explanatory factors initially included in the linear model. Use of continuous and discrete (category) variables follows suggestions of [14], where category variables are categorised in as few as possible meaningful categories, and ordinal variables are assumed to be continuous. Logistic variables are categorised as discrete with $n = 2$ levels.

Factor	Variable type	Notes
Date	Continuous	From day 1 to day 158
Boat	Category ($n = 2$)	Differences in height of sighting platform; possible differences in acoustics of boat engines
Time of day	Continuous	That is, 4:30 pm = 16.5
Behaviour of cetacean	Category ($n = 3$)	Surfacing, feeding, and other
Number of humpback whales	Continuous	
Number of killer whale	Continuous	
Number of white-beaked dolphins	Continuous	
Sea state	Continuous	
Percentage of cloud cover	Continuous	
Weather conditions	Category ($n = 3$)	Sun, Cloud, and rain
Wind direction	Category ($n = 5$)	N, E, S, W, and no wind
Tidal conditions	Category ($n = 2$)	Flood or ebb
Swell height	Continuous	
Visibility	Continuous	
Sea surface temperature	Continuous	
Observer	Category ($n = 3$)	Three different observers recorded results

weather conditions were important. Sea temperature was also important in some cases, and the behaviour being performed by the cetacean was important, when included, for explaining the observed sightings of minke whales.

4. Discussion

In general, we demonstrate that presence-only data limit the explanatory properties of models such as multiple linear regression. However, the inclusion of explanatory factors, which can only be included in presence-only models (i.e., they relate to the actual sightings), can increase the power of such modelling approaches on presence-only data.

One important consideration is that datasets containing multiple zero counts require more complex models than those which do not [12, 13]. As such, presence-only data, by its nature, excludes all zero counts, and means models built using standard linear model techniques can be used for analysis. Such models give a much better and intuitive understanding of model fit using the familiar R^2 variable. While in this study, presence-only data were analysed using generalised linear models for comparison with absence data, a simple log transform of the dependent variable normalised the residuals of the linear model approach.

Given that the current dataset was not collected by citizen scientists, we need to consider how the results might apply to citizen-science-collected data, and whether the technique is valuable if applied to the collection of data using citizen science methods. Firstly, what is clearly important is the amount of data available. A successful citizen science programme could greatly increase the number of records and may result in high-quality predictive models being built on presence-only data. Furthermore, it should also be noted that if larger numbers of participants are taking part in such surveys, the chance of missing an actual sighting of a cetacean on a whale-watching trip will be reduced. As such, if the number of "returns" or submission of data is high for each whale-watching trip, then it could be considered that most actual sightings will have been recorded, hence it is more likely that where no data are present, there were no Cetaceans, rather than this being a false absence.

Explanatory models frequently showed factors such as cloud cover, visibility, and sea state to be important. While these factors could influence actual distribution of Cetaceans, it is more likely that they influence the observer's ability to detect them [10, 15]. While such issues may cause some concern for conversion to citizen-science-collected data, the extra volumes of data collected may be able to be standardised for "detection" conditions, by subsetting the data prior to analysis (i.e., into rough or calm conditions, or into sunny versus overcast conditions).

When behaviour was included in the predictive model, there was a significant increase in its explanatory power. In this study, behaviour helped to explain the number of minke whales present at a particular sighting, with more minke whales present when feeding was occurring than for the other behaviours, likely as minke whales may be more likely to be or remain in the presence of a boat when there is significant food available in the area. However, the use of such an approach is only possible when using presence data, since behaviour cannot be recorded if no individual is seen. However, recent research suggests that untrained volunteers are not good at recording behaviour accurately [16]. Despite this, given that data submission could be via photograph or video, this recognition of behaviour could be verified by researchers (e.g., [1]), as could other accuracy-of-data issues,

Table 2: Factors present in the stepwise-reduced model for each dataset where significant multiple linear regressions were obtained. P indicates that the factor was present in the reduced model, and + or − indicates whether there was a positive or negative relationship of the factor compared to the number of Cetaceans. Where no + or − is present, the factor was a category variable. N/A indicates that this variable was not used as an explanatory factor in the analysis.

Factor	Full dataSet	Presence only	Presence only and behaviour
Date	P+		
Boat	P		
Time of day			
Behaviour of cetacean	N/A	N/A	P
Presence of humpback whales			
Presence of killer whale			
Presence of white-beaked dolphin			
Sea state		P−	
Percentage of cloud cover	P+		
Weather conditions			
Wind direction			
Tidal conditions			
Swell height			
Visibility	P+	P+	
Sea surface temperature	P−		P−
Observer			

such as location and time of the record, which can all be verified using modern digital technology [1, 5].

Results relating to observation of surface sightings of minke whales are clearly dependent on the whales being at the surface. Therefore, factors which influence the dive time of the whales will also be apparent in terms of the surface sightings. For example, surface intervals of minke whales are known to vary throughout the year and throughout the day [17], and this seems to be correlated to the type of food the whales are foraging on and the local bathymetry of the foraging site [18]. In this study, both date and sea surface temperature had opposite effects (with a positive relationship between sightings and date, and a negative relationship with sea surface temperature). Such a finding is consistent with previous findings, indicating that surface intervals may be longer in the spring, since whales may feed on plankton blooms, but may also associate themselves with areas of upwelling (cooler water) where these blooms, or other food, may be more abundant [19].

Overall, presence-only data do provide some useful, biological information regarding the sightings of Cetaceans, and the ability to collect a greater volume of data should offset concerns over its value. However, whether data are collected by citizen scientists or trained scientists, there are limitations of the use of tourist vessels in this type of explanatory variable approach. A first step in many analyses of data using model-reduction approaches is the determination of whether sightings or distributions of a species can be explained by chance, by testing the distribution data against a Poisson distribution [20]. Given that tourist vessels automatically head to previous sightings and communicate with each other as to the location of recently seen whales, any approximation to a regular grid, or random sampling, cannot be assumed, and the fact that distribution of whale sightings could be entirely random cannot be ruled out. Furthermore, although for minke whales, much published research has been conducted from whale-watching vessels (e.g., [18, 19]), the behaviour of other cetacean species to whale-watching boats can be variable, with some species, such as humpback whales, avoiding vessels by increasing dive time [21] and others actively approaching boats [22]. In particular, whale-watching vessels appear to record more "active" behaviour, such a leaping out of the water, and also from younger individuals [23]. However, the use of presence-only data, with recordings and understanding of the implications of the behaviour (avoiding, approaching, jumping, etc.), may still allow useful data on factors such as habitat selection to be collected from whale-watching vessels.

As considered elsewhere, the use of digital technology and internet storage facilities can both increase uptake and accuracy of citizen science work [1, 5], and while collection of presence-only data has some disadvantages, some limitations cannot be improved by the use of trained personnel and require dedicated random surveys. In some cases, presence-only data even have some advantages and are worthy of consideration given the current resource cuts to science budgets and the need to greater engage the public with scientific research.

Acknowledgments

The authors would like to thank the Elding Whale Watching Company, with special thanks to G. Vignir Sigursveinsson and Rannveig Grétarsdóttir without whose support, providing a platform for all survey activities, this research would not have been possible. They are grateful to CSI for funding the data collection in the year 2010, and also to the Faxaflói Cetacean Research volunteer Mirjam Held, who helped with

data collection during the 2010 field season in Faxaflói Bay. They would also like to thank the anonymous reviewer for helpful suggestions to improve the paper.

References

[1] R. Stafford, A. G. Hart, L. Collins et al., "Eu-social science: the role of internet social networks in the collection of bee biodiversity data," *PloS one*, vol. 5, no. 12, p. e14381, 2010.

[2] J. Franklin, *Mapping Species Distributions: Spatial Inference and Prediction*, Cambridge University Press, Cambridge, UK, 2009.

[3] M. W. Tingley and S. R. Beissinger, "Detecting range shifts from historical species occurrences: new perspectives on old data," *Trends in Ecology and Evolution*, vol. 24, no. 11, pp. 625–633, 2009.

[4] C. Hassall and D. J. Thompson, "Accounting for recorder effort in the detection of range shifts from historical data," *Methods in Ecology and Evolution*, vol. 1, no. 4, pp. 343–350, 2010.

[5] C. L. Catlin-Groves, "Submitted to this special issue. The citizen science landscape: from volunteers to citizen sensors and beyond," *International Journal of Zoological Research*. In press.

[6] A. E. Goodenough, A. G. Hart, and R. Stafford, "Regression with empirical variable selection: description of a new method and application to ecological datasets," *PLoS One*, vol. 7, no. 3, Article ID e34338, 2012.

[7] M. J. Whittingham, P. A. Stephens, R. B. Bradbury, and R. P. Freckleton, "Why do we still use stepwise modelling in ecology and behaviour?" *Journal of Animal Ecology*, vol. 75, no. 5, pp. 1182–1189, 2006.

[8] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[9] M. J. Crawley, *Statistics: An Introduction Using R*, Wiley, Chichester, UK, 2005.

[10] P. G. H. Evans and P. S. Hammond, "Monitoring cetaceans in European waters," *Mammal Review*, vol. 34, no. 1-2, pp. 131–156, 2004.

[11] D. Palka, "Effects of Beaufort sea state on the sightability of harbor porpoises in the Gulf of Maine," *Forty-Sixth Report of the International Whaling Commission*, pp. 575–582, 1996.

[12] M. Ridout and C. Demetrio, "Generalized linear models for positive count data," *Revista de Matematica e Estatstica*, vol. 10, pp. 139–148, 1992.

[13] M. J. Crawley, *The R Book*, Wiley, Chichester, UK, 2007.

[14] D. J. Pasta, "Learning when to be discrete: continuous vs. categorical predictors," Paper 248, 2009, SAS Global Forum 2009.

[15] S. Dawson, P. Wade, E. Slooten, and J. Barlow, "Design and field methods for sighting surveys of cetaceans in coastal and riverine habitats," *Mammal Review*, vol. 38, no. 1, pp. 19–49, 2008.

[16] R. L. Williams, S. Porter, A. G. Hart, and A. E. Goodenough, "Submitted to this special issue. The accuracy of behavioural data collected by visitors in a zoo environment Can visitors collect meaningful data?" *International Journal of Zoological Research*. In press.

[17] K. A. Stockin, R. S. Fairbairns, E. C. M. Parsons, and D. W. Sims, "Effects of diel and seasonal cycles on the dive duration of the minke whale (Balaenoptera acutorostrata)," *Journal of the Marine Biological Association of the United Kingdom*, vol. 81, no. 1, pp. 189–190, 2001.

[18] K. Macleod, R. Fairbairns, A. Gill et al., "Seasonal distribution of minke whales Balaenoptera acutorostrata in relation to physiography and prey off the Isle of Mull, Scotland," *Marine Ecology Progress Series*, vol. 277, pp. 263–274, 2004.

[19] P. C. Gill, M. G. Morrice, P. Brad, P. Rebecca, A. H. Levings, and C. Michael, "Blue whale habitat selection and within-season distribution in a regional upwelling system off southern Australia," *Marine Ecology Progress Series*, vol. 421, pp. 243–263, 2011.

[20] A. E. Goodenough, S. L. Elliot, and A. G. Hart, "Are nest sites actively chosen? Testing a common assumption for three non-resource limited birds," *Acta Oecologica*, vol. 35, no. 5, pp. 598–602, 2009.

[21] A. Schaffar, B. Madon, V. Garrigue, and R. Constantine, "Avoidance of whale watching boats by humpback whales in their main breeding ground in New Caledonia," Paper SC/61/WW/6, International Whaling Commission, Cambridge, UK, 2009.

[22] F. Ritter, *Interactions of Cetaceans with Whale Watching Boats—Implications for the Management of Whale Watching Tourism*, M.E.E.R.e.V., Berlin, Germany, 2003.

[23] M. Weinrich, "Are behavioral data from whalewatch boats biased?" Paper SC/61/WW/3, International Whaling Commission, Cambridge, UK, 2009.