JASA EXPRESS LETTERS

Automatic data selection for validation: A method to determine cetacean occurrence in large acoustic data sets

Cite as: JASA Express Lett. **1**, 051201 (2021); https://doi.org/10.1121/10.0004851
Submitted: 23 December 2020 . Accepted: 12 April 2021 . Published Online: 04 May 2021

Katie A. Kowarski, Julien J.-Y. Delarue, Briand J. Gaudet, and S. Bruce Martin

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Automatic data selection for validation: A method to determine cetacean occurrence in large acoustic data sets

Katie A. Kowarski,[a] Julien J.-Y. Delarue, Briand J. Gaudet, and S. Bruce Martin[b]
JASCO Applied Sciences, 32 Troop Avenue, Suite 202, Dartmouth, Nova Scotia B3B 1Z1, Canada
katie.kowarski@jasco.com, julien.delarue@jasco.com, briand.gaudet@jasco.com, bruce.martin@jasco.com

Abstract: Passive acoustic monitoring (PAM) can inform wildlife management by providing information on the distribution of cetaceans. This paper presents an automatic data selection for validation (ADSV) method to effectively identify all species acoustically present in large PAM data sets. The ADSV method involves the application of automated detectors, the automated selection of a portion of data for manual review, and the evaluation/optimization of automated detectors. Using an exemplar data set, results from the ADSV method were compared to a more intensive systematic manual review method. The two methods were found to have similar species occurrence results (hourly occurrence matching 73%–100%).

[Editor: Wu-Jung Lee] https://doi.org/10.1121/10.0004851

Received: 23 December 2020 **Accepted:** 12 April 2021 **Published Online:** 4 May 2021

1. Introduction

Effective wildlife management requires an understanding of the occurrence and distribution of populations across seasons, years, and, ideally, decades. Long-term trend data for cetaceans have traditionally been acquired using visual surveys from ships or aircraft. Visual surveys are challenging, as many species are cryptic in nature, have large distributions that include offshore habitats, and undergo seasonal movements. Passive acoustic monitoring (PAM) has emerged as a method to survey cetacean occurrence and distribution.[1] PAM is inherently limited to sampling vocally active animals whose vocal repertoire has been sufficiently described. Technological advances and cost reductions mean that PAM programs routinely include many autonomous acoustic recorders distributed across large regions operating for extended time periods. Consequently, analyzing the large volumes of data in reasonable timeframes is increasingly becoming a greater challenge than collecting data.

Acoustic analysis methods to determine marine mammal acoustic occurrence vary across PAM programs, depending on the aims, scope, species of interest, and budget of the research group.[2] Analysis protocols incorporate human manual review, the application of automated techniques, or a combination of the two[2] (see the supplementary material[3] for a glossary of data analysis terms). Complete manual review of a large acoustic data set is the preferred method to produce reliable "truth data," but it is time-consuming, is costly, can have inconsistent results across analysts, and at the scale of many monitoring programs is unrealistic. Researchers will often limit this effort by systematically reviewing a portion of the data (this could be a subset of time, only a portion of the frequency range recorded, or both), though such dedicated effort is still beyond the reach of many programs and leaves a large amount of data unrepresented in the results.[2] The development of automated detection methods to identify signals of interest has greatly increased the efficiency of PAM data analysis.[1] Automated detectors suffer from false and missed detections due to conflicting sound sources and signal distortion with propagation. Therefore, some manual validation is required to quantify the performance and reliability of automated detectors,[2,4] where automated detector performance can be described per discrete acoustic signal or per unit time.

A critical question for the design of PAM programs is how much data to analyze for manual validation and how to select them. Thus far, there has arguably been no consensus in the marine PAM community. Some researchers validate less than 1% of their data,[5] while others validate 20%.[6] For some large PAM programs that combine many data sets, a large percentage of data is validated for one data set and less for the remainder, resulting in a range of data validated

[a]Author to whom correspondence should be addressed, ORCID: 0000-0002-1325-8321.
[b]ORCID: 0000-0002-6681-9129.

(~2%–33%).[7] Methods to select data are just as varied as the amount selected, ranging from selecting randomly to selecting based on ambient levels to systematically selecting, for example, every third day.[2]

The acoustic analysis method selected for most monitoring programs often targets a single species or species group. Indeed, in a recent review of baleen whale PAM literature, only 9 of 94 studies attempted to determine the occurrence of more than two species.[2] Describing the acoustic occurrence of one or two species in an area is certainly extremely valuable, but acoustic data sets can hold information on the entire acoustic ecosystem, the majority of which goes undescribed due to limitations in analysis budget and time. By not reporting on all species recorded in a PAM program in a timely manner, researchers risk missing a changing trend in animal occurrence or distribution, limit their scope in an aquatic realm where the ecosystem is inherently interconnected, and reduce the return on their investment.

Here, we describe a method to determine the acoustic occurrence of multiple cetacean species and species groups in large data sets. We propose that after a suite of automated detectors are applied, researchers can use the available information to determine how much data to manually validate (at minimum). We present a method for automatic data selection for validation (ADSV), the concepts of which are applicable for future occurrence studies, be they for one or more species. For an example PAM program from Western Canada, the results of our methodology (in terms of species daily and hourly occurrence) are compared to results of a more manually intensive method, the systematic manual review of a subset of data.

1.1 Example PAM program

Acoustic data were collected by Fisheries and Oceans Canada using a near-bottom static acoustic recorder (AMAR G3; JASCO Applied Sciences, Dartmouth, NS, Canada) from 12 July 2017 to 15 July 2018 on the Gowgaia Shelf off Western Canada (52.39354° latitude, −131.71309° longitude, 743 m water depth) with a 15 min duty-cycled recording schedule of 5 min 41 s sampled at 16 kHz (low-frequency data), 1 min 4 s sampled at 250 kHz (high-frequency data), and 9 min 15 s of sleep mode, resulting in 1.56 terabytes of acoustic data.

2. Multi-species occurrence ADSV methodology

2.1 Automated detection

The first step in the present methodology is to apply a suite of automated detector-classifiers (henceforth referred to as automated detectors; see glossary of terms[3]) to capture all signals of interest expected in the data (Fig. 1). Applying numerous automated detectors (or one automated detector scheme with multiple classifiers) to identify many species, or many vocalization types from one or more species, is already common practice.[8] A range of automated detection techniques has been developed that would be appropriate for the automated detection stage of the present methodology.[1,9,10] The suite of automated detectors applied to the Western Canada PAM data set included those targeting the clicks of delphinids, beaked whales, and porpoise applied to the high-frequency data and those targeting the tonal sounds of baleen and killer whales applied to the low-frequency data (see description in the supplementary material[3]). The present methodology can be applied to a suite of automated detectors targeting different acoustic signals, the same acoustic signals, or a combination of the two.

2.2 Manual validation using ADSV

The results of automated detectors must never be trusted at face value; rather, they should be critically evaluated through manual review of a subset of the data. Previous studies have selected data for manual validation based on automated detector counts (per unit time),[11] time (to capture the entire recording period),[4] and location (to capture all recording sites).[12] Given the known variability in automated detector performance that can occur across these variables,[5,13]

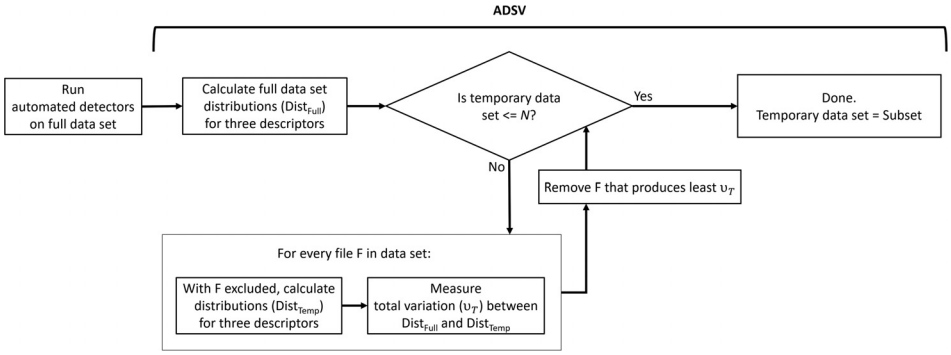


Fig. 1. Flow chart of ADSV algorithm where *N* is the predetermined percent duration of acoustic files from the full data set to be included in the manual analysis subset.

researchers should account for them all when selecting data for manual validation. This is the aim of the ADSV method. The concepts of ADSV are described here, with further details included in the supplementary material.[3]

The ADSV algorithm was developed to automatically select a subset of acoustic files for manual review from the full data set based on automated detections (Fig. 1). The first step in ADSV is generating three histograms that capture the distributions of three descriptors that are based on the number and types of automated detections per unit time in the full data set (Figs. 1 and 2). The three ADSV descriptors are (1) the number of automated detectors that triggered per unit time (diversity), (2) the number of automated detections per automated detector per unit time (counts), and (3) the

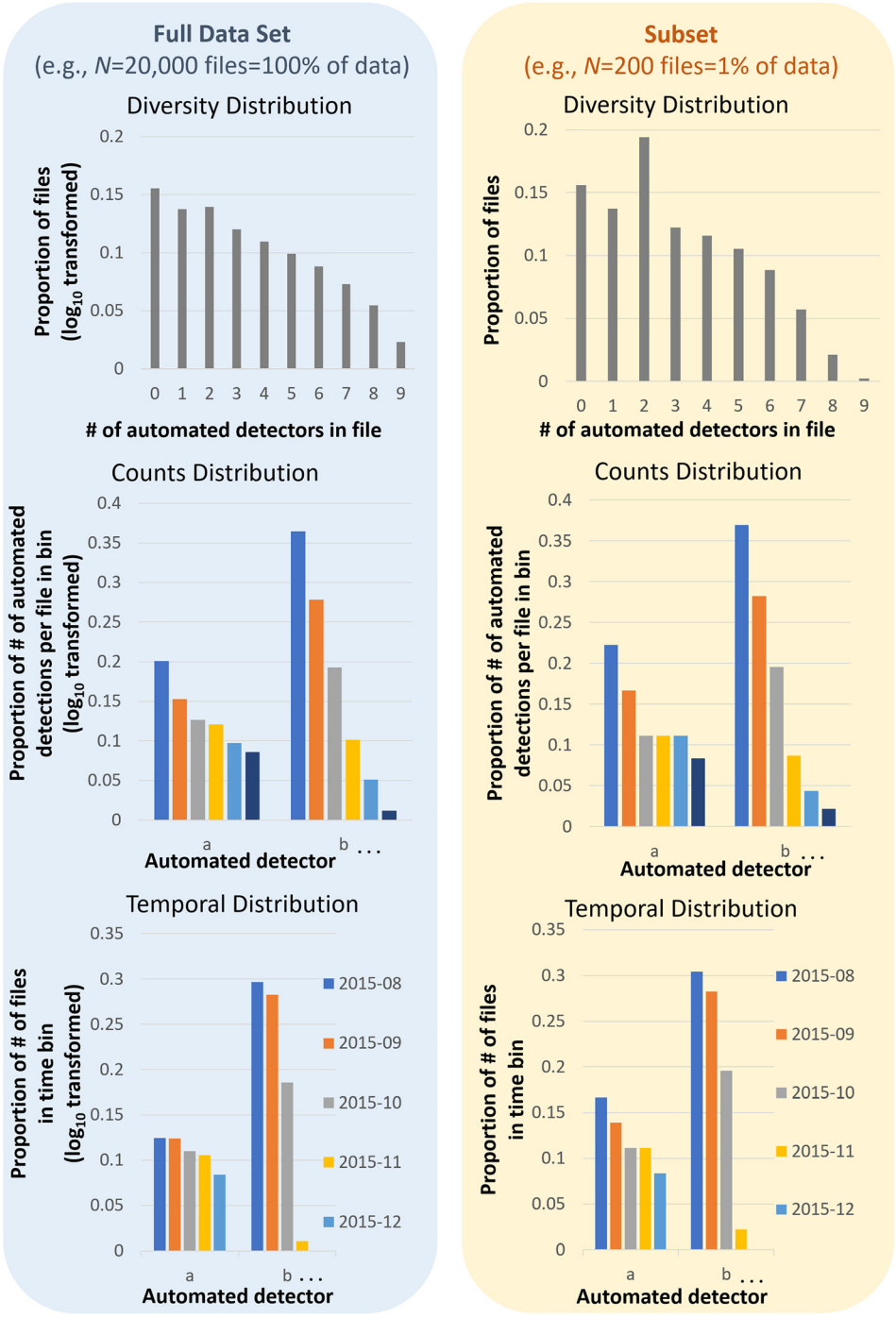


Fig. 2. Example histogram distributions of a full data set and the associated subset selected by ADSV for the three descriptors, where *N* is the predetermined percent duration of acoustic files from the full data set to be included in the manual analysis subset.

distribution of each automated detector's output over time (temporal distribution; Fig. 2; these are explained below). The unit of time for analysis depends on how the data were recorded and subsequently stored, merged, or split; the software used for data analysis; the species potentially present; and the research goals. As most programs and recorders tend to generate files of a fixed short duration, for example, either 1 or 5.5 min in our Western Canadian example, the discrete files can be a convenient choice. The term "file" will be used to mean "per unit time" in this paper.

The first descriptor is the number of automated detectors that were triggered in each file (diversity), where a file in which two automated detectors triggered would have a diversity of 2. By including diversity, periods where automated detectors may have falsely detected vocalizations made by other species are captured. For example, right whale automated upcall detectors often perform differently when humpback whales are singing than when they are not, because some humpback whale vocalizations have characteristics similar to right whale upcalls.[14] Therefore, if both situations occur within a data set, both situations should be manually validated when evaluating automated detector performance. The diversity distribution histogram for the full data set was created by summing the number of files within each bin, where each bin corresponded to the number of automated detectors triggered per file (e.g., top left of Fig. 2). This distribution was transformed using a modified $\log_{10}$ function to reduce the skew of the data (see supplementary material[3]).

The second descriptor is the number of automated detections per file for each automated detector (counts). By including counts, periods with unusually low or unusually high automated detection counts are captured. Variation in counts may represent differences in the signal-to-noise ratios (SNRs) and varying occurrence of conflicting signals. By ensuring that the distribution of reviewed data files includes low counts, researchers can observe if a single detection per file is false or if there are other vocalizations that are low SNR (e.g., distant animal or masked by noise). The counts distribution histogram for the full data set is created from the number of automated detections per file per automated detector. Much like diversity, counts is modified $\log_{10}$ transformed (see supplementary material[3]) to manage the skew of the data, and then the appropriate histogram bin size is applied (Freedman–Diaconis rule with a maximum of 20 bins, e.g., middle left of Fig. 2).

The third descriptor is the temporal distribution of each set of automated detector results across the entire recording period. By incorporating temporal distribution, the many variables that impact the acoustic environment, such as seasonal change, are captured. For example, automated blue whale song note detections may work well during the winter breeding season when most songs are produced, but automated detectors may be falsely triggered by vessel noise during summer. The temporal distribution histograms for the full data set for each automated detector are created by dividing the recording period of the full data set into 12 equally sized timeframes (bins; maximum bin size of 30 days) and summing the number of files with the automated detector triggered in each bin before the data are modified $\log_{10}$ transformed (see supplementary material,[3] e.g., bottom left of Fig. 2).

Once the distributions of the three descriptors for the full data set are calculated ($\text{Dist}_{\text{Full}}$ in Fig. 1), the ADSV algorithm iteratively removes one file at a time from a temporary data set and calculates the distributions of the three descriptors with the file removed ($\text{Dist}_{\text{Temp}}$ in Fig. 1). Histogram distributions for the temporary data set are calculated in the same manner as the full data set with the exception that they are not log transformed. The temporary data set and associated final subset are not transformed to ensure the subset adequately captures the tails of the full data set distributions (files with very low and very high numbers of automated detections). Capturing a range of files with low automated detections is important to determine the minimum number of detections per file required to be confident the species is present (Sec. 2.3).

The variation between the original full data set distributions and the temporary data set distributions is calculated, where variation is the sum of the absolute differences between the histogram bins of the full data set and the associated histogram bins of the temporary data set (see example histograms in Fig. 2 and supplementary material[3] for more detailed variation calculation). In each iteration, the file whose removal produces the least variation between the temporary distributions and the full data set distributions is removed (Fig. 1). The process is repeated until the temporary data set is reduced to a predetermined proportion of the full data set (N), at which point the temporary data set is the subset for manual analysis (Fig. 1). ADSV is run separately on each recording station and deployment period to ensure equal effort across locations. In the Western Canada data set, ADSV was run separately on the high- and low-frequency data (necessary given the different file durations and automated detectors applied).

The question then remains, how much data should be manually validated? Or what is the minimum value of N (N_{min})? For many research groups, this will be determined by the constraining factors of time, budget, and resources. However, it is also possible to compute an optimal value for N_{min} by calculating the variation of the three descriptors between the subset and the full data set with decreasing sample size (N), where low variation indicates little difference between the subset and the full data set (e.g., Fig. 3). The N at which the average variation of the three descriptors is minimal (distribution of subset does not get closer to that of full data set with further decrease in subset size) can be determined. This N_{min} represents the minimal subset size. For the Western Canada data set, N_{min} was approximately 1% of the full data set (Fig. 3); this was the subset size applied for comparing the ADSV selected data to the systematically selected data.

The ADSV sampling method is a type of proportionate stratified sampling made more complex by the three categories (three descriptors) that each contain the entire population (all acoustic files), and each has its own strata (where

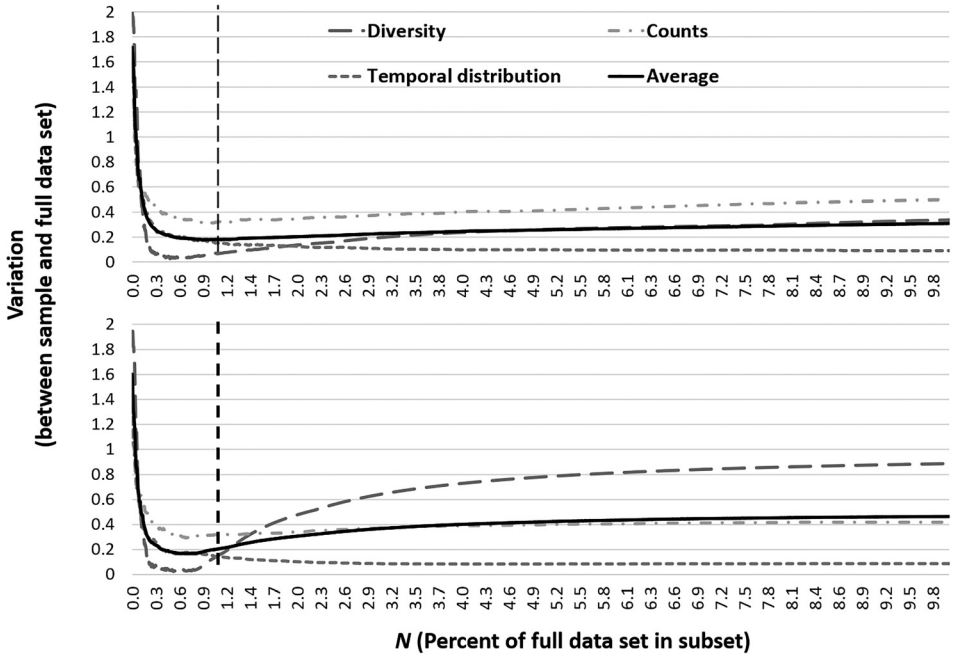


Fig. 3. The variation between the subset and the full data set (where a variation of zero indicates no difference between subset and full data set) for the distribution of automated detector diversity, counts, and temporal distribution and an average of the three variables with increasing percentage of the full data set analyzed (from 0% to 10%, by which point variations plateaued) for the low-frequency (top) and high-frequency (bottom) acoustic data collected off Western Canada. Variations hit a minimum before increasing because full data set distributions of the descriptors are $\log_{10}$ transformed to manage skewness of data, while subset distributions are not. The dashed vertical line indicates the 1% *N* that was applied to the Western Canada data set ADSV.

every acoustic file is in only one stratum per descriptor). The strata (bins in histograms) are based on aspects of automated detector outputs, and the final sample is proportionate to the population as a whole.[15] We propose that such stratified sampling is appropriate for acoustic data because it allows us to maintain the diversity of the population (the diverse range of acoustic events), and it ensures that small subgroups (e.g., rare events) are captured.

Once ADSV provides a subset of data to be manually validated, analysts review the files (or unit of time) to determine the "true" presence or absence of every species. This is achieved by annotating one vocalization from each species (or vocalization type) in every file of the subset. When species or vocalization types are annotated that were not detected, ADSV provides data on false negative events as well as the presence of species for which no automated detectors were applied.

2.3 Automated detector optimization and performance calculation

Once the manual analysis is complete, the automated detector results can be compared to the manual validation results (truth data) for each species (or vocalization type). The automated detector results are restricted to improve automated detector performance. Two stages of automated detector performance optimization are included in the present methodology. In the first stage, temporal restrictions are applied. Manually validated and automatically detected results are plotted as time series to identify any timeframes where species were found to be absent by the manual review but were automatically detected. These presumably false automated detections can be excluded from subsequent analysis. In northern regions, for example, acoustic data can become seasonally inundated with sounds of ice and bearded seals that falsely trigger humpback whale automated detectors. Having never manually confirmed humpback whale vocalizations at that area and time, and knowing that humpback whales prefer open-water habitats, analysts could confidently remove all humpback whale automated detections from northern data sets during those periods. We cannot entirely rule out that a few true positives remain in the excluded periods; therefore, this step is only appropriate for research projects where missing outliers in terms of occurrence is acceptable (e.g., this step may not be appropriate for rare or endangered species, for which validating these detections would be warranted).

In the second stage of automated detector optimization, the present methodology aims to determine if any thresholds need to be applied so that automated detections of a species can be considered valid. Here, thresholds may be the minimum number of automated detections per file that are required to be confident the species is present or a minimum classification confidence score for detectors that generate such measures. The automated and validated results (excluding files where an analyst indicated uncertainty in species occurrence) are fed to a maximum likelihood estimation

algorithm that maximizes the Matthew's correlation coefficient (MCC), which is a measure of detector performance that incorporates true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) (Fig. 4). The algorithm determines an automated detector threshold for each automated detector that maximizes MCC. MCC is used in the present methodology, as it is an appropriate metric in the unbalanced data expected in PAM.[16] The precision (*P*) and recall (*R*) for the resulting threshold are calculated, where *P* is the proportion of files with automated detections that are true positives and *R* is the proportion of files containing the species of interest that are identified by the automated detector (Fig. 4). Examples of automated detector optimization and performance metrics from the Western Canada program are included in the supplementary material.[3]

2.4 Ineffective automated detectors

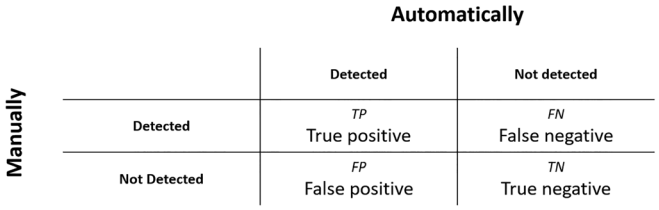
In some instances, once automated detectors have been optimized, the resulting *P* and *R* may indicate an unreliable automated detector. The threshold at which automated detector performance is deemed unacceptable will vary, depending on the requirements of each PAM program. For our purposes, we required a *P* of at least 0.75. In other words, if less than 75% of files with automated detections truly contained the species of interest, then basing species occurrence on automation would be misleading. When automated detectors are deemed ineffective, researchers can (a) present the manually validated results as a minimum occurrence of the species, (b) undertake automated detector development to improve performance, or (c) perform additional systematic manual review that targets the species of concern. The manner in which a research group proceeds will depend on the significance of the species (e.g., whether it is at risk) and the time or expertise that can be allocated to different tasks. No automated detectors in the Western Canada data set were deemed ineffective.

3. Comparing ADSV to systematic manual review

In addition to the ADSV methodology described here that entailed automation and the manual review of 1% of the acoustic data, the Western Canada data set was analyzed via a systematic manual review of 2.5% of the data. The systematic review consisted of analyzing the middle 90 s of every fourth file for the low-frequency data and the middle 45 s of every second file in the high-frequency data. The hourly and daily occurrence results for acoustic signals of seven species were compared between the two methods to determine the percent agreement in presence/absence. To calculate percent occurrence agreement for each species, each timeframe (hour or day) was identified as being in agreement (both methods assigned the species as absent or both methods assigned the species as present) or not in agreement (one method assigned the species as present while the other method assigned the species as absent). The percentage of total recording hours (8858) and days (369) in agreement was calculated and represents the percent occurrence agreement between methods. Occurrence agreement between the methods was high, ranging from 66% to 100%, depending on the species and timeframes considered (Table 1). The variability in agreement can in part be attributed to the systematic manual review only including 2.5% of data, whereas the ADSV technique uses automated detectors that are applied to 100% of data (with 1% manually validated), resulting in ADSV commonly identifying more timeframes (hours or days) with occurrence (Table 1). Imperfect automated detector performance also explains variability in results between methods (Table 1), which highlights the importance of providing automated detector performance metrics (*P*, *R*) whenever marine mammal occurrence results based on automation are presented so that readers can correctly interpret findings.

4. Conclusions

A viable method for multi-species cetacean occurrence analysis in large PAM data sets that combines automation and manual review via ADSV has been presented. We propose that if a subset of data selected for manual validation is



$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(FP+TP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$Recall = R = \frac{TP}{TP+FN}$$

$$Precision = P = \frac{TP}{TP+FP}$$

Fig. 4. Formulas and confusion matrix of automated detector performance metrics. Figure adapted from Kowarski *et al.*[2]

Table 1. The percentage of hours and days that had the same occurrence results (presence or absence) when the Western Canada data were analyzed using the ADSV methodology (1% of data manually reviewed guided by automated detections) versus the systematic manual review of 2.5% of data for seven cetacean species. The final automated detector performance metrics from the ADSV methodology are included for each species as well as the number of hours and days with each species present for each method.

	Final per file automated detector performance		Number of hours (days) with species present		Percent occurrence agreement between methods	
Species	*P*	*R*	ADSV	Systematic manual	Hourly ($N = 8858$)	Daily ($N = 369$)
Baird's beaked whale *Berardius bairdii*	0.95	0.83	35 (28)	18 (16)	99.7	95.1
Blue whale *Balaenoptera musculus*	0.93	0.63	3254 (184)	2887 (158)	88.9	93.0
Fin whale *Balaenoptera physalus*	0.93	0.79	1882 (255)	616 (143)	84.9	68.6
Humpback whale *Megaptera novaeangliae*	0.79	0.81	666 (201)	287 (104)	92.6	65.6
Killer whale *Orcinus orca*	0.97	0.97	110 (58)	51 (22)	99.1	89.2
Porpoise *Phocoena phocoena*	0.98	0.97	629 (271)	617 (261)	93.6	80.5
Sperm whale *Physeter macrocephalus*	0.81	0.43	1454 (250)	3729 (285)	72.3	84.0

representative of the spectrum of soundscapes present in the data, the size of the subset can be minimized; however, manually reviewing as much data as possible is recommended. The concepts of ADSV are broadly applicable and can help research groups attain a more consistent system, or standard, of data selection for manual validation than currently exists. By restricting automated detector outputs post-manual review (e.g., applying temporal restriction or detector thresholds), researchers can optimize detector performance and thus the reliability of results, although possibly at the expense of a few missed detections. The ADSV methodology presented here resulted in similar species occurrence results as a more manual analysis intensive method. Indeed, with less than half the amount of manual analysis effort, our methodology resulted in the same broadscale cetacean trends for the Western Canada data set and, in most cases, identified more instances of species presence. Future work comparing results from these methodologies to one involving the manual review of 100% of data for multi-species occurrence would be valuable. By having an efficient reliable methodology, researchers can better inform management of trends in the occurrence and distribution of cetaceans.

Acknowledgments

We thank Dr. Lynn Lee, Gwaii Haanas Marine Ecologist, for initiating and collaborating on these analyses of passive acoustic recordings from Gwaii Haanas. We thank the cooperative Gwaii Haanas Archipelago Management Board–Council of the Haida Nation, Fisheries and Oceans Canada, and Parks Canada for funding the work as part of their marine monitoring program. Furthermore, we acknowledge Parks Canada's Species at Risk Action Plan Implementation Fund for providing additional project funding and Fisheries and Oceans Canada for technical expertise and operational support to deploy and retrieve the recorders in this unique region. From JASCO, we would like to acknowledge the efforts of Xavier Mouy, Karen Scanlon, Héloïse Frouin-Mouy, and Emily Maxner. We thank the two anonymous reviewers whose suggestions and insight improved the clarity of the manuscript.

References and links

[1]D. K. Mellinger, K. M. Stafford, S. E. Moore, R. P. Dziak, and H. Matsumoto, "An overview of fixed passive acoustic observation methods for cetaceans," Oceanography **20**, 36–45 (2007).

[2]K. A. Kowarski and H. Moors-Murphy, "A review of big data analysis methods for baleen whale passive acoustic monitoring," Mar. Mamm. Sci. **37**(2), 652–673 (2020).

[3]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0004851 for glossary of terms, description of suite of automated detectors, details of ADSV algorithm, and details of automated detector performance.

[4]U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," J. Bus. Res. **70**, 263–286 (2017).

[5]N. E. Balcazar, J. S. Tripovich, H. Klinck, S. L. Nieukirk, D. K. Mellinger, R. P. Dziak, and T. L. Rogers, "Calls reveal population structure of blue whales across the southeast Indian Ocean and the southwest Pacific Ocean," J. Mammal. **96**, 1184–1193 (2015).

[6]S. J. Buchan, R. Hucke-Gaete, K. M. Stafford, and C. W. Clark, "Occasional acoustic presence of Antarctic blue whales on a feeding ground in southern Chile," Mar. Mamm. Sci. **34**, 220–228 (2018).

[7]G. E. Davis, M. F. Baumgartner, J. M. Bonnell, J. Bell, C. Berchok, J. Bort Thornton, S. Brault, G. Buchanan, R. A. Charif, D. Cholewiak, C. W. Clark, P. J. Corkeron, J. J.-Y. Delarue, K. Dudzinski, L. Hatch, J. A. Hildebrand, L. Hodge, H. Klinck, S. Kraus, B. Martin, D. K. Mellinger, H. Moors-Murphy, S. Nieukirk, D. P. Nowacek, S. E. Parks, A. J. Read, A. N. Rice, D. Risch, A. Širović, M. S. Soldevilla, K. Stafford, J. E. Stanistreet, E. Summers, S. Todd, A. Warde, and S. M. Van Parijs, "Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014," Sci. Rep. **7**, 1–12 (2017).

[8]J. E. Stanistreet, D. P. Nowacek, S. Baumann-Pickering, J. T. Bell, D. M. Cholewiak, J. A. Hildebrand, L. E. W. Hodge, H. B. Moors-Murphy, S. M. Van Parijs, and A. J. Read, "Using passive acoustic monitoring to document the distribution of beaked whale species in the western North Atlantic Ocean," Can. J. Fish. Aquat. Sci. **74**, 2098–2109 (2017).

[9]M. F. Baumgartner and S. E. Mussoline, "A generalized baleen whale call detection and classification system," J. Acoust. Soc. Am. **129**, 2889–2902 (2011).
[10]H. Klinck, D. K. Mellinger, K. Klinck, N. M. Bogue, J. C. Luby, W. A. Jump, G. B. Shilling, T. Litchendorf, A. S. Wood, G. S. Schorr, and R. W. Baird, "Near-real-time acoustic monitoring of beaked whales and other cetaceans using a Seaglider^{TM}," PLoS One **7**, e36128 (2012).
[11]K. A. Kowarski, J. J.-Y. Delarue, B. Martin, J. O'Brien, R. Meade, O. Ó., Cadhla, and S. D. Berrow, "Signals from the deep: Spatial and temporal acoustic occurrence of beaked whales off western Ireland," PLoS One **13**, e0199431 (2018).
[12]D. Risch, S. C. Wilson, M. Hoogerwerf, N. C. F. van Geel, E. W. J. Edwards, and K. L. Brookes, "Seasonal and diel acoustic presence of North Atlantic minke whales in the North Sea," Sci. Rep. **9**, 3571 (2019).
[13]D. Risch, C. W. Clark, P. J. Dugan, M. Popescu, U. Siebert, and S. M. Van Parijs, "Minke whale acoustic behavior and multi-year seasonal and diel vocalization patterns in Massachusetts Bay, USA," Mar. Ecol. Prog. Ser. **489**, 279–295 (2013).
[14]S. E. Parks, A. Searby, A. Célérier, M. P. Johnson, D. P. Nowacek, and P. L. Tyack, "Sound production behavior of individual North Atlantic right whales: Implications for passive acoustic monitoring," Endanger. Species Res. **15**, 63–76 (2011).
[15]V. L. Parsons, "Stratified sampling," in *Wiley StatsRef: Statistics Reference Online*, edited by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels (Wiley, New York, 2017).
[16]S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," PLoS One **12**, e0177678 (2017).