U.S. Fish & Wildlife Service and U.S. Geological Survey

Testing Procedures, Performance Criteria and Approval Process for Automated Acoustic Bat ID Software Programs associated with the Range-wide Indiana Bat Summer Survey Guidelines

Overview

Evaluation of software for use in Indiana bat (*Myotis sodalis*; MYSO) and northern long-eared bat (*Myotis septentrionalis*; MYSE) presence/probable absence (p/a) acoustic surveys is a collaborative effort between the U.S. Fish and Wildlife Service (Service/USFWS) and the U.S. Geological Survey (USGS), Virginia Cooperative Fish and Wildlife Research Unit (VCFWRU). Performance criteria needed to achieve software approval are set by the Service, with input from experienced bat ecologists, statisticians, and regulatory specialists. For software to be reviewed and tested, developers must submit their program along with an official submittal form (available on the Service's Automated Acoustic Bat ID Software Program webpage)[1] to the Service. Developers may submit software to the Service to be tested at any time and at no cost. All submissions and inquiries regarding software testing must be directly made to the Service. Once submitted to the Service, the submittal form will be reviewed, and if complete, the software package and submittal form will be forwarded to the VCFWRU for official testing. Test results from the VCFWRU are reviewed by the Service and provided to the software developers with determination of acceptability (Figure 1). In the interest of improving software performance, the Service, with VCFWRU, will discuss results with software developers. However, the Service asks that software developers not use the review process as a beta-testing platform.

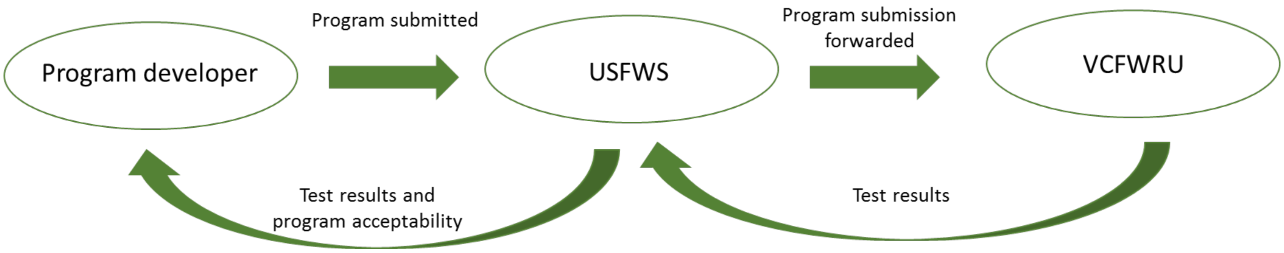
Program developer
Program submitted
USFWS
Program submission forwarded
VCFWRU
Test results and program acceptability
Test results

Figure 1. Relationship between software developers, the USFWS, and the USGS Virginia Cooperative Fish and Wildlife Research Unit in automated bat acoustic identification software evaluation and testing.

If the Service determines test results meet performance criteria, the tested software version is approved for use (but, see notice below) for MYSO and MYSE p/a surveys under the settings used in testing. If results do not meet performance criteria, software are not approved for use in official MYSO and MYSE p/a surveys; failure to meet performance criteria does not preclude software from being revised and resubmitted for testing or from software being available for other purposes. Regardless of test outcome, VCFWRU test results are published on the Service's webpage. If developers plan to issue a new version of

[1] Submittal form is available a on the Service's Automated Acoustic Bat ID Software Programs webpage: https://www.fws.gov/midwest/Endangered/mammals/inba/surveys/inbaAcousticSoftware.html

their software containing **any** change to their classification algorithms, maximum likelihood estimate (MLE) calculations, or other code that could influence species classification, they must submit it for testing and meet performance criteria before it will be considered "Service-approved."

NOTICE: The Service has made significant improvements to software testing procedures (further described below). Periodically re-testing programs will encourage ongoing software development in step with improvements to standardized testing from expansion of test call libraries, and ongoing developments in bat identification science. Developers are welcome to re-submit programs without changes to classifiers if they believe classifier performance is sufficient to meet the Service's current performance requirements.

The Master Test Library and Selection of Test Sets

Since acoustics were first allowed as a valid p/a survey option for MYSOs, the Service and USGS have partnered to establish a Master Test Library (MTL) containing > 2,500 of zero-cross and full-spectrum call files recorded from known bat species that occur across or within portions of the MYSO and MYSE ranges. The MTL also contains many other sound recordings commonly encountered during field surveys (e.g., insect calls). The VCFWRU maintains the MTL.

Prior to testing each new software submission, the VCFWRU generates 10 unique test sets (i.e., Test Sets 1-10) by randomly selecting (without replacement) call files of each species included in the test from the MTL (Figure 2). Test sets contain call files from MYSO, MYSE, little brown (*M. lucifugus*), southeastern (*M. austroriparius*), eastern small-footed (*M. leibii*), gray (*M. grisescens*), tri-colored (*Perimyotis subflavus*), eastern red (*Lasiurus borealis*), hoary (*Lasiurus cinereus*), silver-haired (*Lasionycteris noctivigans*), evening (*Nycticeius humeralis*), big brown (*Eptesicus fuscus*), and Rafinesque's big-eared bats (*Corynorhinus rafinesquii*), as well as "noise" files. To better reflect the realities of field-recorded data, a substantial proportion of the MTL is comprised of commonly encountered noise files of various types and are included in each test set. Categories of noise files used in testing include files with multiple bat species, feeding buzzes, flying squirrel calls, random noise (e.g., static, rain), structured noise (e.g., insect calls), and unidentifiable bat calls (e.g., single passes, highly fragmented passes), with numbers of each category present in the test determined independently. Determination of "unidentifiable" status of bat calls within a file inherently is subjective, so files judged as unidentifiable have been reviewed by multiple experienced bat acoustic ecologists.

Proportions of individual species pass files included in randomly generated test sets have been scaled to represent approximate relative abundances observed across the MYSO range [intermediate between pre-white-nose syndrome (WNS) and post-WNS onset populations] and normalized to big brown bat and eastern red bat relative activity levels (Figure 3). Normalizing to big brown bats and eastern red bats allows the test to combine appropriate numbers of files of species that may not commonly co-occur (e.g. *M. grisescens*, and *M. austroriparius*). A single-species pool has been used for testing due to the infeasibility of testing and interpreting all possible individual high-frequency species pools, and for the challenging testing condition it presents. Actual number of calls per species, and thus total calls included in a given test set, are allowed to vary within a pre-specified range. Testing under the full complement of *Myotis* species present across the entire MYSO range represents the most difficult test scenario for accessing accuracy.

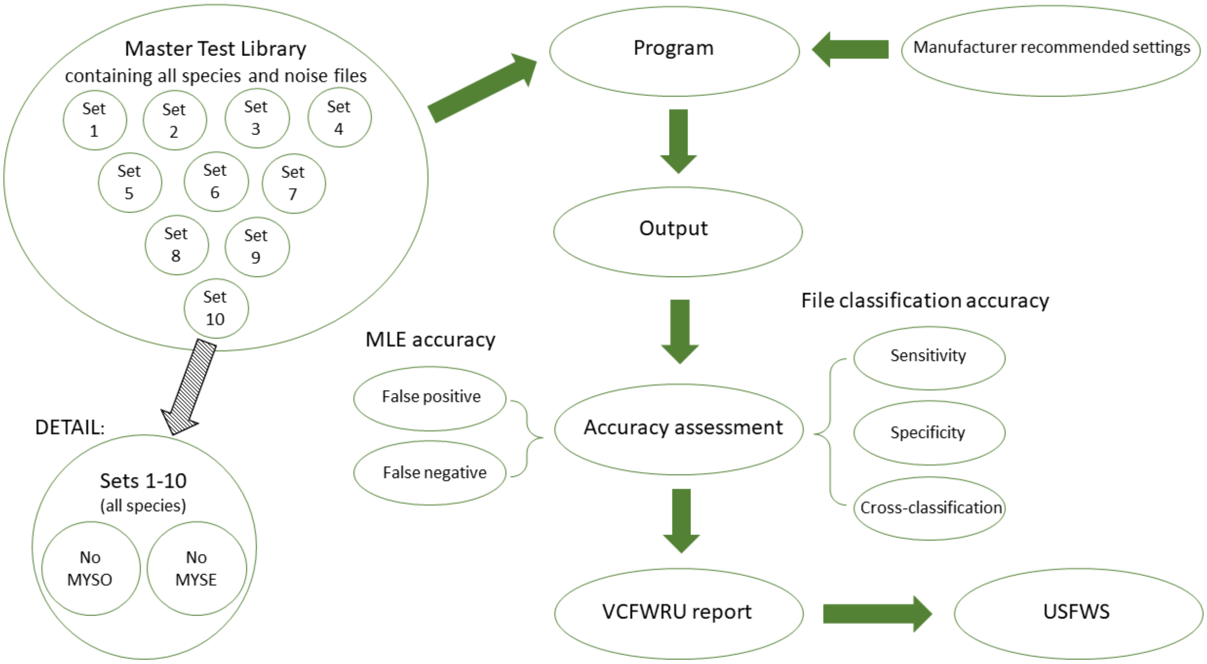

Figure 2. Overview of the testing procedure used by the USGS Virginia Cooperative Fish and Wildlife Research Unit in testing automated bat acoustic identification software submitted to the USFWS for use in Indiana bat (*Myotis sodalis;* MYSO) and northern long-eared bat (*Myotis septentrionalis*; MYSE) acoustic surveys.

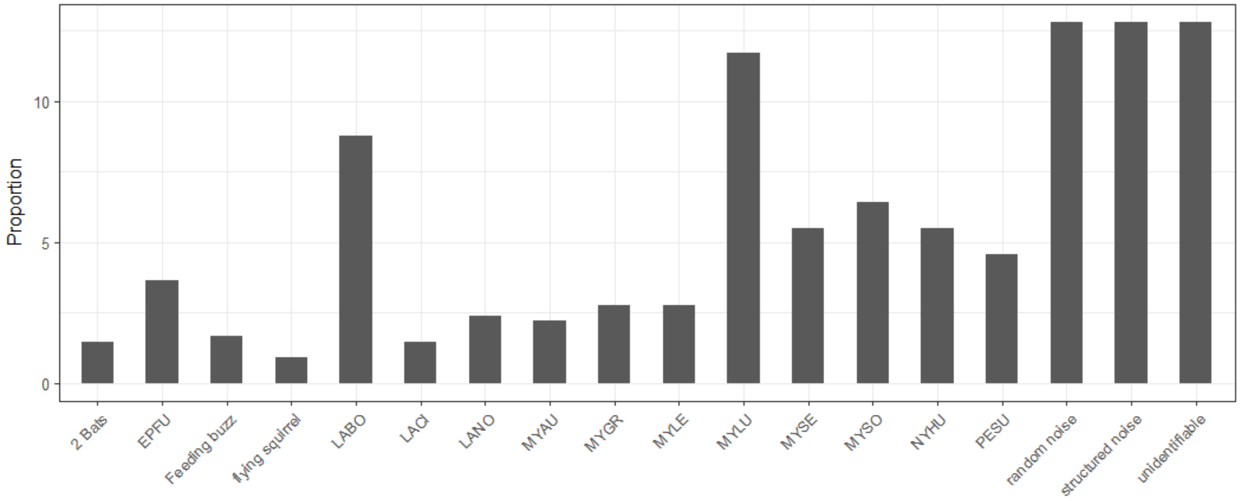

Figure 3. Example of proportions of individual file types that may be included in a randomly generated automated bat acoustic identification software test set. Note that exact proportions will vary randomly within a pre-specified range across tests.

Test Design and Performance Criteria

The 10 randomly selected test sets are run through three different tests. The first test is simply a run of the original/unmodified test sets as an “all species in” test (false negative test; Test Run #1 – ALL). For the second test, all MYSO calls are removed from each of the 10 test sets (false positive test; i.e., Test Run #2 – No MYSO). Similarly, all MYSE calls are removed from the test sets for the third test (false positive test; i.e., Test Run #3 – No MYSE). Therefore, Test Sets 1-10 (modified and unmodified) are tested three separate times for a total of 30 unique tests. This provides a nominal 20 tests for false negative for both MYSO and MYSE, and 10 tests for false positives.

The performance of each software version is tested using the developer’s requested settings (or default in absence of a recommendation) and evaluated based on the accuracy of returned MLE *p*-values for target species presence and absence (i.e. false negative and false positive detections). A MLE *p*-value of 0.05 has been set as the threshold for assessing software accuracy with *p*-values ≤0.05 indicating a species is likely present and *p*-values >0.05 indicating probable absence.

To obtain the Service’s approval for either or both MYSO and MYSE p/a surveys, a new software version must <u>meet or exceed</u> (as applicable) each of the following performance criteria:

1. **100% accuracy** rate in the MLE assessment of **MYSO presence** across all 10 test sets for Test Run #1 – ALL and Test Run #3 – No MYSE (i.e., no false negatives are allowed in the 20 tests where MYSO calls are present), and

2. **100% accuracy** rate in the MLE assessment of **MYSE presence** across all 10 test sets for Test Run #1 – ALL and Test Run #2 – No MYSO (i.e., no false negatives are allowed in the 20 tests where MYSE calls are present), and

3. **80% accuracy** rate in the MLE assessment of **MYSO absence** across the 10 test sets under Test Run #2 – No MYSO (i.e., ≤20% false positive rate/no more than 2 out of the 10 can be wrong).

4. **80% accuracy** rate in the MLE assessment of **MYSE absence** across the 10 test sets under the Test Run #3 - No MYSE (i.e., ≤20% false positive rate/no more than 2 out of the 10 can be wrong).

In short, failure to correctly classify true presence in any test or true absence of one or both species in >2 test sets in Test Run #2 - No MYSO and/or Test Run #3 - No MYSE will result in program failure for one or both species, respectively. Conversely, if the above performance criteria are met or exceeded (as applicable), the Service will approve the “passing” program version (using specified settings) for official p/a survey use for one or both species, as appropriate.

The Service’s previous acoustic software approval tests allowed a 90% false negative rate, and set no false positive rate requirements. The Service selected these new accuracy thresholds to minimize the potential for “take” arising from false negatives, while also minimizing false positives that could unnecessarily affect public and private activities. Requiring zero false negatives (100% accuracy) in acoustic software output was also set considering the currently required minimum level of effort (LOE) for acoustic p/a surveys is set to achieve a 90% confidence level (i.e., MYSOs are likely to be missed up to 10% of the time when actually present)[2].

Because software with adjustable parameters is tested using a developer's recommended/requested settings, it is only approved for official use at those specific settings (presuming it passes the test). In addition to MLE p-values, USGS test reports include some other metrics such as producer's[3], user's[4], and overall accuracy in file-level classification results as an additional explanatory tool for assessing results.

Current Test Approach and Future Development

The current randomized testing approach replaces the previous (2014-2017) approach of using multiple static test sets, and incorporates more call and noise files. The randomized test approach provides several benefits to the Service, users, and developers. Random selection of files in each test set introduces considerable variability into the testing and therefore a better understanding of how software perform as a wider and unpredictable range of variation in call files are analyzed. Likewise, incorporation of variance in relative species proportions ensures rigorous testing of the underlying species classification tables used by software programs to calculate MLE *p*-values. For the Service, randomization allows for more testing over a longer term using a finite number of call files, thus preserving the ability to robustly determine program acceptability for use in acoustic surveys for threatened and endangered bats. For both the Service and developers, randomization helps limit training of software to the test, an inevitable consequence of using static test sets long-term. Randomized test set assembly is anticipated to comprise the primary approach to software testing, but it is the intention of the Service to continuously update the master test library's composition and size as more reference call files are acquired, and to more accurately reflect post-WNS bat community composition. In practice, this may make testing more difficult over time, but should ensure accurate determination of species presence/absence and thus help ensure appropriate conservation measures are taken.

While inclusion of more noise files and random selection of call files makes the current testing procedure more reflective of real-world conditions, it remains unlikely that the currently available calls within the master test library represent the full repertoire of calls each species makes in the wild. To address this issue, the Service will continue to explore additional lines of software testing that include additional variation in call files for potential consideration in future testing protocols. For example, the Service currently is exploring use of field-recorded calls from passive detectors that coincide with physical captures in close spatial proximity. Additionally, the Service is evaluating use of passively recorded call files to assess false positive detections in a field-recorded library when species presence reliably can be excluded by geography or known historic range (e.g. MYSO detections in Newfoundland). Because acoustic detection of a species cannot be guaranteed even when the species is captured, and range expansions and wayward individuals do occur, software results from these approaches would be considered at lesser weight than the randomization approach. Limited initial testing indicates that these

[2] For more information see *Addendum 1 - Methods to Evaluate and Develop Minimum Recommended Summer Survey Effort for Indiana Bats: White Paper*. Available at https://www.fws.gov/midwest/Endangered/mammals/inba/inbasummersurveyguidance.html

[3] Producer's accuracy is 100% - commission error of the target species in the classification; the probability that a file of a given species will be identified as that species. Example – 80% probability that a species X file will be identified as species X.

[4] User's accuracy is 100% - omission error of the target species in the classification; the probability that a file identified as a given species is that species. Example – 85% probability that a file identified as species X is species X.

approaches may not be applicable for establishing program acceptability, but may be useful in better exploring and understanding software performance.

How to Contribute Additional Call Files to the Master Test Library

In the interest of improving the overall number and diversity of call files included in the MTL (see Figure 3), the Service encourages individuals to contribute additional call files (full-spectrum and zero-crossing) of known species. Call files will be considered for inclusion if species identity is unequivocally known. Example methods of generating unequivocal identifications include recordings of hand release and/or light tagged bats. Passively collected and manually identified calls may be considered on a case-by-case basis pending discussion with the Service, but in general are not a preferred source of calls to be used in testing. Files used by developers to train software programs should not be submitted for inclusion in the test library. To permit future development of software testing (e.g. regional test libraries), contributors are requested to included relevant metadata along with file submissions using the call submittal form available on the Service's Automated Acoustic Bat ID Software Programs webpage[5].

If you are interested in contributing call files to the test call library, please contact
Mike Armstrong (Mike_Armstrong@fws.gov), Robyn Niver (Robyn_Niver@fws.gov), or Andy King (Andrew_King@fws.gov).

[5] https://www.fws.gov/midwest/Endangered/mammals/inba/surveys/inbaAcousticSoftware.html