Date: August 17, 2022
Title of Article: A pipeline for assessing the quality of images and metadata from crowd-sourced databases.
Manuscript: **https://www.biorxiv.org/content/10.1101/2022.04.29.490112v1**
External Files:
Name of the Corresponding Author: Jackie Billotte, M.S**.**
Email Address of the Corresponding Author: jackie.billotte@colostate.edu

Dear Reviewers,

Thank you for your careful reading of the manuscript, positive comments, and constructive critiques. I feel that the revisions and responses to the reviewers' comments substantially improve the quality of the manuscript and address some key issues.

I have addressed your individual comments below. Major changes to the manuscript include the following:

1. Use of MatLab to run BRISQUE analysis for image quality has been changed. Image quality analysis is now run using *pybrisque*. The analysis can now be run directly through the same Rmd file as the other pipelines.
2. Due to using *pybrisque* to run the image quality analysis, users no longer need to download any image files. This will greatly reduce the storage space necessary to run the program.
3. A small sample dataset is available on github, allowing users to see the formatting for input files, practice running the pipeline, and view the formatting of the output documents.
4. Figure 1 and Figure 3 have been changed to better demonstrate the workflow of the pipelines.

Again, thank you for providing your insight and recommendations and your consideration of this manuscript.

Sincerely,
Jackie Billotte, M.S.
Ph.D. student, Colorado State University
jackie.billotte@colostate.edu
(720) 590-2373

Detailed response to reviewers

Round #1

1. *by Matthias Foellmer, 22 Jun 2022 15:39*

2. **On our way to harness the full potential of iNaturalist and other databases**
    a. **Dear Jackie Billotte,**
    b. In this manuscript you present a protocol to efficiently evaluate the quality of images and associated metadata of any pre-specified taxon using self-written and co-opted R and MATLAB scripts. Importantly, the usefulness of the approach extends to e.g. assessing the range expansions of invasive species or evaluating the presence of endangered species, potentially supporting conservation efforts. Given the rapidly increasing popularity of iNaturalist and similar databases to which the general public contributes, this paper can make a very valuable and timely contribution. The reviewers agree with this view and provide thoughtful suggestions for improving the clarity and facilitating the use of the various workflow steps. It would be fantastic if you could make BRISQUE work in Python, since, as one of the reviewers points out, few readers will probably have access to MATLAB, which is expensive.
        i. **Thank you for the suggestion to incorporate the Python version of BRISQUE into the image quality assessment. I have done so in the revised version. I agree that providing a free open source version of BRISQUE makes this a much more accessible method. Additionally, the Python version of BRISQUE forgoes the need to download load images from iNaturalist or GBIF and allows the user to screen the images directly from the URL.**
3. In additions to the reviewers' comments, I have the following suggestions:
    a. L20: "… but lower in observations…" – wouldn't "for" be more appropriate than "in"?
        i. **I agree and this is corrected in the revision.**
    b. Data acquisition:
        i. You state that "For the Araneae case study, I searched for and downloaded observations for each family under the order Araneae on iNaturalist on July 21, 2021 ('iNaturalist', 2014). I then searched for and downloaded observations classified only to the order level." Please explain why you employed this strategy. One would think that a single query for Araneae should give all relevant results, with all observations determined to the various taxonomic levels. After all, downloading data for 100+ families one-by-one seems an arduous endeavor I would want to avoid.
            1. **This strategy was implemented due to the limit iNaturalist places on the number of observations that can be downloaded at once. iNaturalist limits the observations to 200,000 observations per download. This has been noted in the revised version of this manuscript (line 81) as it can be quite an undertaking for very large datasets.**
        ii. Datasets to be downloaded from iNat can easily be very large without narrowing down the search criteria. Please detail your search strategy for at least one example. Specify the settings in the filter and the Export

Observations page, so that the reader can reproduce your search results (see also the reviewers' comments). On lines 87ff you only state the minimum requirements with respect to the fields to be included.

1. **Additional information and screen captions of the export settings have been added to the new version of the manuscript.**

iii. When I tried to run your R code for obtaining the data directly from iNat for taxon_name = "Araneidae", I got the error message "Error in get_inat_obs(query = NULL, taxon_name = "Araneidae", taxon_id = NULL, : Your search returned too many results, please consider breaking it up into smaller chunks by year or month." So simply searching for a given family doesn't work, highlighting the need for a more detailed description of your search strategy.

1. **Unfortunately, it appears the R-package that was used has a few bugs when used in newer versions of Rstudio. I kept running into this and similar errors while running it. So, it has been removed from the code. I would be interested in adding it back in once the package has been debugged.**

iv. Your site Observation_Database_Assesment on GitHub.com currently (when I checked) only has the basic R and MATLAB code posted, but no other files. Please add example searches and data sets.

1. **Example input and output files are now available on github. The sample set is a small set of observations from the family Xenoctenidae. The output files show examples of the outputs produced by running the pipelines, as well.**

c. R and MATLAB code: please make sure you provide sufficient annotation so that all steps and their implementation can easily be understood even by the not-so-proficient coder.

i. **Additional annotations have been added to the file on github. As a frequent R and Python user I understand how necessary annotations can be to users. If any additional annotations would be necessary, please let me know.**

d. I hope you find the reviewers' and my comments helpful. I'm looking forward to reading your revision.

Thank you for your thoughtful feedback.

Matthias Foellmer, NYC, 22-Jun-2022

Reviews

4. *Reviewed by Catherine Scott, 06 Jun 2022 14:05*
   a. This is a valuable and timely contribution. As an arachnologist interested in using iNaturalist data, I think it provides some very useful tools and guidelines for processing and using these data.

i. **Thank you and thank you for your helpful comments. I have addressed each of your comments below.**

b. **Major comments:**

i. Unlike R, which is freely available, MATLAB is a paid software so it is not accessible to all. If possible, it would be preferable to have a script for the image analysis in a freely available software. It seems that BRISQUE can be implemented in Python: https://github.com/bukalapak/pybrisque. I do really like the suggestion that iNaturalist automatically score image quality--this would make filtering out useful observations much easier!

1. **Thank you for suggesting the use of *pybrisque*. I agree that providing an open-source, free version of BRISQUE makes this method much more accessible. Additionally, *pybrisque* can be called from the R-markdown file, removing the need for users to download images directly. Your suggestion greatly improved the workflow for image quality assessment.**

ii. While not critical for this paper, which is meant to describe methods, it would be really nice to have an example of the utility of some of the methods, perhaps for a particular family (one of the smaller ones). Going through each of the steps and making the specific small dataset and code available for readers to reproduce the analyses to familiarize themselves with the pipeline, knowing the expected results, would be very valuable. It would also be good to show an example where the GBIF and iNAT ranges do not match, and a look at whether it's because of a mislabeled observation or a true range expansion.

1. **I agree this would be very helpful for users. I have added a sample dataset with sample input and output files to github. The set is limited to 75 observations from iNaturalists and includes a taxonomic reference and GBIF data and URLs. While I was not able to locate a small enough data set that would also demonstrate what range expansion would look like as an output, I will continue to try.**

c. **Minor comments:**

i. since data were non-Normal, it might be more appropriate and informative to report medians and ranges rather than means and SE

1. **The medians and ranges are now reported in the results section in the new version of the manuscript.**

ii. Figure 3 caption is cut off

1. **This is corrected in the revised version**.

iii. line 162: Araneidae must be a typo--presumably this should be one of the smaller families that start with A

1. **Thank you for noticing this. This has been corrected. It was meant to be Archoleptonetidae.**

iv. line 168: "I found 158,129 of the 156,842 downloadable observations" check numbers, have they been reversed?

1. **Again, thank you for bringing this to my attention. The numbers have been placed in the correct order.**

v. line 173: it would be helpful to have some explanation of what it means for an observation to be accurate or precise.
   1. **An explanation for an observation being accurate and/or precise was added at lines 187 and 190 respectively.**

vi. line 206: it is not quite correct that "requires that an observation reach a threshold of three votes from users to confirm an identification" as research-grade. Instead, iNaturalist states that "Observations become "Research Grade" when the community agrees on species-level ID or lower, i.e. when more than 2/3 of identifiers agree on a taxon." In practice this means that an observation can become research grade after exactly two identifiers agree on an ID.
   1. **The clarification has been made in the revised version of the manuscript on line 222.**

d. Note: I did not have a chance to try to run any of the code myself.

5. ***Reviewed by Clive Hambler, 01 Jun 2022 07:38***

a. The variable quality that citizen science observations are rapidly accumulating. The paper and novel method it describes examine aspects of the validity of the accompanying 'metadata' for online images (including crowd-sourced identification at various taxonomic levels). It checks whether each image uses a standard (selected) taxonomy. It enables automated contrasts of the stated geographical location of images with the existing documented range of the taxon. It automatically assesses image quality as another potential filter on images of a taxon. This method will enable more rapid filtering of large databases such as iNaturalist for taxa of interest, retrieving taxa of interest and removing the less reliable records. It will also enable rapid detection of range changes, such as through introductions. It helps researchers and others make the most of the efforts of numerous citizen science image providers and assessors. The pipeline can be used to check the consistency of taxonomic nomenclature between image metadata (such as a crowd-sourced identification) and the preferred taxonomic list for a set of images (in this case spiders on iNaturalist).
   i. **Thank you, I appreciate you taking the time to provide valuable and useful feedback, which were of great help in improving the manuscript. I have addressed each comment below.**

b. The taxonomic accuracy of metadata for images is compared with an 'official' database, in this case the World Spider Catalogue. However, for many taxa there are disputed taxonomies at several levels including species, genus and family. Some additional comments on how to include capacity for expansion of the pipeline to include alternative taxonomic lists would be helpful.

    i. **Thank you for the suggestion. I have included examples of other sources for a taxonomic reference, and this is addressed beginning at line 277 of the revision.**

c. Geo-location metadata of images can be checked for major errors and were quite precise for spiders in this case study. The method permits potentially erroneous species records to be flagged for expert attention if they are in a surprising location. The pipeline can help filter images automatically by computer-assessed image quality, complementing scores given by crowd-sourced identification, and thus helping researchers select the most reliable. The author sees data-volume as the biggest barrier to crowd-sourced data. Perhaps this is true for taxa that can be identified from images, but such taxa are a tiny fraction of all recorded taxa. For the Araneae, used in this paper as a case-study, there are only a few percent of species, even in a fauna as small and relatively well-known as that of Britain, for which I would accept photographic evidence of a species' identity. Most spider species, even as adults, require microscopic identification of a dead specimen. The value will be higher for genus and family level, or for screening for those taxa where photographs can permit reliable identification of some individuals. The method is transferrable and for other major taxonomic groups there may be fewer, or more, limitations.
    i. **I agree different taxa have their own limitations when using this pipeline, especially with taxa rich in cryptic species. Species-level identification using photographic evidence is not always reliable for spiders, as well as many other invertebrates outside of detailed examination and or barcoding. I have addressed this further in the Discussion section of the revision. While examination of image quality does not mean a taxonomic label is accurate, it does provide a method for quantifying the quality of images used and allows users of the pipeline to remove poor-quality images without having to look at each individually.**

d. The methods are not easy to follow for a non-programmer, and the figure captions (such as Figure 1) need to be simpler and more self-explanatory. I am taking the programming methods on trust since I do not have the specialist knowledge to assess them, but other reviews can assess validity. My focus is on the scientific applications, assuming the methods to be sound.
    i. **New figures outlining the workflow in more detail are added to the revision (Figure 1 and Figure 3). There have also been modifications to the methods section to clarify how each step is working. I have also uploaded a sample dataset and input files to github along with samples of the outputs. The inclusion of the sample set will allow users to try the program and provide a template for the data they wish to use.**

e. Based on the novel analysis and case-study of spiders, the author makes helpful suggestions on limitations, potential improvements and applications of the method on databases more generally. There are potential applications of this tool, so long as one remains aware of the taxonomic limitations. I think it could

more rapidly alert people to the spread of harmful invasive species such as the false widow Steatoda nobilis, by flagging images for expert validation - perhaps including those from newspaper archives. It could detect other range-expansions of conservation or ecological interest. From sets of images before and after a disturbance it could perhaps examine shifts in the balance of taxa - for example declines in orb or scaffold web building spider families when vegetation structure is simplified. Other applications will doubtless be discovered. There are complimentary proposals for those designing citizen science studies, such as what data are most valuable or most essential to keep records consistent and accessible.

i. **Thank you for the excellent suggestions for how this pipeline can be utilized in the future. I have addressed possible future uses in the Discussion. Also, thank you for seeing the broad potential for this pipeline.**

f. A few minor issues in presentation:

i. I suggest the Abstract be simpler and more like the end of the Discussion. Some of the punctuation could be improved, most importantly in the Abstract, and there are a few typos (eg large should be larger in the geo-tagging methods section; "was" should be were in Figure 3 caption). I suggest clarification in the Abstract: "genus level and the highest image quality according to the BRISQUE scores" should presumably be: genus level and had the highest image quality according to the BRISQUE scores. Perhaps one instance of "observations taxonomic" should be observations' taxonomic in the discussion. For me, the PDF preview and download had a formatting problem which cut off parts of some figure captions. Please describe what a 'pipeline' is for the novice!

1. **Thank you for bringing these to my attention. The errors have been corrected in the revision.**