Leveraging Expert Analysis to Remove Duplicates in GLOBE Mosquito Habitat Data

Isabella Levine, Kevin Balagtas, Mireya Galván, Lauren Farrar, Jose Hidalgo

SEES Mentors: Kavita Kar, Faguni Gupta, and Vishnu Rajasekhar

NASA SEES Mosquito Mapper Virtual Internship

**Table of Contents**

Abstract …………………………………………………………………………………......3

Introduction ............……………………………………………………………………………..4

Methodology ………..…………………....……………………………………………………....5

Results ……………....………………………………………………………………...…...6

Discussion …………………………………………………………………………………..…7

References ………………….…………………………………………………………....…….9

Acknowledgments….…………………....…….………………………………...…………....….10

**Abstract**

In the past two decades, citizen science has gained recognition due to its benefit and cost-effectiveness, and it has shown its potential to harness observations in wide geographies and over long time periods. Citizen science does, however, need to be verified by researchers hoping to use that data in the future. Within the GLOBE Mosquito Habitat data set, there is a concern of duplicate records, and this research explores the undertaking of removing those duplicates while creating a roadmap for future removal of duplicate data. This research is grounded in the theory of expert review scoring, as all reviewers underwent 10 weeks of NASA Earth Systems training, and this method of scoring was by the law of averages.

*Keywords:* Citizen Science, duplicates, Mosquito habitat

**Introduction**

In the past two decades, citizen science has gained popularity due to its benefit and cost-effectiveness (Aceves-Bueno et al., 2017). Citizen science allows scientists to tackle research questions with large spatial and temporal scales. Although it has gained recent popularity, most are still unsure of the accuracy of citizen science data. For example, 16 case studies reviewed by Finn Danielsen, Neil Burgess, and Andrew Balmford in 2005 “only provided cautious support for volunteers’ ability to detect changes in populations, habitats, or patterns of resource use” (Danielsen et al., 2005, Aceves-Bueno et al., 2017).

Another obstacle citizen scientists face is duplicate data. Duplicates are a major issue, as they can occur with all types of data. An effective method of identifying duplicate data is by focusing on images. As experiments have shown, determining whether or not two images are duplicates is a subjective task. Duplicates are such if they have the same content and composition, and were taken from the same angle and range (Jaimes et al., 2003). However, images do not have to be identical to be duplicates. The differences between the images determines whether or not they are duplicates. Differences between the images can be characterized in terms of the scene (what is being shot), the camera (the device used to take the photo), and the image (the final product) (Jaimes et al., 2003). These characteristics can help identify duplicate data within a set of images.

This study was undertaken in order to create a replicable protocol for removing duplicates in GLOBE mosquito habitat citizen science data. This research analyzed seven separate data sets with seven expert observations to identify duplicates and assign confidence scores.

**Methodology**

 Duplicates of images can skew the larger data set in a certain direction, thus they must be removed. In this project, the term “duplicate” would pertain to any data captured with more than 10 overall entries that shared the same MGRS Latitude and MGRS Longitude, same date, and same larvae count.

The data examined was from the GLOBE Observer app, or Global Learning and Observations to Benefit the Environment Program. The app uses citizen science to gather large amounts of data from people across the globe. The mosquito observations are collected by taking a picture of the site, counting the number of eggs, adults, pupae, and larvae, and then identifying the species. The data was then examined by experts in NASA SEES, or STEM Enhancement in Earth Science. The experts had been through 10 weeks of mosquito habitat training.

The initial data had 16,910 rows, however only data from North America was used. The data included information about the date measured, the date updated, date published, ID, organization, elevation, longitude and latitude, country, larvae number, egg number, adult number, and species. The data also included URLs to images of the mosquito site. The data was then searched for duplicates using Python and separated into duplicate instances.

There were 7 instances of duplicates found from the script. A thorough expert inspection was enacted to understand whether the data found was truly duplicated. Beyond a visual inspection, an image matching protocol was followed. The protocol started with an expert review, where each expert had to rate each instance on how confident that the images were from the same site, on a scale from 1-10, and were eventually averaged to provide a confidence score to substantiate the meta data.

The correlation in wording and scoring of data sets resulting from both the expert review and the objective review can be evidenced as two effective methods in gaining a general aspect of separating duplicate data from non-duplicates. The expert review detection method had individuals score images without output from other experts, yet showed a relatively close range in scores from 1-10 with only one outlying data set. The scale from 1-10 was utilized by the experts in this method, with a score of 1 symbolizing the least likely for the images to be duplicates and a score of 10 symbolizing that the images were the most likely to be duplicates. This method of scoring was founded in strong literature supported by the law of averages. The second method, employing an objective review, also showed a strong similarity in the experts’ descriptions of each image in their folder. The images with either the most or least of the same characteristics described by the experts tended to correspond to the average score given in the expert review.

**Results**

After comparing the scores from the expert review and the objective review, the results yielded strong similarities in the duplicate images. Beginning with the expert review, the scores that were set individually by each expert amounted to an average standard deviation of 1.3 between each data set, excluding the fourth data set. The low standard deviation showed a similarity in the visual perception of the supposed matching images by the experts. The fourth data set had a wide variety of scores, ranging from 3-10, being the only one with a larger standard deviation than 1.3. Because the fourth data set was significantly different than the others in the scale of its scoring, further discussion and analysis would be needed. Having multiple similarities in the data sets scored by human experts can be utilized to give a general perspective of which data sets are more likely or less likely to be duplicates.

|  |  |  |
| --- | --- | --- |
|  | Averages of Scores (by folder) | Standard Deviation of Scores (by folder) |
| Duplicate Image Folder 1 | 7.29 | 1.25 |
| Duplicate Image Folder 2 | 9.00 | 1.00 |
| Duplicate Image Folder 3 | 9.43 | 0.98 |
| Duplicate Image Folder 4 | 6.71 | 2.75 |
| Duplicate Image Folder 5 | 2.71 | 1.11 |
| Duplicate Image Folder 6 | 2.29 | 1.11 |
| Duplicate Image Folder 7 | 8.14 | 0.90 |

**Discussion**

The protocol we established was built for future evaluations to identify duplicate data points in a data set. This has established a path for future scientists to begin replicating the procedure on future data. This method integrated examining the scene, the camera, and the final product with each image in a data set to determine its likelihood of being a duplicate.

During our time researching and analyzing data, we used a threshold of more than 10 duplicates in the dataset, and that number could be lowered in the future. We used a method of analyzing images to confirm if they had been duplicated.

Citizen science is the future of all spectrums of science; all of the information that was previously inaccessible to one scientist alone can be gathered quickly by many volunteers in different locations, environments, and times. Using data collected by citizen scientists, large research questions can be answered using different perspectives and conditions, giving the field a wide range of applications.

Although some scientists have expressed that citizen science is sometimes unreliable, our protocol fills the gap in citizen science when duplicate data could possibly be collected. The probability of duplicate data being produced is high in all fields of science. With the proper tools and methods the number of duplicates could be minimal. The cost of citizen produced data is relatively inexpensive to collect.

**References**

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017, 9 29). *The Accuracy of Citizen Science Data: A Quantitative Review*. Ecological Society of America. Retrieved 7 26, 2021, from https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/bes2.1336

Danielsen, F., Burgess, N. D., & Balmford, A. (2005, 10). *Monitoring Matters: Examining the Potential of Locally-based Approaches*. Springer Link. Retrieved 7 26, 2021, from https://link.springer.com/article/10.1007/s10531-005-8375-0

Jaimes, A., Chang, S.-F., & Loui, A. C. (2003, 12 18). *Detection of Non-Identical Duplicate Consumer Photographs*. Retrieved 7 26, 2021, from https://drive.google.com/file/d/1iaoWLO\_1nQgYpssY09SJWjc16L3Lig9F/view

**Acknowledgements**

We would like to express our gratitude to the SEES program for allowing us this amazing opportunity and teaching us so much!

Thank you to the Earth System Explorers/Mosquito Mappers mentors Rusty Low, Peder Nelson, Cassie Soeffing, Erika Podest, Becky Boger, Surya Khudan and peer mentors Kavita Kar, Faguni Gupta, Vishnu Rajasekhar, Pratham Barbaria, and Matteo Kimura.

This project included; Kavita Kar, Faguni Gupta, Vishnu Rajasekhar, Isabella Levine, Kevin Balagtas, Mireya Galvan, Lauren Farrar, Jose Hidalgo

Mireya Galvan wrote the discussion section of our research paper. Mireya also contributed her expert review on photos to determine their likelihood of being duplicates in a data set. Mireya learned that not only in the study of mosquito habitats but in all fields of science having duplicates in data is highly likely, and an issue. Additionally, Mireya learned the importance of citizen science and its benefits on all areas of science.

Lauren Farrar wrote the introduction section of our research paper. Her section included research on citizen science and its growing popularity, and the problem of duplicate data in research. Lauren also provided her expert review on photos to determine their likelihood of being duplicates. Lauren learned that only a small percentage of citizen scientists are able to detect changes in populations, habitats, or patterns of resource use.

Jose Hidalgo contributed to our project by creating our powerpoint presentation and utilized his technology skills to put together our presentation video. Jose learned a lot about duplicates and why they are so harmful to data.

Isabella Levine wrote the methodology part of our paper that included a detailed description of how we went about performing our research and analysis of duplicate data. She also provided her expert review on photos to determine their likelihood of being duplicates. Isabella learned the Importance of citizen science and how it can help scientists get data on a very large scale.

Kevin Balagtas gathered and wrote the results of our duplicate data analysis. Kevin also contributed his expert review on photos to determine their likelihood of being duplicates. Kevin learned to analyze sets of data from our own mosquito trap projects to create an accurate depiction and help contribute to not only our team, but the scientific community as a whole with our results.