International Cartographic Association
Association Cartographique Internationale

Auto-filtering validation in citizen science biodiversity monitoring: a case study

Maryam Lotfian [ab] *, Jens Ingensand [b], Olivier Ertz [b], Simon Oulevay[b], Thibaud Chassin[bc]

[a] *Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy; maryam.lotfian@polimi.it*
[b] *University of Applied Sciences and Arts Western Switzerland, Yverdon-Les-Bains Switzerland; [firstname.lastname]@heig-vd.ch*
[c] *EPFL, LASIG, CH-1015 Lausanne, Switzerland; thibaud.chassin@epfl.ch*

* Corresponding author

Abstract: Data quality is the primary concern for researchers working on citizen science projects. The collected data by citizen science participants are heterogeneous and therefore must be validated. There are several validation approaches depending on the theme and objective of the citizen science project, but the most common approach is the expert review. While expert validation is essential in citizen science projects, considering it as the only validation approach can be very difficult and complicated for the experts. In addition, volunteers can get demotivated to contribute if they do not receive any feedback regarding their submissions. This project aims at introducing an automatic filtering mechanism for a biodiversity citizen science project. The goals of this project are to first use an available historical database of the local species to filter out the unusual ones, and second to use machine learning and image recognition techniques to verify if the observation image corresponds with the right species type. The submissions that does not successfully pass the automatic filtering will be flagged as unusual and goes through expert review. The objective is on the one hand to simplify validation task by the experts, and on the other hand to increase participants' motivation by giving them real-time feedback on their submissions. Finally, the flagged observations will be classified as valid, valid but uncommon, and invalid, and the observation outliers (rare species) can be identified for each specific region.

Keywords: Citizen Science, data validation, machine learning, biodiversity, automatic filters

1. Introduction

Citizen science, or public participation in scientific projects, is a term defined as the participation of non-professionals in scientific activities(Cohn, 2008). The history of citizen science goes back to the 1900s, the project Breeding Bird Survey (BBS) (www.pwrc.usgs.gov/bbs). The initial point for citizen science was environmental fields but with developments in mobile technology especially over the past few years, citizen science has expanded to many disciplines resulting in hundreds of citizen science mobile or web applications (Schade and Tsinaraki, 2016) in different fields such as astronomy (Lintott et al., 2008), medicine (Curtis, 2015), etc. Accordingly, the number of volunteers participating in citizen science projects has increased significantly. Citizen science participants are coming from diverse scientific backgrounds, age, culture, expertise, etc., therefore, the collected data in such projects are quite heterogeneous. As a result, data quality and validation play an essential role in any citizen science project (Dickinson et al., 2010). To ensure data quality various number of validation mechanisms have been used in different citizen science projects. Wiggins et al. (Wiggins et al., 2011) have done a survey to discover the most used validation approaches among various citizen science projects. In their survey, 840 emails were sent, among which 63 surveys were completed with a response rate of approximately 8%. As a result of the survey, a list of different validation approaches used in these projects was produced. The results illustrated that the most common data validation methodology with 77 percent among all the surveyed projects was expert review.

Expert review usually is done several months after data submission, when the data is going to be utilized for further analysis (David N. Bonter and Cooper, 2012). As a result, the participants miss the opportunity to receive quick feedbacks on their submissions, which might result in demotivating them in their contribution. In addition, expert validation can be overwhelming and time consuming with massive quantities of observations at the end (Kelling et al., 2011). Therefore, auto-filtering or timely screening of the submissions seems to be a solution to address the latter issues in validation of user-generated data in citizen science projects. There are few citizen science projects, which use automatic data verification approaches. The two main well-known ones are "Project FeederWatch (PFW)" (Bonney and Dhondt, 1997) and "eBird" (Wood et al., 2011). Researchers of PFW have used historical data to create a checklist of "allowable" bird species for each US state and Canadian province, and a "smart filter" then is developed based on the defined checklist (David N Bonter and Cooper, 2012). The smart filter could get violated in two situations: 1) the observation does not appear in the predefined checklist, and 2) the number of reported species exceeded the maximum allowed counts depending of the

Proceedings of the International Cartographic Association, 2, 2019.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan. This contribution underwent
single-blind peer review based on submitted abstracts. https://doi.org/10.5194/ica-proc-2-78-2019

species type, month and region. If a submission is flagged as unusual after passing the smart filter, the participant receives a warning message and is asked to either confirm or change the data entry.

Project eBird is using a similar approach as PWF to detect outliers or unusual observations but with an additional step before flagging the submissions (Kelling et al., 2011). In this project, the observations are treated in two steps: first, the unusual observations are detected based on the historical database (as in PFW), and second, a machine learning approach is used to classify the observers based on their expertise to experts or novices. Finally, the classification of observers is used to decide whether the unusual observations should be flagged or not. If a submission dose not pass successfully the first step, then it will be flagged as unusual only if the observer is a novice.

This article aims at introducing a citizen science project to collect biodiversity observations with an automated validation functionality. In this project, three automatic filtering mechanisms have been proposed. The first two methods, similar to the approaches used in the mentioned projects above, are using an available database of species to check the validity of observations regarding the time and location of data entry. The third method is using machine learning and image recognition techniques to verify whether the submitted images of the species corresponds with the right selected species type. Among the three auto-filtering approaches, the utilization of machine learning techniques is drawing our attention and it is our central focus in this project. In other words, , our main goal is to gather evidence if machine learning on the one hand simplifies validation task and improves the results of a citizen science project in terms of data quality and on the other hand helps motivating citizens to continue contributing to a project. Moreover, if combining machine learning techniques with citizen science proves successful in the context of our particular project, we would like to investigate whether or not machine learning could be useful to other citizen science projects as well.

The remaining of this paper is structured as follows: in the first part, the combination of the two concepts of citizen science and machine learning is discussed. Afterwards, our previous work, which leaded us to develop this citizen science project is being presented followed by introducing the current project explaining the three proposed automatic filtering approaches. Finally, the discussion and future perspectives are presented.

2. Machine learning and citizen science

Volunteers have always been the core of citizen science projects to help the researchers by collecting or post-processing data. Nowadays with the advances in the artificial intelligence and machine learning this question arises that if computers can be banded together with volunteers and scientists to help them out with collection and analysis of large amount of data in citizen science projects? The combination of citizen science and machine learning has been discussed in few papers during the past few years (Beaumont et al., 2014; Keshavan et al., 2018; Sullivan et al., 2018). A very recent project led by the University of Minnesota, has claimed that computer is the new partner of citizen science projects (University of Minnesota, 2019). In this research, data scientists and citizen science experts have collaborated with ecologists interested in studying wildlife by establishing hidden cameras in the nature to collect images of passing wild species. The photos then will be classified depending on their study' goals using machine learning algorithms, and volunteers will be involved in classifying only the photos that will be hard for the computers to identify. They concluded that this approach can help the researchers to speed up the data processing, and reduce the tasks of volunteers in citizen science projects by allowing them to concentrate only on the rarer and more complicated classifications. While this is an interesting project, our goal in this paper is to use machine learning' techniques not to replace the role of participants, since they are the heart of citizen science projects, but to work in parallel with them, and to facilitate the validation or data collection tasks in these projects.

3. Previous work

BioSentiers is a project developed by the development and GIS team of the university of Applied Sciences and Arts, Western Switzerland in collaboration with the University of Teacher Education Vaud. This project aims at raising children's knowledge about biodiversity as well as connecting them with the nature through an augmented reality mobile application (Ingensand et al., 2018). In this project, the pupils can observe the natural environment through the camera of a smartphone or tablet, with the virtual elements representing the biodiversity Points of Interest (POI) overlaying on the screen (Figure 1). By tapping on the virtual biodiversity POIs on the screen, the pupils are guided about their distance from the actual species in the nature. As they get close enough to the species, they tap once more on the virtual icon and a page with information about the species opens. Finally, they can read the information given in the application and compare different features with the species in the nature. This project was tested in November 2017 with 15 pupils aged around 9-12 years old, and the results were presented in AGILE conference 2018. One interesting feedback after presenting this word was how we update the POIs and how to collect more new species. This question was an inspiration for us to develop a citizen science project to first validate the already existing POIs or in other words updating BioSentiers database by verifying if the species are still available or not (which is valid mainly with respect to trees or flowers), and second to collect new observations. The following section introduces the citizen science project, with the proposed validation approaches.

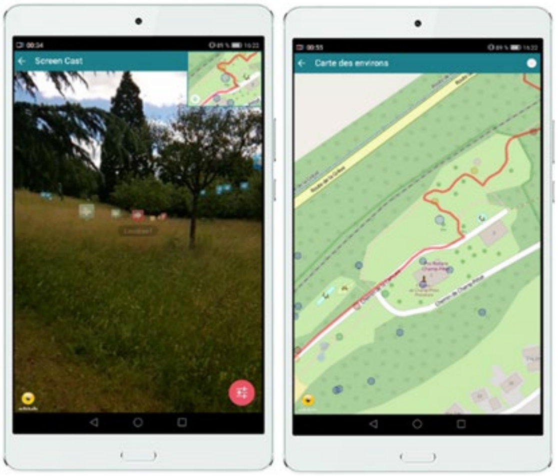

Figure 1: Screenshot of the AR view (left) and 2D map (right) of the BioSentiers application

4. BioSenCS

BioSenCS is a citizen science project to collect biodiversity observations with real-time validation functionalities through a web application. While the initial biodiversity points of interest (POI) for BioSentiers application were collected by biologists, BioSenCS is designed with the objective of updating them as well as collecting new observations by citizens not necessarily experts. In this work, three automatic validation mechanisms are proposed: first, checking the periods, which the species are visible, second checking the common areas, which the species exist, and finally using a machine learning API called Clarifai (https://clarifai.com/). The objective here is on the one hand to increase participating motivation by giving real-time feedbacks to volunteers, and on the other hand to simplify end-validation task (expert review) in our project.

The general workflow of the automatic validation is as follows (Figure 2): First, the user makes a submission in our online platform, the submission goes through the automatic filtering and in case of no warning it successfully passes the filter as valid observation. However, if a warning occurs during the automatic filtering procedure then the submission is flagged as unusual and the user will receive the first feedback. In this stage, the user has two possibilities, one is to modify the observation and resubmit and the other is to force-confirm the submission without making any changes. In case of confirming the submission, the observation is automatically forwarded for expert review. The experts in this project are the local biologists who has defined the initial database for BioSentiers project. In this part, the experts can verify whether or not the flagged observations are valid, valid but rare or invalid. Moreover, in case more information regarding the species is required by the experts (before assigning the observation to a category), the volunteers then will receive a second feedback, which is asking them to provide more supporting documents such as more images or reports regarding the species physical characteristics (e.g. color, size, etc.). BioSenCS is being developed in the Django Web framework using PostGIS for the persistence of data. The four main categories of interest of species at this phase of the application are tree, flower, butterfly, and bird.

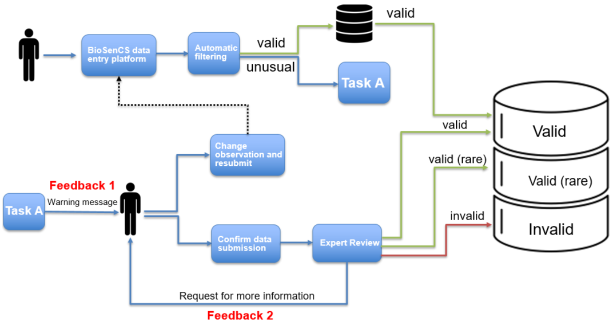

Figure 2: General architecture of applying automatic validation to collected biodiversity observations in BioSenCS platform

4.1 Visibility period filter

In the database of BioSentiers, biologists have defined the periods where the species are visible. When a participant makes a submission the auto-filter verifies whether the observation time matches with the visibility period of the specific species or not. If the observation time is not within the predefined period then the participant will receive a warning message indicating that the observation does not seem to be present in this season/time, and is requested to make sure the submission is done correctly, e.g. controlling the selected species name/type. Figure 3 illustrates the procedure to control the validity of the observation depending on its visibility period.

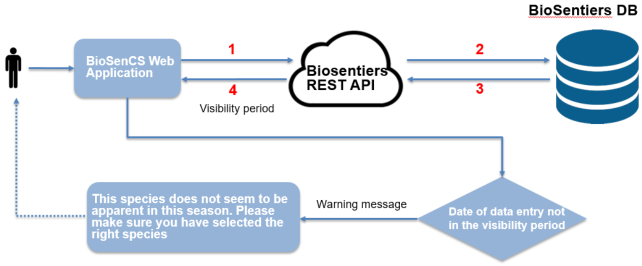
BioSentiers DB
BioSenCS Web Application
1
2
3
4
Biosentiers REST API
Visibility period
This species does not seem to be apparent in this season. Please make sure you have selected the right species
Warning message
Date of data entry not in the visibility period

Figure 3: architecture of submission auto-filtering based on species visibility period

4.2 Common location filter

Our study area is composed of eight zones with four different landcover types including urban zone, park, natural wetland and natural forest. Depending on the zone and its landcover characteristics, the species living in one zone differ from one another. Considering this fact, to explore the spatial distribution of the species in the region of interest, a cluster map for each specific species type is generated using the POIs in the BioSentiers database indicating common habitats for the species. Submissions out of these common areas are considered as unusual, and the participant will be asked to ensure the observation is correctly added or the location is correctly selected. In addition to the cluster map, the participants are not allowed to submit observations located out of the eight regions/zones of interest. Likewise, observation entries that cannot be logically or geographically valid are discarded e.g., a tree in a lake. Figure 4 illustrates the procedure to control the validity of the observation depending on their location.

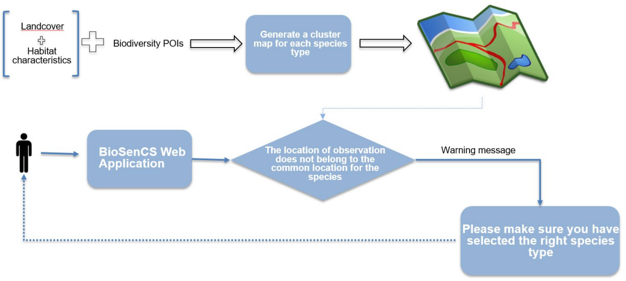
Landcover
+
Habitat characteristics
Biodiversity POIs
Generate a cluster map for each species type
BioSenCS Web Application
The location of observation does not belong to the common location for the species
Warning message
Please make sure you have selected the right species type

Figure 4: Architecture of auto-filtering validation based on species common location

4.3 Machine learning filter

Participants are requested to select a species type before uploading the observation photo. This approach, which is the main focus of this work, uses image recognition techniques to verify if the species type selected matches with the uploaded observation photo. In this project (for the initial phase), Clarifai API has been used. Clarifai was used here due to the fact that it has been mentioned as a powerful computer vision tool in some scientific articles (Szegedy et al., 2015; Wu et al., 2015). The API processes the submitted photo of the species and generates a set of prediction tags with their probabilities using computer vision algorithms. BioSenCS checks the tags with the probability higher that 90 percent, and if the species type does not belong to any of the predicted tags for the mentioned probability threshold, the participant will receive the warning message, and is asked to verify if the species type is selected correctly. Figure 5 illustrates the procedure to control the validity of the observation photo and species type using Clarifai trained models.

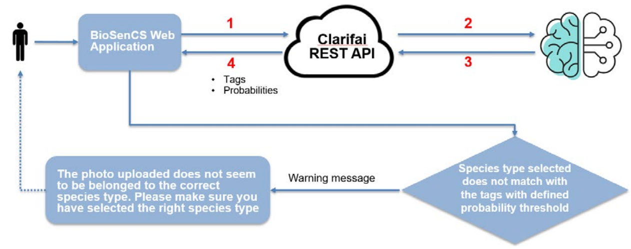
BioSenCS Web Application
1
2
Clarifai REST API
4
3
• Tags
• Probabilities
The photo uploaded does not seem to be belonged to the correct species type. Please make sure you have selected the right species type
Warning message
Species type selected does not match with the tags with defined probability threshold

Figure 5: Architecture of machine learning auto-filtering validation using Clarifai API

5. Discussion

The approaches presented here are considered as our initial hypotheses, which must be verified when the application is ready to be tested. The first hypothesis is that *automatic filtering reduces the amount of erroneous observations* since the warning messages sent to the participants will give them the opportunity to modify and correct wrong submissions resulting in more simplified end validation task. The second hypothesis is that *real-time feedbacks will keep the volunteers motivated to contribute and sustain their participation*. The third hypothesis is that *this automatic validation helps volunteers to learn better about different species while contributing to a particular project*. Moreover, the last hypothesis is that *the automatic filter helps us to easier categorize rare species and to define the outliers in our observation database*. For instance if a particular bird is observed in a region which is not common, there will be records which are of utmost importance for biologists or bird watchers. As another example, if a tree is observed in a zone surrounded by all different families of trees, it must be considered as an outlier in that region. Another point to be investigated is the probability threshold defined for Clarifai API. The threshold is selected based on the trend of checking test images, but there is no scientific reasoning behind it. Therefore, we would like to understand if that is the right way to flag a submission or not. Moreover, we would like to explore the output of the machine learning validation not only with extreme images (images containing only one species), but also with images that contain several species but the participant aims at submitting one particular and has selected the species type accordingly. For instance, if the volunteer has selected the species type as bird, and uploaded a photo, which contains a bird sitting on a tree with the ground covered with flowers. If the model gives a probability less than 90% to this submission the user will get a false warning. Therefore, the question here is if the selected probability threshold is the right way to validate the images or not since the main objective is to simplify not complicate the validation task in citizen science.

6. Summary and future perspectives

Ensuring data quality has always been a critical issue when dealing with user generated contents, and specially citizen science projects. Data validation can be an overwhelming and time consuming task if it is only done by expert controls in the absence of any other validation approaches. This paper has presented an automatic filtering procedure of the collected biodiversity observations for our citizen science project. In this project, three auto-filtering mechanisms has been proposed. Therefore, we are going to verify if the auto-filtering approaches, specifically machine learning, would help us on one side to attract the volunteers to contribute and sustain participation, and on the other side to reduce erroneous observations and facilitate expert review at the end. Moreover, another focus of this work is to explore whether the combination of machine learning and citizen science can result in simplifying the data collection and validation tasks. The future objective of this work is to understand how the combination of the three proposed filters can be applied in the application, and if they must be prioritized. In addition, in this work the level of expertise of the participants is not taken into account and the data entry from all the users is being treated equally. Therefore, a classification of the participants depending on their expertize or their level of contribution can be an interesting area for future studies of this work. Finally, the region of interest for this project is of small scale, and the goal is to investigate if this particular approach presented here will work for large regions as well.

7. References

Beaumont, C.N., Goodman, A.A., Kendrew, S., Williams, J.P., Simpson, R., 2014. The milky way project: Leveraging citizen science and machine learning to detect interstellar bubbles. Astrophys. Journal, Suppl. Ser. 214.

Bonney, R., Dhondt, A.A., 1997. FeederWatch. In: Cohen, K.C. (Ed.), Internet Links for Science Education: Student---Scientist Partnerships. Springer US, Boston, MA, pp. 31–53.

Bonter, D.N., Cooper, C.B., 2012. Data validation in citizen science: A case study from Project FeederWatch. Front. Ecol. Environ. 10, 305–307.

Cohn, J.P., 2008. Citizen Science: Can Volunteers Do Real Research? Bioscience 58, 192–197.

Curtis, V., 2015. Motivation to Participate in an Online Citizen Science Game: A Study of Foldit. Sci. Commun. 37, 723–746.

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010. Citizen Science as an Ecological Research Tool : Challenges and Benefits.

Ingensand, J., Lotfian, M., Composto, S., Ertz, O., Oulevay, S., Oberson, M., Da Cunha, M., Piot, D., 2018. Augmented reality technologies for biodiversity education – a case study. 21st Int. Conf. Geogr. Inf. Sci. (AGILE 2018) 1–5.

Kelling, S., Yu, J., Gerbracht, J., Wong, W.K., 2011. Emergent filters: Automated data verification in a large-scale citizen science project. Proc. - 7th IEEE Int. Conf. e-Science Work. eScienceW 2011 20–27.

Keshavan, A., Yeatman, J.D., Rokem, A., 2018. Combining citizen science and deep learning to amplify expertise in neuroimaging 1–21.

Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P., Vandenberg, J., 2008. Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Mon. Not. R. Astron. Soc. 389, 1179–1189.

Schade, S., Tsinaraki, C., 2016. Survey report : data management in Citizen Science projects.

Sullivan, D.P., Winsnes, C.F., Åkesson, L., Hjelmare, M., Wiking, M., Schutten, R., Campbell, L., Leifsson, H., Rhodes, S., Nordgren, A., Smith, K., Revaz, B., Finnbogason, B., Szantner, A., Lundberg, E., 2018. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nat. Biotechnol. 36, 820.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going Deeper With Convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

University of Minnesota, 2019. Citizen science projects have a surprising new partner, the computer: New machine learning techniques could pave the way for larger, more timely projects.

Wiggins, A., Newman, G., Stevenson, R.D., Crowston, K., 2011. Mechanisms for data quality and validation in citizen science. Proc. - 7th IEEE Int. Conf. e-Science Work. eScienceW 2011 14–19.

Wood, C., Sullivan, B., Iliff, M., Fink, D., Kelling, S., 2011. eBird: Engaging birders in science and conservation, PLoS Biology.

Wu, R., Yan, S., Shan, Y., Dang, Q., Sun, G., 2015. Deep Image: Scaling up Image Recognition.