An Interactive System for Intrinsic Validation of Citizen Science Data for Species Distribution Mapping and Modelling Applications

Hossein Vahidi*’**[1], Brian Klinkenberg**, Wanglin Yan*

* EcoGIS Lab., Keio University, Japan
** Lab for Advanced Spatial Analysis, University of British Columbia, Canada

Abstract. This paper presents a conceptual model for assurance of the quality of species occurrence observations in citizen science projects. To this end, we adopted the notion of trust as an indicator of VGI quality and defined the concept of trustworthiness of VGI as a function of four main contexts: consistency with habitat, consistency with neighbors, consistency with the temporal life cycle of the species, and the competence and credibility of volunteers. In this sense, the quality of an observation is quantified in terms of the level of the trustworthiness of the submission by using fuzzy set theory. Moreover, the different possible ex-post and ex-ante architecture of the proposed system is briefly discussed to empower the end user (data consumer), expert reviewers, and volunteers (data producers) to perform more robust and precise VGI quality assurance practices. Finally, the paper ends with concluding remarks and some tips for future research directions.

Keywords. Volunteered Geographic Information (VGI), Citizen Science Data Quality Assessment, Species Distribution Mapping and Modelling

1. Introduction

In the current research trend on volunteered geographic information (VGI), much concern has been directed to the issue of data quality and validation of the crowdsourced data, focused on the issues of accuracy, credibility, and the possibility of vandalism in the crowdsourced data. The issue of VGI data quality assessment and evaluation has become very sensitive and vital, as we usually do not have access (or we have limited access) to the real-world data to collect the ground truth when the geographical and temporal dimensions of the data are relatively large.

In the context of ecological studies, the habitat of a species (plant or animal) is mapped and modelled by using the collected data on the species’ presence (i.e., occurrence data), which include a sample of locations with known presence of the target species (Merow et al. 2013).

Conventionally, presence records are collected by experts and authoritative sectors to ensure the quality of the data; however, recently, because of the popularity of citizen science programs in ecological studies, non-experts are also enabled to contribute to the process of data collection through participation in VGI and citizen science projects for biodiversity/conservation observations (e.g., E-Flora, iNaturalist, and eBird). Legions of citizen scientists record their

[1] Corresponding author (E-mail: vahidi@sfc.keio.ac.jp)

LBS 2016
13th Conference on Location-Based Services
Vienna, 14–16 November 2016

Published in “Proceedings of the 13th International Conference on Location-Based Services”, edited by Georg Gartner and Haosheng Huang, LBS 2016, 14-16 November 2016, Vienna, Austria.

observations of all types of species by identifying and recording the location of the observed species as well as other relevant attributes and meta data about the target species (e.g., taxon, time of observation).

Various factors, such as uncertainties in the measurement of the spatial component of observations by amateur naturalists, the complexity of taxonomy and species identification by non-experts, different understandings of the concept of quality from experts' and non-experts' points of view and diverse motivations, and the knowledge level and background of the participants in citizen science projects, may trigger questions about the quality of VGI data in biodiversity observation projects (Ali and Schmid 2014). To control the quality of the crowdsourced data in the context of citizen science projects, several approaches have been implemented and tested in previous practices.

Cross-referencing the VGI on species presence with the existing authoritative data, checking the consensus and agreement of reports at each location, and expert or community-based (by the participation of other volunteers) data quality control of the user submissions are the most popular approaches adopted for data quality assurance in citizen science biodiversity observation projects (Goodchild and Li 2012).

Nevertheless, the lack of authoritative data in the biodiversity domain and the existing uncertainties within the available datasets usually may avoid the use of the cross-referencing method in most VGI activities. However, when the spatio-temporal extent and the diversity of species are relatively large and there is a large data space in comparison with the number of active volunteers, not all localities in the study area may be observed frequently by the different participants, so the consensus-based approach to data quality assessment may not be applicable. Furthermore, the validation of all VGI submissions based on expert or community-based data quality control method could be very time and energy consuming (and costly in the case of using recruited experts) and mostly impractical, as the relative number of skilled human resources is usually limited, not all members of the community have the motivation or skill to participate in such a process, and sometimes there is not enough information for validation of the submission as well as the opportunity for in-situ data collection when it is applicable (particularly in wild and remote areas).

In this paper, we propose the general schema of an interactive system for intrinsic validation of VGI species occurrence datasets to reduce the dependence of data quality assessment processes on authoritative data as well as the participation of experts and the community in the process of VGI data quality control.

2. Related Work

One of the most common approaches for the assessment of the VGI quality is to compare VGI with ground truth reference datasets (i.e., authoritative dataset) (Barron et al. 2014). However, high-quality authoritative datasets for conducting extrinsic quality assessment are often not accessible/usable because of the lack of such data, costs, and licensing restrictions (Mooney et al. 2010) or the nature of the problem. Therefore, in cases where the direct cross-referencing approach is not applicable, researchers have explored more intrinsic approaches to evaluate the quality of VGI by using other proxies and indicators for quality measures (Senaratne et al. 2016).

Goodchild and Li (2012) discussed a geographic approach as an intrinsic VGI quality assurance method that relies on identifying rules that connect different information based on their location to evaluate whether an attribute of a VGI submission is reported correctly at a certain location.

The notion of trust has been used in a number of previous studies (Bishr and Janowicz 2010, Bishr and Mantelas 2008) as a proxy for data quality assessment of VGI contributions by making a link between the notion of spatial data quality and the established notion of interpersonal trust. It is expected that the trusted contributors provide more trustworthy VGI than less trusted ones (Yan et al. 2016). Thus, the trustworthiness of VGI can be substituted with traditional quality indicators of spatial data (e.g., completeness, logical consistency, and positional accuracy) (Yan et al. 2016), particularly when authoritative data are not available and the extrinsic quality assessment approach is not applicable.

Yan et al. (2016) presented a system to ensure the quality of VGI acquired for the means of species surveillance. In this system, they adopted trust as a proxy of VGI quality by defining trust as a function of the provenance of user expertise and the fitness of a submission according to geographic context. The quality of VGI is quantified in terms of the level of the trustworthiness of a submission by using fuzzy set theory.

Bordogna et al. (2016) broke down the existing adopted approaches for improving the quality of VGI by considering the time of their application as related to the time of VGI item creation into two main categories: ex-ante and ex-post. The former category refers to all quality assurance approaches that perform the quality improvement task before the creation (i.e., final submission) of VGI to prevent the creation of low-quality VGI. The latter category includes approaches in which the data quality improvement task is undertaken after the collection of VGI and a cleaning and enhancement activity is executed after the VGI is created (i.e., submitted) (Bordogna et al. 2014).

In the next section, we propose a conceptual framework for the intrinsic quality assessment of VGI in citizen science biodiversity projects using the trust notion as a proxy for data quality in the framework of both ex-ante and ex-post VGI quality management mainstreams.

3. Methodology and Conceptual Framework

A conceptual model is developed for the intrinsic quality evaluation of participants' observations in citizen science biodiversity projects.

In the simplest and non-interactive architecture of such a system, the observation records for a particular species are submitted to the system by non-expert volunteers who are participating in the citizen science project.

To ensure the quality of all the submitted data in database of such a project for the end user of the data (i.e., data consumers), we adopted the notion of trust as an indicator of VGI quality. In this context, we defined the concept of the trustworthiness of VGI as a function of four main contexts: consistency with habitat, consistency with neighbors, consistency with the temporal life cycle of the species, and the competence and credibility of volunteers (*Figure 1)*.

In the next sections, the four different components of VGI trustworthiness and the proposed rule-based fuzzy system to estimate the trustworthiness value will be briefly presented and the different possible architecture of the system will be discussed.

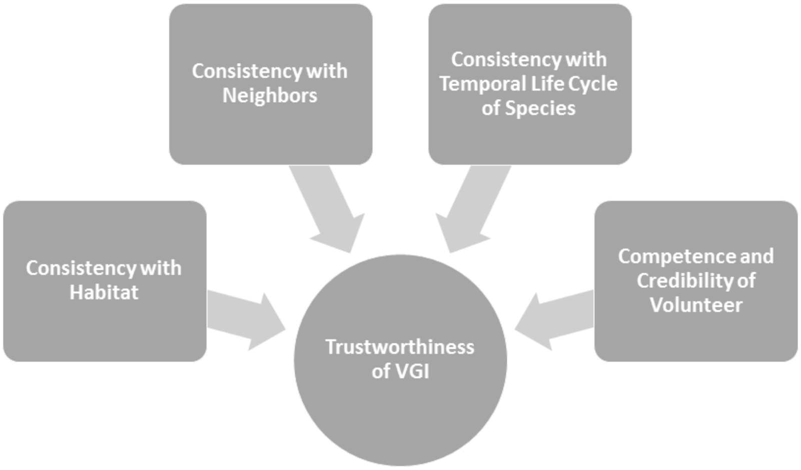
Consistency with Neighbors
Consistency with Temporal Life Cycle of Species
Consistency with Habitat
Competence and Credibility of Volunteer
Trustworthiness of VGI

Figure 1. Components for the assessment of trustworthiness of the VGI in the proposed approach

3.1. Components and General Schema of the Proposed System for the Quality Assurance of Crowdsourced Species Occurrence Observations

To assess the validity of an observation (i.e., a purported report on the observation of a particular species at a certain location) in the database via the "consistency with habitat" metric, the submission is compared with a generated reference layer that indicates the suitability of the landscape at each location to be inhabited by a particular species or demonstrates the probability of the occurrence of a particular species at a particular locality. The generated reference layer is estimated by using an ecological niche modelling approach on the basis of estimating the similarity of the environmental conditions at unknown localities in the landscape to the environmental conditions (e.g., temperature, precipitation, and topography) at the locations of known occurrence of a species (Hijmans and Elith 2016). The few required occurrence records for training the ecological niche model can be adapted from authoritative sources (e.g., herbarium datasets) or high-quality crowdsourced datasets from previous projects. By cross-checking a purported species taxon that was reported by a contributor with the estimated possibility (or probability) of the occurrence of the species in that location on the generated reference layer, one can evaluate whether the reported taxonomy of the submission is plausible.

Nevertheless, as the results of ecological niche modelling approaches are error-prone, the proposed conceptual model is empowered by other intrinsic contexts for the evaluation of VGI submissions.

It is widely known that the occurrence of a particular species in the geo-graphical space is the function of the environmental conditions at that location in the landscape. Furthermore, according to the first law of geography (Tobler 1970), "everything is related to everything else, but near things are more related than distant things." Hence, in the proposed conceptual model, the submissions that were tagged in the proximity of other submissions are considered more consistent with their neighbors and get higher scores.

Any species (plant and animal) has a particular temporal regime in its life cycle. For instance, it is widely known that flowers and seasons are closely related to each other and most flowers are season-specific, so they occur in specific time periods in the year. Or it is known that some animals go into hibernation during the cold winter months and wake up in the warm season. Thus, a purported observation has to be consistent with the temporal life cycle of the declared species.

The user's expertise and experience or the quality of previous contributions of a volunteer as well as the declared confidence in the quality of submissions by the volunteer could be indicators of the competence and credibility of the volunteer, and higher levels of these indicators increase one's expectation of higher levels of the trustworthiness of VGI.

Finally, to handle the uncertainties and ambiguities inherent in the four different trust indicators and to evaluate the quality of VGI in terms of the trust concept, we adopted a rule-based fuzzy system in the proposed conceptual model.

Figure 2 demonstrates the general schema of an automatic system for ensuring the quality of the observations of a particular species in a citizen science project. In this architecture, all the stored observations of the participants in the database of the system are evaluated in terms of the trustworthiness of records in an ex-post approach. Thus, if the trustworthiness of a record meets the minimum requirements of the end user (i.e., data consumer) that is characterized in terms of an acceptance threshold, it is reported to him/her as qualified data.

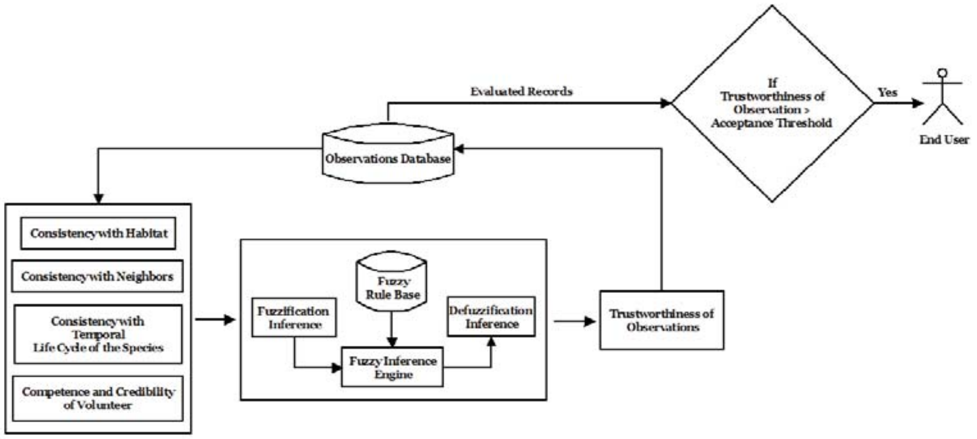

Figure 2. The proposed ex-post approach for the quality assurance of species occurrence observations in citizen science projects for the end user

3.2. An Ex-post Approach for Supporting Expert Reviewers in the Process of Quality Assurance

The proposed method can be utilized as a decision support system for empowering the expert reviewers who are in charge of the quality assurance task in a citizen science project. In the proposed ex-post approach, the trustworthiness of a single submission is evaluated, and if it does not meet the defined requirements of the experts, it is flagged for further manual review by the experts in the system (also, in a similar methodology, the estimated trustworthiness level of a submission can be used by the experts as an indicator of the quality of the VGI) (*Figure 3*). The suggested decision support system may help us to establish a semi-automatic VGI quality assurance system and reduce the workload of the expert reviewers and enhance the precision of the VGI quality assessment approach.

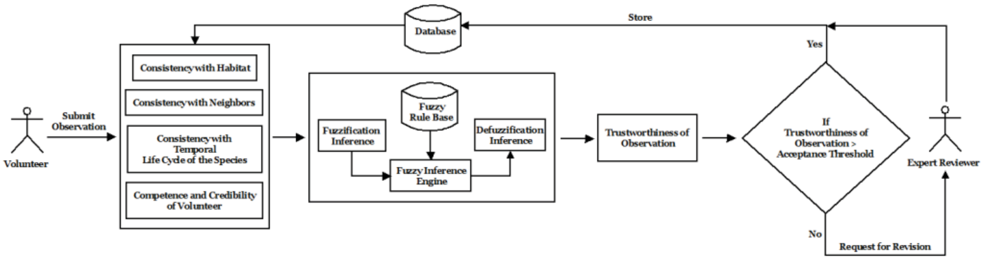
Store
Database
Yes
Consistency with Habitat
Consistency with Neighbors
Consistency with Temporal Life Cycle of the Species
Competence and Credibility of Volunteer
Submit Observation
Volunteer
Fuzzy Rule Base
Fuzzification Inference
Defuzzification Inference
Fuzzy Inference Engine
Trustworthiness of Observation
If Trustworthiness of Observation > Acceptance Threshold
Expert Reviewer
No
Request for Revision

Figure 3. The proposed ex-post approach for supporting the expert reviewers for the quality assurance of species occurrence observations in citizen science projects

3.3. An Ex-ante Approach for Supporting Volunteers in the Process of Quality Assurance

Figure 4 shows the schema of an ex-ante representation of the proposed model for the quality assurance of species occurrence observations in citizen science projects. In this architecture, upon the submission of an observation by a volunteer, the system evaluates the trustworthiness of the submission during the entity creation process and before storing it in the database of the system and alerts the volunteer (i.e., data producer) to revise the submission if it does not meet the defined acceptance threshold for an observation. The volunteer receives feedback from the system interactively that enables him/her to evaluate the validity of his/her submission and modify it if it is applicable. Moreover, such a system encourages the user to learn more about the living environment and increase his/her expertise by its indirect training process.

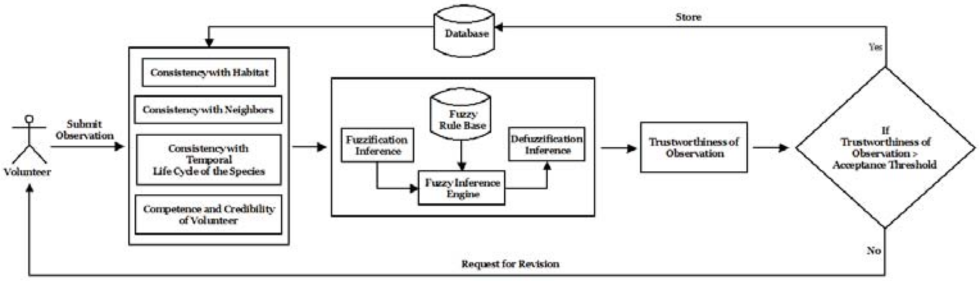
Store
Database
Yes
Consistency with Habitat
Consistency with Neighbors
Consistency with Temporal Life Cycle of the Species
Competence and Credibility of Volunteer
Submit Observation
Volunteer
Fuzzy Rule Base
Fuzzification Inference
Defuzzification Inference
Fuzzy Inference Engine
Trustworthiness of Observation
If Trustworthiness of Observation > Acceptance Threshold
No
Request for Revision

Figure 4. The proposed ex-ante approach for supporting the volunteer in producing qualified species occurrence observations in citizen science projects

4. Conclusion and Future Work

In this paper, we presented a conceptual model for the quality assurance of species occurrence observations in citizen science projects. To this end, we adopted the notion of trust as an indicator of VGI quality and defined the concept of the trustworthiness of VGI as a function of four main contexts: consistency with habitat, consistency with neighbors, consistency with the temporal life cycle of the species, and the competence and credibility of volunteers. In this sense, the quality of an observation is quantified in terms of the level of the trustworthiness of the submission by using fuzzy set theory.

Moreover, the different possible ex-post and ex-ante architectures of the proposed system were discussed to empower the end user (data consumer), expert reviewers, and volunteers (data producers) to perform more robust and precise VGI quality assurance practices.

The indicators of trustworthiness of VGI are not limited to the four aforementioned factors, so further investigation is required to define all the effective components of VGI trust in citizen science biodiversity observation projects.

Furthermore, in the proposed ex-post and ex-ante architectures for supporting the expert reviewers and volunteers in the process of VGI quality assurance, the completeness of the observation space is increased by increasing the number of submitted observations in the system over time. Thus, the consistency with neighbors rate for a submission might be changed by tagging more observations in the proximity of it. In addition, the first submitted observations may get a low value for the context of consistency with neighbors, as no observation was recorded in the proximity of them. Therefore, in future work, different options (such as enrichment of the data space by using existing authoritative sources or high-quality crowdsourced datasets from previous projects as well as defining the fuzzy rules in a dynamic approach) must be studied to address this issue.

In future steps, the two aforementioned ex-post and ex-ante architectures can be integrated with a recommender system to advise the expert reviewers and contributors to select the suitable taxon and scientific name for the observation at a certain locality.

References

Ali, A. L. and Schmid, F., 2014. Data Quality Assurance for Volunteered Geographic Information. *In:* Duckham, M.*, et al.* eds. *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings*. Cham: Springer International Publishing, 126-141.

Barron, C., Neis, P. and Zipf, A. 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS,* 18(6), 877-895.

Bishr, M. and Janowicz, K., Can we trust information?-the case of volunteered geographic information. ed. *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume*, 2010.

Bishr, M. and Mantelas, L. 2008. A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal,* 72(3), 229-237.

Bordogna, G.*, et al.* 2014. A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. *Information Sciences,* 258, 312-327.

Bordogna, G.*, et al.* 2016. On predicting and improving the quality of Volunteer Geographic Information projects. *International Journal of Digital Earth,* 9(2), 134-155.

Goodchild, M. F. and Li, L. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics,* 1, 110-120.

Hijmans, R. J. and Elith, J., 2016. *Species distribution modeling with R* [online]. Available from: http://www.idg.pl/mirrors/CRAN/web/packages/dismo/vignettes/sdm.pdf [Accessed 07.09.2016.

Merow, C., Smith, M. J. and Silander, J. A. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography,* 36(10), 1058-1069.

Mooney, P., Corcoran, P. and Winstanley, A. C., 2010. Towards quality metrics for OpenStreetMap. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems.* San Jose, California: ACM, 514-517.

Senaratne, H.*, et al.* 2016. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 1-29.

Tobler, W. R. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography,* 46(sup1), 234-240.

Yan, Y., Feng, C.-C. and Wang, Y.-C. 2016. Utilizing fuzzy set theory to assure the quality of volunteered geographic information. *GeoJournal*, 1-16.