# Towards the use of Citizen Sensor Information as an Ancillary Tool for the Thematic Classification of Ecological Phenomena

Laura Kinley

Nottingham Geospatial Institute, The University of Nottingham, United Kingdom
psxlk@nottingham.ac.uk

**Abstract.** The combination of Volunteered Geographic Information (VGI) with remote sensing classification techniques is addressed rarely, yet has masses of potential in the domain of improving data collection and annotation for environmental monitoring. This position paper delineates the benefits of using VGI within ecological research and identifies key research challenges in gathering ecologically robust data from citizens. The importance of VGI design and sampling typologies in understanding the patterns of and mechanisms of improving data quality are emphasised. Finally, future work in addressing the quality of crowd ground truth information for map generation in a remote sensing context is outlined with the hope that the traits of VGI can be aligned to meet the authoritative rigor required for it to be of use within ecological research applications.

**Keywords:** Volunteered Geographic Information, Spatial Data Quality, Biodiversity Distribution Mapping, Remote Sensing

## 1 Introduction

Ecologists and biogeographers face a great challenge in accurately examining the spatial distribution of vegetation assemblages over large spatial and temporal scales. Many traditional global, pan-continental and national land cover maps rely on the classification of satellite sensor imagery via remote sensing, which can be constrained by opportunities for error propagation and the time and expense required to collect training samples – it does not scale well [1]. The inconsistent and broad classification schemes used denote that a generalised view of species distribution is often generated, detrimental to our understanding and management of the environment. Subsequently, there is much disparity between existing global land cover maps, meaning ecosystem and land use science lacks the data to achieve high detailed comparative analyses [1]. Broad ecotones complicate landscape mapping further; with gradual transitions between geomorphic, adaphic and hydrologic gradients [2] presenting cartographers a great challenge to map to an appropriate level of detail.

The involvement of citizen sensors could facilitate the collection of unprecedented quantities of ecological information but there are naturally strong concerns over the validity of using amateur derived data in ecological research [3]. This paper discusses concepts important to the use of raster focused VGI with rigor from a spatial quality perspective within ecological applications.

## 2 Volunteered Geographic Information as a Research Tool

Over the course of the last decade the advancement and increased accessibility of geospatial technology has led to a blossoming in the quantity of geospatial information generated and shared across the web and via mobile applications. It allows us to better "observe, analyse and visualise" [4] our changing world and denotes that for some areas researchers now have access to richer geo-information that can be more accurate, more up-to-date and more complete than professional sources [5]. Merging data from a variably skilled crowd with that of specialist teams collecting conventional data seems absurd from a quality perspective however VGI (as introduced by Goodchild, [6]) is increasingly used as an ancillary source for both use in lieu of and to reinforce Professional Geographic Information (PGI) where PGI is unavailable or deemed insufficient [34].

Crowdsourcing is proliferating into domains demanding an increasingly high level of expertise, particularly where the timeliness of information is a necessity. Originally confined to basic identification tasks such as the 1930s Land Utilisation Survey of Britain [7] and the Christmas Bird Count, citizen sensing has matured through the emergence of Web 2.0 to the extent that citizens contribute towards complex issues traditionally confined to expert analysis. CrowdHydrology engages citizens in the collection of hydraulic measurements such as stream stage [8] and crowdsourcing methods have even been deployed in the creation of architectural 3D building models [9]. VGI has an emerging role in researching our environment, presenting a "powerful opportunity" [10] to understand current and future environmental changes.

## 3 Typologies of Citizen Sampling in Ecology

There is no ubiquitous, 'one approach fits all' solution with regards to citizen sensing quality assurance and as such it is important to understand the variance between types of VGI. There are diverse arrays of overarching project goals [11], projects permitting varying levels of freedom through varying power hierarchies, and projects focused upon differing stages of research inquiry. Something particularly pertinent to ecological research is the design and focus of sampling. Nichols & Williams [12] present targeted monitoring - based on *a priori* hypotheses and surveillance monitoring - which is not guided by such hypotheses, as two strands of conservation observation which can also be applied to citizen observation [13]. Targeted crowdsourcing is a

structured, directed approach traditionally associated with citizen science projects; specific questions are answered as contributors are instructed to collect or analyse information for well-defined hypotheses. Surveillance crowdsourcing has a generalised, unstructured aim, such as the establishment of a broad-scale environmental sensor network, signifying that extensive and often unexpected spatial patterns can be detected. Citizen Science often has a distinct hypothesis led goal opposed to surveillance crowdsourcing which can be much more amorphous in approach.

A second broad type of VGI, which can be split into structured hypothesis led contributions or unstructured surveillance approaches, is indirect VGI. It involves the use of openly licensed or creative commons data to contribute to hypothesis verification or unstructured surveillance, which was not intentionally created for the purpose of doing so, often forming a serendipitous linkage between a scientific problem and a semi-relevant, existing source. Indirect VGI sources that can be used within scientific analyses include ancillary information from Location Based Social Networks (LBSNs) and Web 2.0 content sharing platforms. Gschwend & Purves [14] show how the language used to describe Flickr photographs can, to an extent, relate to the undulations in a Digital Terrain Model, and textual information from LBSNs can be mined such as the detection of forest fires from geo-located Twitter data [15]. A structured instance of indirect VGI is the usage of OpenStreetMap which whilst having suggested ontologies and purpose, can be used in many separately conceived applications. Acknowledging the non-uniformity in types of VGI is crucial for addressing next steps with the data from a quality perspective.

## 4 Is there a Role for VGI in Remote Sensing Classification?

Despite the associated uncertainties [16], VGI has proven some utility in the domains of disaster and community mapping [17] and is increasingly used within the ecological sciences as a means of procuring information; indeed, why should researchers not utilise the world's largest research team [18] to counter for the flaws in ground truth acquisition within traditional Spatial Data Infrastructures (SDIs)?

### 4.1 Uncertainty & Error in Ecological Sampling

Ecology holds sampling and survey design in high esteem as the description of the spatial and temporal scales at which biodiversity is distributed forms the focus of research. As such, systematic and scientifically defensible biodiversity sampling is important. Limiting factors such as the model type, parameters and quality of the data employed within analyses denote that this is not always the case. Often, ecological and biogeographical research data are not collected in a standardised manner; the spatial resolution, temporal regularity, units of information collection and the captured degree of complexity varies from study to study [10]. It is a challenge to obtain an appropriate quantity of data with a trade-off between the need for highly accurate

data (spatial resolution, count of detected elements and biological validity) and the cost in both time and technical skills required to gather information at an appropriate level [2, 19]. Ecologists are thus faced with a further trade-off, as to whether sampling effort should be conducted on one area continuously or many areas sporadically [20].

One of the largest sources of error in studies of diversity and distributions is the variation in sample size; a change in sample size from 50 to 300 can alter the outcomes of subsequent habitat conservation target analyses by up to 45% [21]. Global land cover datasets indispensible for distribution analyses, such as the GLC-2000, MODIS and GlobCover have large spatial discrepancies between them [2]. Alongside the variance in sensors and calibration methodologies, a reason for the huge disparities in automatic land cover classification is the lack of sufficient in-situ data for the development of these products. Ground data can be logistically challenging and time consuming to acquire [22] exemplifying how traditional approaches can be flawed and signifying that alternate methodologies may be of value.

### 4.2 Uncertainty and Error in VGI

The trend of citizens creating geospatial information brings forth questions of quality assurance and sustainability. There are multiple stages at which error can propagate within crowdsourced content; the presence of error within the original interpretation of a phenomena (whether outdoors or through identification from a secondary source on the Internet) can be exacerbated by the improper description or notation of phenomena (whether a digitisation or a mislabeled or misallocated attribute). Based upon comparisons with PGI guided by the ISO 19157 spatial data quality principles, VGI has limited and varied accuracy with its completeness, lineage, logical consistence, attribute accuracy and positional accuracy [16, 23] limiting usability in professional contexts.

Despite the potential advances in addressing the prominent issue of under-sampling, VGI presents additional problems with regards to sampling bias. Citizen sensors are less likely to collect information in a systematic and consistent manner as the spatial scales over which they are encouraged to do so can be very broad - particularly with regards to the indirect and direct surveillance approaches described. Sampling effort will also vary geographically as digital exclusion [16] denotes that the coverage of citizen sensing is heterogeneous. The consequences of incorporating extra uncertainty and misrepresenting distribution within the outlined data collection steps are numerous [24]; inappropriate decisions could be made regarding the foci of conservation efforts and funding and the misrepresentation of the ecosystems people depend on can lead to low quality environmental management that depreciates the value of our ecosystems' services.

### 4.3 Can VGI Address the Flaws of Ecological Sampling?

VGI is increasingly seen as a valid data retrieval method within the ecology and biogeography communities; Peters [25] portrays non-traditional knowledge as integral for science driven synthesis and calls for its integration with traditional sampling strategies to provide crucial feedbacks for the determination of future ecological sampling requirements. Its emergence signifies that given appropriate quality control measures, ecologists now have the capability to attain data for areas and purposes previously prohibitively expensive to attain at the appropriate quantitative level.

Technology is not yet sufficiently advanced to provide a near perfect digital representation of reality. The representivity of ecological studies is dependent upon the completeness of data [26] yet the complexity of the environment denotes that all measurements are erroneous to a certain, unknown extent. As such, VGI is often critiqued for not meeting an impossible ideal. If VGI can be of a 'good enough' quality, it can be argued that the additional information it provides can be used to detect and correct for bias in traditionally obtained samples. It is extremely important to understand the tradeoffs between obtaining a high quantity of intrinsically flawed data and in obtaining high quality, verified data. There is a clear case to use VGI where the need for the data is greater than the impact of the potential risks to using the crowdsourced input or if the risks of using the information can be mitigated through quality control.

## 5 Fusing VGI and Remote Sensing

Despite the potential to vastly increase the quantity of ground truthed training data, VGI is rarely combined with remote sensing techniques. It has been used frequently for the validation of thematic map outputs (such as the use to verify global land cover products through the GeoWiki platform – [1]) yet is rarely incorporated within the classification algorithms used to *generate* thematic maps. Schnebele and Cervone [27] take verification a step further by refining a flood hazard map according to the presence of VGI, though its combination with image classification has not yet been practiced within the domain of geospatial science. Given the importance of the temporality of ground truth information, the use of VGI could be imperative to improving the timeliness of land cover change detection, particularly in areas infrequently surveyed as ground truth and in the domain of post-disaster management where the addition of crowd data derived extra training sites could vastly improve the authenticity of image classification. If crowd data is to be used within machine learning classification the research challenge lies in ensuring that the training samples are of extremely high quality. Finding methodologies of appropriately obtaining and weighting training data could promote the use of crowd interpretations in a broad range of applications and forms the focus of underway research.

# 6 Fostering Quality in VGI Derived Image Classification

The research community has developed insight into how accurate VGI can be [16, 23, and 28] but the best practices for using and merging different types of VGI with authoritative data are yet to be fully defined. The level of trust ecologists can place in diverse crowd generated content is variable; incomplete and inaccurate information may be supplied which must be accounted for prior to usage in formal applications. A benefit in using VGI with remote sensing classification is that weightings may easily be applied. Diversity in quality makes research into the intricacies of weighting submissions integral and is rarely investigated in terms of best practices for the various components of citizen sensor typologies. The following strategies will be trialed in selecting the most appropriate instances of training data within future research in the context of habitat mapping.

## 6.1 Variance in Structure & Codes of Practice

Kodric-Brown & Brown [26] depict how "there is no substitute for first hand field experience with organisms and habitats" emphasizing that whilst it is too much to expect that all contributors to the science have first-hand expert knowledge, some must. This approach applied to the citizen sensing means that to achieve quality output from the community, contributors must be spearheaded by experienced professionals to avoid any serious errors of interpretation and application. Designing crowd projects to conform to standard protocols of data collection via the distribution of precise and strict instructions could aid interoperability and increase the accuracy of submissions by removing room for error.

Hypothesis led, directed approaches often yield very focused and useful results in ecology [12]. Tasks can be designed to automate certain recordings such as location via a device's GPS and minimise error by guiding the user through a task. The design of VGI studies through platforms such as Crowdcrafting permit researchers to set up a specified number of tasks which can be controlled in terms of the number of participants contributing, the precise geo-location of the sensing and the type of data entered. There is little freedom in what, where and how something is sensed which provides a stark contrast to the indirect data scraped from LBSNs and unstructured directed approaches such as OpenStreetMap. Here the data is produced with little or no hypothesis led guidance resulting in extremely variable responses, few of which are wholly relevant to the research goals. Indirect VGI content is ubiquitous and despite not being necessarily fit for purpose, usage of this data could broaden sample size and with authoritative direction could provide a more appropriate distribution. It is important to understand how project structure and codes presented to the crowd can affect and inform best practices for quality induction with regards to the gathering of crowd training data. Future research will involve comparing indirect and direct ap-

proaches in terms of the applicability of the training data they yield in habitat classification.

### 6.2 User characteristic filtering

Low quality submissions can outweigh the benefit provided by accurate submissions. Looking at the accuracies of users within the crowd can minimise the impact of inaccurate input through informing annotator aware models in machine learning [29]. Dickinson et al. [13] suggest the exclusion of contributions from new participants and from participants that submit erroneous and erratic reports. There is little evidence to suggest that frequency of submission has an impact upon quality; the inference that high frequency annotators have more skill has been shown to be insignificant following analysis of intra-annotator accuracy over time [30]. Inter-annotator accuracy is very heterogeneous owing to variance in motivation and ability, signifying that the weightings of submissions should be adjusted on a source-dependent basis so as to provide representative analyses [27, 31]. Knowing the likelihood of a sample's accuracy is imperative; the addition of the most skilled volunteers (judged on annotator accuracy) through probabilistic multi labelling approaches have been shown to improve consensus labeling accuracy [32].

Filtering methods discussed within the literature have great potential but are dependent upon the presence of annotator metadata. Anonymysed submissions (often from unstructured and indirect sources) are problematic, indeed, how can researchers best aggregate these responses to pick out and eliminate submissions great inefficiencies in the data collection chain? Unfortunately in the case of indirect crowdsourced information, the details of the user submitting the information are often inaccessible or insufficiently detailed to inform many filtering techniques. Assuming for and predicting non-uniformity within the crowd through learned probabilistic models is difficult, particularly where little metadata exists, yet could be of great use and will be explored in selecting high quality ground truth samples.

### 6.3 Ancillary information

Comparing crowd submitted content to existing sources of information is inappropriate when the 'ground truth' information itself is missing or of poor quality, particularly in the case of detecting changes to existing features as no authoritative ground truth may exist. In this situation we must refer to logic based mechanisms of validation which depend on known facts. An example in the field of ecology would be the comparison of the location of a geo-tagged species with existing certified knowledge depicting its known geographic range and physical tolerances. If this geo-tagged species is within the statistically significant bounds of a historically established range one can determine the degree of likelihood that the submission is legitimate. If it is not within a statistically significant range (or buffer zone) and without the appropriate metadata

as evidence then the submission can be regarded as of inappropriate quality. Building an open, interoperable and comprehensive database of such variables could be extremely important as we begin to encourage and automate the introduction and convergence of volunteered content in a range of traditionally authoritative and closed domains. It is hoped that applying these traditional knowledge based strategies to crowdsourced data sets will assist in assigning weights and probabilities of accuracy to citizen sensors.

## 7 Conclusions and Future Research

Direct and indirect sources of VGI have been presented as a baseline for comparing and evaluating the potential that crowd generated content has in training classifiers for thematic map production. This fusion of weighted crowd interpretation and remote sensing classification techniques is proposed as an under-explored mechanism of fostering quality in the domain of habitat mapping and change detection, moving away from the heavily explored area of analysing vector VGI. Specifically the integration of crowd interpretations with training data in remote sensing classification will be explored. The next step in this research will be to conduct experiments which have been designed to explore the quality control mechanisms raised within this paper. A Crowdcrafting Application has been developed to obtain user derived habitat classifications based on the use of a JNCC Phase 1 habitat classification system [33] and photographs taken at several sites within the New Forest, Hampshire (United Kingdom). User-accuracy metrics derived from the comparisons of the thematic classifications against that of an ecologist will be incorporated as weighted training data prior to the classification of remotely sensed imagery for the same area. This will enable the determination of the value of directed non-traditional data sources in a habitat distribution context. It is hoped that the research will inform the emerging field of VGI quality enhancement and in particular, lead to a greater understanding of how VGI can be seen as an asset to remote sensing classification.

## 8 Acknowledgements

Laura Kinley is supported by the Horizon Doctoral Training Centre at the University of Nottingham (RCUK Grant No. EP/G037574/1) and is supervised by Professor Mike Jackson and Professor Giles Foody.

# 9 References


1. Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., & Obersteiner, M. (2012). Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, 31, 110-123
2. Arnot, C., Fisher, P. F., Wadsworth, R., & Wellens, J. (2004). Landscape metrics with ecotones: pattern under uncertainty. *Landscape Ecology*, 19(2), 181-195.
3. Cohn, J. P. (2008). Citizen science: Can volunteers do real research? *BioScience*, 58(3), 192-197.
4. Dalby, S. (2012). Geo 2.0: digital tools, geographical vision and a changing planet. *The Geographical Journal*, 178(3), 270-274.
5. Devillers, R., Bégin, D., & Vandecasteele, A. (2012). Is the rise of Volunteered Geographic Information (VGI) a sign of the end of National Mapping Agencies as we know them? *GIScience 2012 workshop "Role of Volunteer Geographic Information in Advancing Science: Quality and Credibility"*, Columbus, OH, September 18, 2012
6. Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.
7. Jiang, M., Bullock, J. M., & Hooftman, D. A. (2013). Mapping ecosystem service and biodiversity changes over 70 years in a rural English county. *Journal of Applied Ecology*.
8. Lowry, C. S., & Fienen, M. N. (2013). CrowdHydrology: Crowdsourcing Hydrologic Data and Engaging Citizen Scientists. *Ground Water*, *51*(1), 151-156.
9. Uden, M., & Zipf, A. (2013). Open Building Models: Towards a Platform for Crowdsourcing Virtual 3D Cities. In *Progress and New Trends in 3D Geoinformation Sciences* (pp. 299-314). Springer Berlin Heidelberg.
10. Hampton, S. E., Strasser, C. A., & Tewksbury, J. J. (2013). Growing Pains for Ecology in the Twenty-First Century. *BioScience*, *63*(2), 69-71.
11. Wiggins, A., & Crowston, K. (2011, January). From conservation to crowdsourcing: A typology of citizen science. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on* (pp. 1-10). IEEE.
12. Nichols, J. D., & Williams, B. K. (2006). Monitoring for conservation. *Trends in Ecology & Evolution*, *21*(12), 668-673.
13. Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics*, *41*, 149-172.
14. Gschwend, C., & Purves, R. (2011). Comparing Flickr tags to a geomorphometric classification. *In Proceedings of the Nineteenth Annual GIS Research UK Annual Conference, Portsmouth, United Kingdom* (pp. 174-78).
15. De Longueville, B., Smith, R. S., & Luraschi, G. (2009, November). OMG, from here, I can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. *In Proceedings of the 2009 International Workshop on Location Based Social Networks* (pp. 73-80). ACM.
16. Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design*, 37(4), 682.
17. Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. International Journal of Digital Earth, 3(3), 231-241.
18. Irwin, A. (1995). *Citizen science: A study of people, expertise and sustainable development.* Psychology Press.

19. Godet, L., Fournier, J., Toupoint, N., & Olivier, F. (2009). Mapping and monitoring intertidal benthic habitats: a review of techniques and a proposal for a new visual methodology for the European coasts. *Progress in Physical Geography*, 33(3), 378-402.
20. Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and distributions*, 16(3), 354-362.
21. Metcalfe, K., Delavenne, J., Garcia, C., Foveau, A., Dauvin, J. C., Coggan, R. & Smith, R. J. (2013). Impacts of data quality on the setting of conservation planning targets using the species–area relationship. *Diversity and Distributions*, 19(1), 1-13.
22. Townshend, J. R., Masek, J. G., Huang, C., Vermote, E. F., Gao, F., Channan, S., & Wolfe, R. E. (2012). Global characterization and monitoring of forest cover using Landsat data: opportunities and challenges. *International Journal of Digital Earth*, 5(5), 373-397.
23. Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, 14(4), 435-459.
24. Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jimenez-Valverde, A., Ricotta, C. & Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, *35*(2), 211-226.
25. Peters, D. P. (2010). Accessible ecology: synthesis of the long, deep, and broad. *Trends in ecology & evolution*, *25*(10), 592-601.
26. Kodric-Brown, A., & Brown, J. H. (1993). Incomplete data sets in community ecology and biogeography: a cautionary tale. *Ecological Applications*, 736-742.
27. Schnebele, E., & Cervone, G. (2013). Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Science*, *13*(3),669-677.
28. Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110-120.
29. Rodrigues, F., Pereira, F., & Ribeiro, B. (2013). Learning from multiple annotators: Distinguishing good from random labelers. *Pattern Recognition Letters*.
30. Foody, G. M. (2013). Rating crowdsourced annotations: evaluating contributions of variable quality and completeness. *International Journal of Digital Earth*, (ahead-of-print), 1-21.
31. Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. A., Scott, E. M., ... & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology & Evolution*, 25(10), 574-582.
32. Kumar, A., & Lease, M. (2011, February). Modeling annotator accuracies for supervised learning. In *WSDM Workshop on Crowdsourcing for Search and Data Mining* (pp. 19-22).
33. JNCC, (2010), Handbook for Phase 1 habitat survey - a technique for environmental audit, ISBN 0 86139 636 7
34. Du, H., Anand, S., Alechina, N., Morley, J., Hart, G., Leibovici, D., & Ware, M. (2012). Geospatial information integration for authoritative and crowd sourced road vector data. *Transactions in GIS*, *16*(4), 455-476.