Assessing the Accuracy of Volunteered Geographic Information Derived Habitat Classification

Laura Kinley[1], *Giles Foody*[2] *& Mike Jackson*[1]

[1] Nottingham Geospatial Institute, The University of Nottingham, Jubilee Campus,
Nottingham, NG7 2TU, UK

[2] School of Geography, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK

Email: psxlk@nottingham.ac.uk

Abstract

Crowd-assisted annotation is trialled as a mechanism of producing accurate thematic habitat maps. Volunteered Geographic Information (VGI) classifications of ground-based photographs (made using the UK's Phase-1 Habitat Survey nomenclature) are gathered and filtered according to metrics of annotator accuracy and user-confidence in labelling. Interpretations for 150 sites initially classified by a professional ecologist, in subsets of varying class distinguishability were then subjected to accuracy assessment techniques to determine their utility as training data for the classification of remotely sensed imagery. Initial results show user-confidence as an encouraging indicator of label accuracy and VGI to be a promising substitute for authoritative training samples where gold-standard data is inaccessible.

Keywords: Volunteered Geographic Information, Habitat Classification, Accuracy-Filtering.

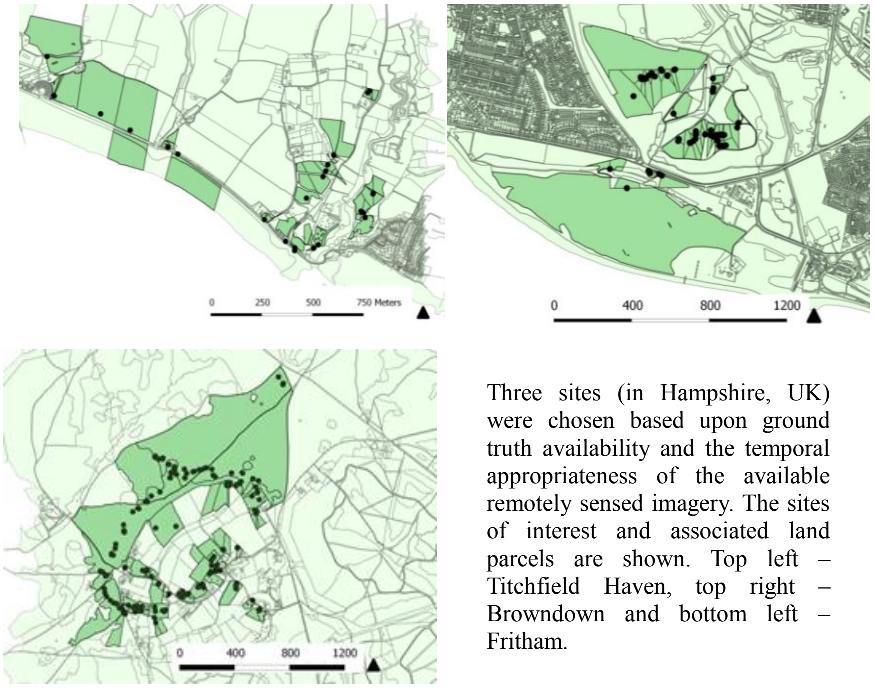
1. Introduction

Producing accurate maps of habitat loss and fragmentation is integral to understanding and managing the environment. Many authoritative habitat and land cover maps could benefit from additional ground truth to enable improved identification of species distributions and transition gradients (Dickinson *et al.*, 2010). VGI presents opportunities for improving current practices of data collection and annotation, providing extensive sample coverage alongside viewpoints from multiple 'sensors' that would not otherwise be economically feasible to collect. Minimal quality control practices do however prompt concerns over accuracy and completeness (Heipke, 2010); VGI typically exhibits considerable quality variance ranging from imprecise digitisation to poor label accuracy making usage for planning and scientific purposes challenging from a data quality perspective.

The research community has developed some insight into how accurate VGI can be and how it can be utilised for validating existing datasets (Goodchild and Li., 2012); however, there is scope for VGI and the more directed approach of citizen science to be used in image classification. In combination with machine learning, focused sampling has the potential to produce training samples which, with appropriate weighting, (such as down-weighting less accurate contributors and emphasising the contributions of accurate annotators) could yield timely outputs of appropriate thematic accuracy. The parameters under which crowdsourced annotations should be collated and weighted by accuracy to reduce label noise have been explored in machine learning (Lease., 2011) yet are rarely applied in the domain of VGI. This research aims to assess the potential value of VGI in habitat classification with the hope of informing authoritative data collection methods.

2. Data & Methods

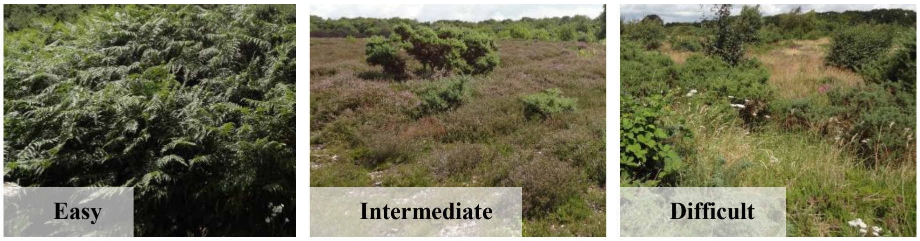
The UK's Phase-1 Habitat Classification scheme (JNCC, 2010) is used as a case study for the incorporation of VGI into the remote sensing training stage. 1,795 crowd habitat classifications of 150 georeferenced, ground based photographs taken within the New Forest (Figure 1) have been gathered via a web application and compared to an ecologists in field interpretations as the assumed gold standard. The area has a high diversity of flora and one of y urope's most extensive areas of heathland signify that timely habitat change detection is important.



Three sites (in Hampshire, UK) were chosen based upon ground truth availability and the temporal appropriateness of the available remotely sensed imagery. The sites of interest and associated land parcels are shown. Top left – Titchfield Haven, top right – Browndown and bottom left – Fritham.

Figure 1: Crowd classification test sites and associated Ordnance Survey land parcels

The photographs were randomly stratified by habitat class and classification difficulty level which was determined by transition gradients present within the photographs and uncertainty ratings provided by the ecologist (Figure 2). Broad and sub-level habitat classifications were sourced from 83 participants ranging from the amateur-enthusiast to professional ecologists who were allocated an experience category following a short questionnaire. Annotators were timed and screen-recorded to establish any associations of task time and attention to the provided training materials with annotation accuracy. Levels of user-confidence in each classification were recorded to establish its utility in filtering the training samples in addition to user attention to quality-control tasks incorporated to identify spurious annotations detrimental to overall labelling accuracy.



Broad Class(es)	Tall Herb & Fen	Heathland / Woodland & Scrub	Heathland / Woodland & Scrub / Coastal
Sub-Class(es)	Bracken	Dry Dwarf Shrub Heath / Dense Scrub	Dry Dwarf Shrub Heath / Scattered Scrub / Coastal Heathland
Certainty	Certain	Fairly Certain	Uncertain

Figure 2: Example images from each category with ecologist's broad and sub habitat classification and certainty rating

Cohen's Kappa statistics were used to investigate the reliability of inter-annotator agreement and annotator accuracy and per-term consistency was assessed in terms of the annotator's expertise level. Given the differing degrees of correctness within the interpretation of a habitat (owing to hierarchy within the classification system and inherent uncertainties from multi-class ownership), a linguistic scale (Woodcock and Gopal, 2000) is also used to assess annotation accuracy.

3. Results

3.1 Accuracy of crowd classification

When all of the classifications are taken into consideration the broad level accuracy is 64.40% (overall Cohen's Kappa agreement: 0.55, p=<0.01) and the sub level accuracy is 30.08% (overall Cohen's Kappa agreement: 0.31 p <0.0005). When solely the modal classification for each image is used, the accuracies rise to 77.33% and 47.33% respectively. Given that Remote Sensing classifications are typically deemed to be of suitable accuracy when classification accuracies of 85% are consistently seen across multiple classes (Foody, 2002); the sub level classifications are very poor and the broad level classifications are below adequate.

Broad scale inter-class overall accuracy varies by up to 39.89% (Table 1) suggesting that the crowd could be better suited to identifying certain habitat types. At the broad classification level 'woodland and scrub' and 'heathland' are the most commonly misclassified habitat type (43.07% and 45.95% of classifications were incorrect respectively). The crowd were most successful at identifying the 'open water' habitat followed by the 'tall herb and fern' category. Analysis of the sub level accuracies shows further disparity in inter-class accuracy with certain classes such as arable land being consistently identified by the crowd (72.06% accuracy) whereas for more indiscriminate classes such as acid and neutral grassland varieties the crowd classified the images inadequately (14.64% and 15.38% respectively).

Table 1: Broad Class Accuracy

Group	User's Accuracy	Producer's Accuracy	Std. Dev.	Overall Accuracy
Woodland & Scrub	61.79%	59.58%	49.55	56.93%
Grassland & Marsh	73.31%	68.79%	47.16	66.73%
Tall Herb & Fern	66.45%	69.10%	46.24	69.21%
Heathland	65.27%	57.98%	49.97	54.05%
Open Water	93.94%	93.94%	24.23	93.94%
Miscellaneous	65.29%	71.43%	45.38	71.16%

Inter-group broad scale accuracies vary by up to 12.91% with ecologists & ornithologists as the most accurate annotators (Table 2). There is significant variance between ecologists, ornithologists and the group of participants with no background or interest in ecology (observed Z values of 2.55 and 2.03 respectively) showing that filtering crowd annotations by group membership could improve the accuracy of crowd derived classifications.

Table 2: Cohen's Kappa Analysis

Group	Kappa	Variance	Z statistic	Overall Accuracy
Interested Amateur	.592	0.0035	10.01	65.63%
Ecologists	.652	0.0024	13.31	74.45%
No Interest	.468	0.0028	8.84	61.54%
Ornithologists	.613	0.0023	12.78	70.00%
Students	.600	0.0025	12.00	68.79%

Landscapes are not crisply segmented (Rocchini, 2007) and hierarchical classification systems also signify that interpreters can be correct to varying degrees. Relative similarity scores based upon Woodcock and Gopal (2000) linguistic scales were thus attributed to each classification based on the hierarchical nature of the classification system and the identification of habitats that are commonly juxtaposed. When taking fuzzy accuracy methods into account, the percentage accuracies for the sub-classes (aggregated by broad class in Table 3) improve significantly with the highest improvement being seen for 'open water' and 'grassland and marsh' categories which encompass a great number of hard to distinguish sub-categories.

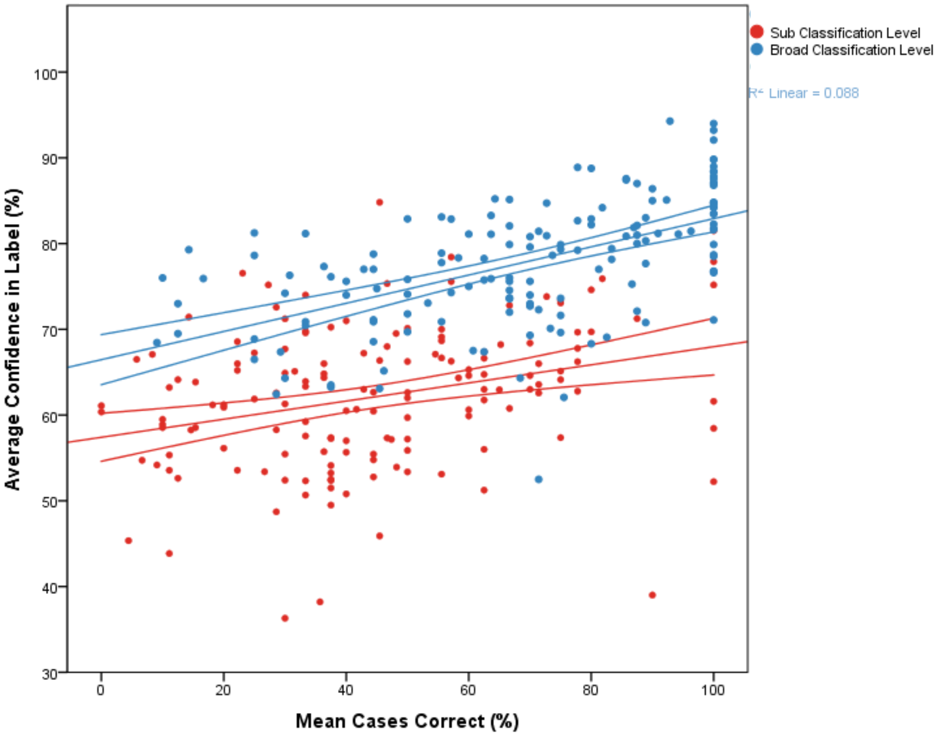
Table 3: Percentage Accuracies and Percentage Improvements - using less stringent data as determined by a linguistic scale.

	Matches Using (Cumulative %)			Improvement	
Habitat Class	**Max M**	**Right R**	**Somewhat S**	**(R-M)**	**(S-M)**
Woodland & Scrub	31.7	59.5	87.8	27.8	56.1
Grassland & Marsh	23.3	66.7	67.1	43.4	43.8
Tall Herb & Fern	38.7	69.2	93.7	30.5	55
Heathland	16.8	55.7	98.4	38.9	81.6
Open Water	36.4	93.9	100	57.5	63.6
Miscellaneous	39.3	71.2	71.2	31.9	31.9

3.2 Correlations with ancillary information

Small though statistically significant negative correlations (p= <0.005) were found with both the number of clicks per classification and time taken to perform each classification at both the broad hierarchical level (a correlation of -2.44: clicks, -2.45: time) and sub level (-3.62: clicks, -1.81: time). A greater number of clicks denotes that more time was spent looking at the associated classification definitions webpage. The negative correlations of both the time taken and number of clicks per classification with accuracy are perhaps the result of hesitation and uncertainty indicating a complex classification site. Pearson’s correlation coefficient tests show a medium positive correlation (0.56, p= <0.01) with user-confidence at the broad classification level and a medium positive correlation at the sub hierarchical level (0.366, p= <0.01) signifying that user-confidence could be an adequate indicator of annotation accuracy.

Graph 2: Relationship between classification accuracy and label confidence at broad and sub hierarchical levels



3.3 Qualitative Interpretations

Problems such as the low image resolution highlighted by participants (Table 3) could provide an explanation for the low sub-scale accuracy; certainly the sub grassland and hedgerow categories require high resolution imagery to classify them to an appropriate degree of accuracy. Border areas were also highlighted as something which prohibited classification; this information highlighting uncertain areas could however be of use in determining ecotonal boundaries and provide a good indicator as to where further professional sampling effort is required.

The qualitative data also shows interesting interpretation variations owing to the fact that certain international participants were unfamiliar with certain categories. This suggests that familiarity or exposure to a certain habitat could promote accuracy and the presence of between-group variance in case labelling, however, no statistically significant variance in accuracy was found between UK and non-UK participants.

Table 4: Interpreter's comments regarding classification uncertainties

Theme	Participant Comments
Low image resolution	*Can't see enough detail to be any more specific* about the vegetation - *I'm making inferences* from the Salix in the surrounding area. Certainly not a single-species hedge - diversity looks good but *not easy to assess.* As it is quite heavily grazed I don't find it possible to tell whether this is dry or humid heath (*can't see whether Molinia is present* but I suspect not). Also it may have a reasonably high base status but *not enough detail in the grassland* to be sure.
Border areas causing uncertainty	The image is located at the *border between two habitats*. The habitat is *dominated by two different types* of grass, which are grouped together in patches Hard to tell as *many different types in same area.*
Lack of context	I have to choose "Intertidal" for coastal open water, although *I have no way of knowing.* *Perspective issue* - more bracken in landscape but not in photograph.
Classification system inadequacies	To me, this looks to be heathland overgrown with bracken but *the only classification that allows for "bracken" is "tall herb and fern*" It's certainly a grassland community *but impossible to classify further with any degree of accuracy* given that I'm viewing on a phone and have no species list.
Lack of participant knowledge / understanding	*Not sure if I'm familiar enough with UK ecosystems* to answer. Interesting. I have to look up half the expressions though, like the difference between mire and marsh which is *pretty clear to me in German, but not in English.*

4. Conclusion

The proliferation of geo-technology and popularity of citizen-sensing suggests that ecologists can access valuable additional data. There are opportunities to move beyond the use of VGI for validation, filtering VGI according to indicators of accuracy and integrating the annotations within remote sensing classification. This paper examines the utility of VGI as ancillary data for remote sensing classification and shows how datasets can be synthesised that, whilst not reaching the accuracies of authoritative data, could be seen as an asset to ecological monitoring and inventory. Next steps include the use of VGI for the supervised classification of four-band aerial imagery captured around the time of the initial survey.

Acknowledgments

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. Thank you to Mercedes Torres and Ordnance Survey for the provision of the ground-based photography and authoritative classifications.

References

Dickinson, J., Zuckerberg, B., & Bonter, N. (2010), Citizen Science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics*, 41, 149-172.

Foody, G. M. (2002), Status of land cover classification accuracy assessment. *Remote sensing of environment*, *80*, 185-201.

Goodchild, M.F. & Li, L. (2012), Assuring the quality of volunteered geographic information. *Spatial statistics*, 1, 110-120.

Heipke, C. (2010), Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(6), 550-557.

JNCC, (2010). Handbook for Phase 1 habitat survey -a technique for environmental audit, ISBN 0861396367

Lease, M. (2011), On quality control and machine learning in crowdsourcing. *Proceedings of the 3rd Human Computation Workshop at AAAI*, 97–102

Rocchini, D., & Ricotta, C. (2007), Are landscapes as crisp as we may think? *Ecological Modelling*, *204*, 535-539.

Woodcock, C. & Gopal, S. (2000), Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographical Information Science*, *14*, 153-172.