Mining user-generated geographic content: An interactive, crowdsourced approach to validation and supervision

Frank O. Ostermann Gustavo A. Garcia-Chapeton Raul Zurita-Milla Menno-Jan Kraak

Faculty of Geo-InformationScience and Earth Observation (ITC)
University of Twente
PO Box 217, 7500 AE Enschede, The Netherlands
{f.o.ostermann, g.a.garciachapeton, r.zurita-milla, m.j.kraak}@utwente.nl

Abstract

This paper describes a pilot study that implements a novel approach to validate data mining tasks by using the crowd to train a classifier. This hybrid approach to processing successfully addresses challenges faced during human curation or machine processing of user-generated geographic content (UGGC), namely quality control, reproducibility, sustainability, scaling, data quality, overfitting, and training costs. We test the approach on mining UGGC to derive information on local places as humans perceive them. Specifically, we retrieve Flickr image metadata, enrich it semantically by building term vectors using a controlled vocabulary, cluster it spatially, let online participants rate those clusters, classify them into noise and places by using both semantic and cluster characteristics, let online participants supervise the classification by annotating the results, and use their feedback to improve clustering and revise the trained model. The results show that the approach is feasible and suggest future studies to improve it, while also indicating that mining places from UGGC requires more than a single source.

Keywords: crowdsourcing, user-generated geographic content, places, data mining, supervised machine learning, hybrid geoprocessing.

1 From user-generated content to user-supervised analysis

This paper describes a pilot study that implements a novel approach for crowdsourcing the validation of clusters created from user-generated geographic content, or UGGC (Craglia et al., 2012). This validation is done by letting the crowd train a classifier.

Although UGGC has proven its utility for a variety of tasks and purposes (Fast and Rinner, 2014; Granell and Ostermann, 2016; Haworth, 2016), high semantic and syntactic heterogeneity, and unknown provenance and production parameters have a negative impact on its fitness-for-use.

A common strategy has been to crowdsource curation tasks (Sui et al., 2012): Human volunteers check UGGC accuracy, assign labels, and prioritize further processing. Despite encouraging results, this approach lacks quality control and reproducibility (Camponovo and Freundschuh, 2014), guaranteed sustainability, and efficient scaling up (Morrow et al., 2011).

Another approach is to employ data mining (DM) and machine learning (ML) techniques to select, filter, classify, and enrich UGGC. Here, at least three main challenges persist: Dependency on input data quality for unsupervised DM and ML (Kanevski et al., 2008), overfitting of the learning model (Butler, 2013), and training costs for different contexts and tasks (Spinsanti and Ostermann, 2013).

By combining human and computational analysis of UGGC, we aim to address those challenges. Such a hybrid processing approach could improve UGGC's fitness-for-use by exploiting contextual, local, or traditional knowledge from the human supervisors to ensure meaningful results, while using computation to reduce reliance on volunteers and to manage big data sets. The primary aim of this study is to test the overall feasibility of the approach. The secondary aim is to discover distinct urban places from meta-data of georeferenced photographs, contributing to the development of geospatial representations of our environment that account for different perspectives and the vagueness of human place conceptualization. To search for places, we look into the tags and descriptions of Flickr images, using a controlled vocabulary of terms for activities, qualities, and elements of places (Purves et al., 2011). We combine this semantic enrichment with clustering based on spatial proximity, and use a supervised classifier to remove noise from meaningful results. A web interface presents the results to human study participants for an interactive and iterative map-based validation and supervision.

2 Pilot Study on Geotagged Photos

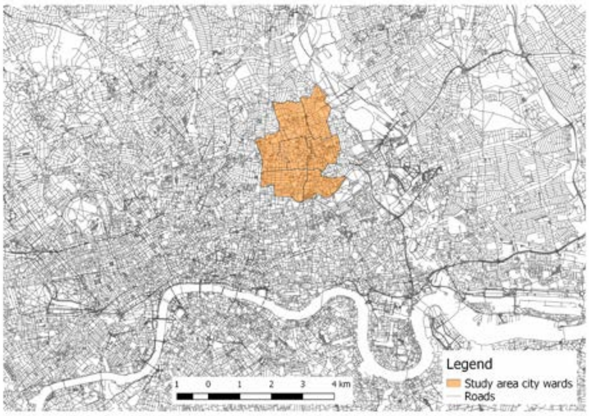
2.1 Study Subject and Area

Shared image content on platforms such as Flickr, Panoramio, and Instagram has received a substantial amount of research interest, because photographs often possess a strong semantic link between image and descriptive text, and geographic location (Sigurbjörnsson and Van Zwol, 2008). Flickr provides a mature and accessible API. It allows users to tag and describe images, and offers rich EXIF1 metadata. Most studies use a pre-defined, coarse geospatial granularity, or do not validate all of the results, which our approach intends to address.

[1] EXIF (Exchangeable Image File Format) is a standard used by digital cameras to record technical information of the camera's status when shooting a photograph.

As a study area, we chose seven London City wards (Queensbridge, Dalston, Hackney Downs, Leabridge, Victoria, Hackney Central, Chatham, shown in Figure 1), because of their rich and diverse urban fabric, the abundance of UGGC, the absence of major touristic hotspots, and their outer administrative boundaries forming a mostly convex hull to reduce edge effects.

Figure 1 Study area of seven City Wards within London



2.2 System Architecture and Set-up

The pilot study work flow consists of eight main phases, which Figure 2 shows in an overview, and which are explained in more detail below

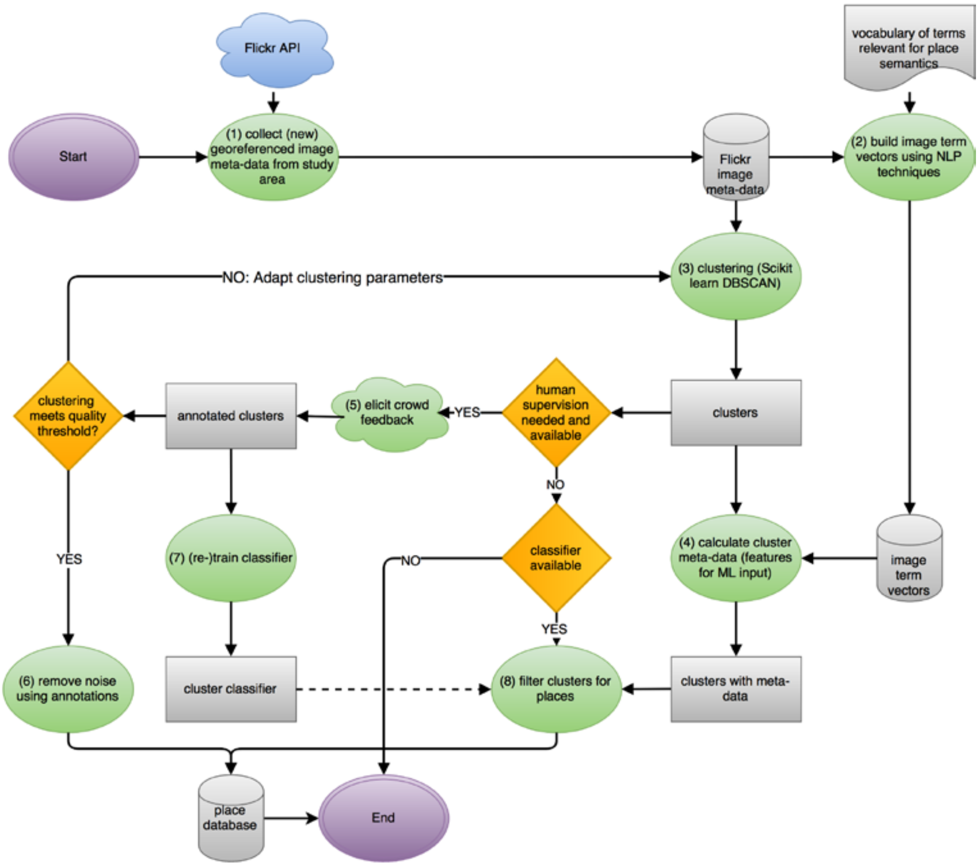
Phase 1 collects metadata of georeferenced Flickr images from the study area using the public API, and stores the retrieved information in a PostGIS database.

Phase 2 finds place-related terms and builds term vectors, by using a controlled vocabulary that consists of 107 activities (e.g. “party”, “football”, “exhibition”), 313 elements (e.g. “church”, “station”, “graffiti”), and qualities (e.g. “dark”, “royal”, “woods”). For every image, its tags, title and descriptions were parsed to find vocabulary terms through lexical matching, and a term vector was constructed.

Phase 3 searches for spatial clusters. Images that contain vocabulary terms are input for the clustering, using the DBSCAN algorithm from the Python Scikit-learn framework. DBSCAN can deal well with varying density of points as well as irregular shapes of clusters, and Scikit-learn could be reliably integrated into our workflow. Further, it is computationally inexpensive and fast, and has been used successfully in many studies. The clustering was spatial, using longitude and latitude as only features.

Phase 4 computes thematic and spatial cluster characteristics. For each cluster, we calculated several thematic and spatial characteristics to be used as input features

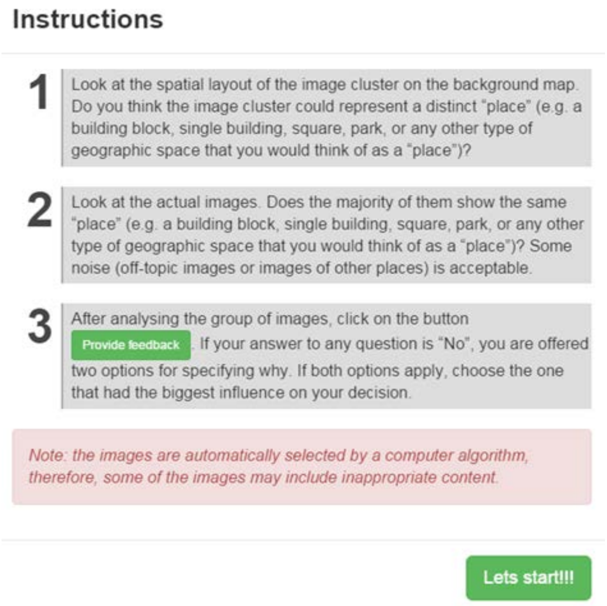
Figure 2 Pilot study prototype implementation and workflow


Flickr API
vocabulary of terms relevant for place semantics
Start
(1) collect (new) georeferenced image meta-data from study area
Flickr image meta-data
(2) build image term vectors using NLP techniques
NO: Adapt clustering parameters
(3) clustering (Scikit learn DBSCAN)
clustering meets quality threshold?
annotated clusters
(5) elicit crowd feedback
YES
human supervision needed and available
clusters
NO
YES
(7) (re-)train classifier
NO
classifier available
(4) calculate cluster meta-data (features for ML input)
image term vectors
YES
(6) remove noise using annotations
cluster classifier
(8) filter clusters for places
clusters with meta-data
place database
End

for the ML classifier: First, the average and median cosine similarity. Cosine similarity is a common metric for comparing the semantic similarity between two term vectors, and equals the cosine of the angle between the two term vectors. To measure the internal similarity of a cluster, we calculated its average and median (to mitigate the effect of a single outlier image within an otherwise homogeneous cluster) cosine similarity from all its image pairs. Second, the number of images and unique contributors might indicate which image clusters define a distinct geographic place. Additionally, we computed the average and median silhouette (Rousseeuw, 1987) scores of all items for each cluster. The silhouette coefficient measures how similar an object is compared to the other objects in its cluster. It ranges from -1 to +1, with high values indicating a poor match with other clusters and a good match with objects in its own cluster.

Phase 5 crowdsources the validation and annotation of clusters. For the first workflow iteration, we considered all found clusters to be potential places. The web interface presents these clusters one at a time to a human supervisor. The screenshot in Figure 3 shows the initial instructions presented to a human supervisor. The interface then presents the first cluster, with the location of the images shown on a map (using OpenStreetMap as a base map) on the left side of the interface, and the actual images shown in a gallery on the right side (see Figure 4). It is possible to select images on the map and gallery. Once a supervisor clicks on the “Provide feedback” button, s/he can comment on the spatial layout and thematic consistency of the cluster (see Figure 5). If any question is answered “No”, additional feedback options appear. For question 1, those are “Wrong shape” and “Too large”. The option “Too small” was discarded after initial tests, since even if all images from exactly the same location, their content can cover much more geographic area. For question 2, the additional options are “There is more than one place shown” and “There are too many images that are not about a place at all!”

Figure 3 Initial instructions to human supervisors



Instructions

1 Look at the spatial layout of the image cluster on the background map. Do you think the image cluster could represent a distinct “place” (e.g. a building block, single building, square, park, or any other type of geographic space that you would think of as a “place”)?

2 Look at the actual images. Does the majority of them show the same “place” (e.g. a building block, single building, square, park, or any other type of geographic space that you would think of as a “place”)? Some noise (off-topic images or images of other places) is acceptable.

3 After analysing the group of images, click on the button Provide feedback . If your answer to any question is “No”, you are offered two options for specifying why. If both options apply, choose the one that had the biggest influence on your decision.

Note: the images are automatically selected by a computer algorithm, therefore, some of the images may include inappropriate content.

Figure 4 Interface showing typical cluster to be labelled

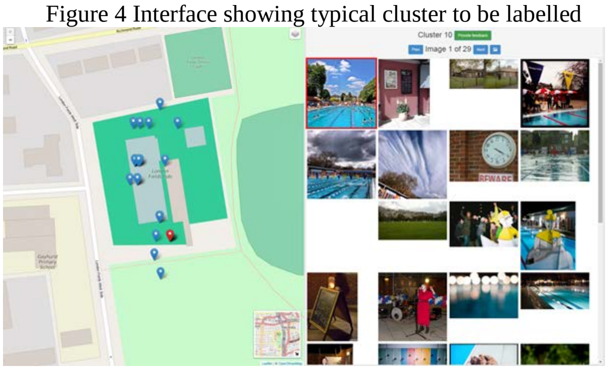
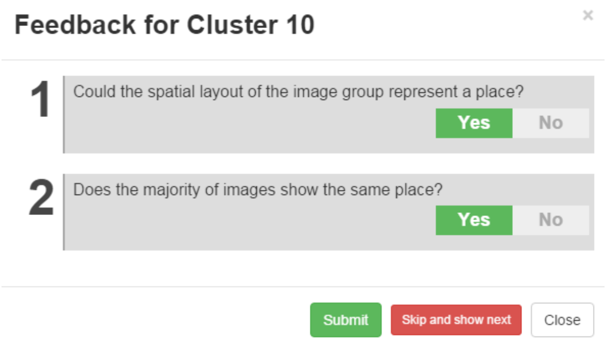


Figure 5 Supervisor feedback for a cluster



Feedback for Cluster 10

1 Could the spatial layout of the image group represent a place?

Yes No

2 Does the majority of images show the same place?

Yes No

Phase 6 assesses supervision results and removes noise. If the investigator(s) consider the results satisfactory and complete, non-place clusters can be removed as noise and the remaining stored in a UGGC place database. If not, the supervisor feedback leads to adjusting the clustering parameters. For the pilot study, we did not define a stable and satisfactory result. Instead, the first iteration used parameters that lead to geographically big clusters, and the second iteration parameters resulted in smaller (more compact) clusters. The aim was to account for the unknown and varying scale of places, and to test the supervisors’ feedback.

Phase 7 (Re-)trains a classifier to detect places from noise. The characteristics derived in phase 4 form the feature space for the ML algorithm to classify the clusters into “place” or “non-place”. For starters, we choose a simple J48 decision tree learner implemented in Weka[2], which has performed well on previous occasions (Spinsanti and Ostermann, 2013).

Phase 8 filter clusters using the classifier. If there is no or insufficient human supervision available, the system could fall back on a previously trained classifier to filter noise from the clusters without human supervision.

2.3 Results

Initially, we used as search query a bounding box that included the entire Greater London Area and resulted in metadata on 5,182,330 geo-referenced photos uploaded until December 2014. After building term vectors and filtering for the study area, 16632 items remained.

[2] http://www.cs.waikato.ac.nz/ml/weka/

The initial run used DBSCAN parameters of eps=0.0005 and a minimum number of 10 images per cluster. The resulting 77 clusters were then shown to the human supervisors (n=5, with some annotators skipping certain clusters where they felt not confident enough to provide feedback) using the web interface described in the previous section. As expected, there was some disagreement between the supervisors. The small number of supervisors allowed a simple majority vote (i.e. the most common answer is assigned to that cluster, with the cluster being dropped from further analysis in case of ties). The results of the first round of annotations are shown in Table 1. These results indicate that 55% of the clusters contain one or more possible places (categories A and C), and 45% contain mostly noise (category B).

Table 1 Frequency of annotator responses to first clustering (n = 77 clusters, majority vote in case of inter-rater disagreement, m = 5 annotators; x = 15 excluded if no majority vote available; Q = question)

Shape spatial cluster (Q1)	**Fre-quency**	**Places (Q2)**	**Fre-quency**
0 (correct)	42	A (one)	22
1 (wrong shape)	14	B (none)	28
2 (too big)	6	C (several)	12

To remove that noise, the supervisor labels were used to train a J48 classifier, using a 10-fold stratified cross-validation to estimate performance. Excluding ambiguous clusters (those without majority rater agreement), and using all features described under phase 4, the resulting classifier performance is estimated to correctly classify 71% of all instances. The average recall is 79% if we consider only Type II errors (false negatives) as clusters that contain one or more places but were classified as noise. Table 2 shows the full confusion matrix

Table 2 Combined confusion matrix of initial cluster classification

Majority Label	**Classified as**		
	A	**B**	**C**
One place (A)	17	3	2
Too much noise (B)	6	21	1
More than one place (C)	2	4	6

The supervision results indicated that some clusters consist of more than one place and cover too much area. Therefore, a second iteration used modified clustering parameters to allow for smaller clusters (eps = 0.0003, min = 5), resulting in 210 clusters (Figure 6 and Table 3).

The ratio of signal-to-noise remains similar: 52% of the clusters contain one or more places, and 48% contain mostly noise. Applying the original classifier to predict the second set results in an overall decrease of performance: only 49% are correctly classified, with a recall of only 45%, indicating a large number of false negatives (see also Table 4). Some performance degradation is to be expected, given that the original training data set was created with different clustering parameters.

Figure 6 Distribution of images in study area, colored according to cluster attribution

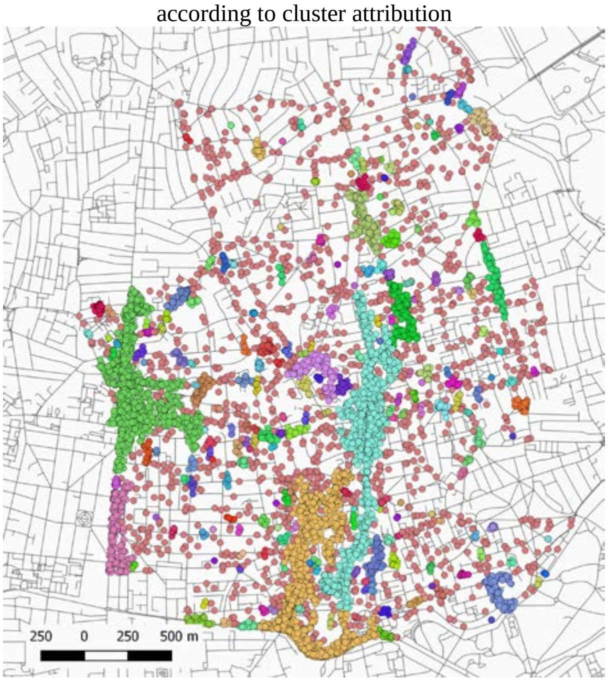


Table 3 Frequency of annotator responses to second clustering (n=210 clusters)

Shape spatial cluster (Q 1)	**Fre-quency**	**Places (Q 2)**	**Fre-quency**
0 (correct)	189	A (one)	91
1 (wrong shape)	11	B (none)	101
2 (too big)	10	C (several)	18

Table 4 Combined confusion matrix of second iteration using first classifier

Majority Label	**Classified as**		
	A	**B**	**C**
One place (A)	38	45	8
Too much noise (B)	28	63	10
More than one place (C)	1	15	2

However, training a new J48 model using the second dataset results in mixed performance: The performance estimation of stratified 10-fold CV is only 49.5% correctly classified instances, however with an improved recall (again categories A and C combined) of 66% (see Table 5).

Table 5 Combined confusion matrix of second iteration using second classifier

Majority Label	Classified as		
	A	B	C
One place (A)	58	29	4
Too much noise (B)	53	44	4
More than one place (C)	8	8	2

3 Discussion

An initial, purely unsupervised DM approach to detect places produced too many clusters, emphasizing the need for reduction to meaningful places through ML classification. The numerous instances to label, and the iterative character of the search for good hyper-parameters support our approach of crowdsourced validation and supervision. Despite the pilot study's limited scope, the approach should scale well: The computational costs were low, with clustering taking less than a minute of run-time on a common-off-the-shelf business laptop. All employed software is free and open source, and mature enough that a user with a moderate IT-skills can set up the prototype within few hours. Plugging-in different data sources or using other algorithms only requires few manual adjustments. This makes the approach suitable for citizen science projects not having a strong or dedicated computer or data science expertise. The web interface proved easy to navigate and work with. Supervisor labelling required less than a minute per cluster. However, temporal and resource constraints led to choosing a systems-centered design perspective, and annotator feedback indicated that the questions could be formulated clearer. We consider these limitations acceptable for an initial pilot study.

While the pilot study fulfilled its primary aim of demonstrating a feasible approach to hybrid geoinformation processing, its secondary aim of searching for meaningful places suffered from the clustering and classification performance, especially at a finer granularity. Although initial clustering was good enough to result in high inter-rater agreement, the number of false negatives (misclassified as non-place related) is too high. Further, a considerable share of images had either no assigned cluster or was part of large mega-clusters. Taking the temporal dimension into account (Birant and Kut, 2007) might help to detect ephemeral events and distinguish them from persistent features. Finding places through UGGC is a complex task, and many images have only very few and quite generic terms in their textual descriptions. We expect that more features for the classification and ancillary data from other UGGC sources or socio-demographic data from authoritative sources will improve the results.

4 Conclusions and outlook

In this paper, we presented a pilot study to demonstrate the feasibility of a hybrid approach to geoinformation processing. Using exclusively open source software and algorithms, it collects UGGC from a photo-sharing platform (Flickr), adds term vectors for semantic enrichment, and clusters it using the DBSCAN algorithm. The resulting clusters are presented in a web-interface that allows asynchronous validation by multiple human supervisors. The responses are used to improve clustering parameters and train a classifier to remove false positives.

The pilot study highlighted several issues that future research should address. (i) UGGC requires a system architecture that supports stream processing, e.g. flexible spatial and temporal bounding of the clustering (i.e. a single new UGGC should not trigger a re-clustering of the whole study area). (ii) The crowdsourced supervision needs a sustainable organization, so that more training sets can be labelled. An option is active learning, where the learning algorithm chooses which instances human annotators should label next, thereby maximizing the impact of human annotation and remaining flexible towards new instances. However, the type of learning might also depend on the expertise of the supervisors (Settles, 2009). Finally, recruiting annotators has to be a systematic and sustainable process, i.e. relying on research on establishing successful and lasting collaborative frameworks (Eveleigh et al., 2014).

Acknowledgements

We would like to thank the volunteers that contributed to the pilot study for their efforts, and the reviewers for their valuable comments that improved the paper.

References

Birant, D., Kut, A., 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data Knowl. Eng. 60, 208–221. doi:10.1016/j.datak.2006.01.013

Butler, D., 2013. When Google got flu wrong. Nature 494, 155–156.

Camponovo, M.E., Freundschuh, S.M., 2014. Assessing uncertainty in VGI for emergency response. Cartogr. Geogr. Inf. Sci. 41, 440–455. doi:10.1080/15230406.2014.950332

Craglia, M., Ostermann, F.O., Spinsanti, L., 2012. Digital Earth from vision to practice: making sense of citizen-generated content. Int. J. Digit. Earth 5, 398–416. doi:10.1080/17538947.2012.712273

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., Cox, A.L., 2014. Designing for Dabblers and Deterring Drop-outs in Citizen Science, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14. ACM, New York, NY, USA, pp. 2985–2994. doi:10.1145/2556288.2557262

Fast, V., Rinner, C., 2014. A Systems Perspective on Volunteered Geographic Information. ISPRS Int. J. Geo-Inf. 3, 1278–1292.

Granell, C., Ostermann, F.O., 2016. Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. Comput. Environ. Urban Syst. doi:10.1016/j.compenvurbsys.2016.01.006

Haworth, B., 2016. Emergency management perspectives on volunteered geographic information: Opportunities, challenges and change. Comput. Environ. Urban Syst. 57, 189–198. doi:10.1016/j.compenvurbsys.2016.02.009

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2008. Machine Learning Algorithms for GeoSpatial Data. Applications and Software Tools, in: Proceedings of the 4th Biennial Meeting of the International Congress on Environmental Modelling and Software

Morrow, N., Mock, N., Papendieck, A., Kocmich, N., 2011. Independent Evaluation of the Ushahidi Haiti Project. DISI - Development Information Systems International.

Purves, R., Edwardes, A., Wood, J., 2011. Describing place through user generated content. First Monday Vol. 16 Number 9 - 5 Sept. 2011.

Rousseeuw, P.J., 1987. Silhouettes - a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. J. Comput. Appl. Math. 20, 53–65.

Settles, B., 2009. Active Learning Literature Survey (Computer Sciences Technical Report No. 1648). University of Wisconsin, Madison.

Sigurbjörnsson, B., Van Zwol, R., 2008. Flickr tag recommendation based on collective knowledge, in: Proceedings of the 17th International Conference on World Wide Web. ACM Press, Beijing, China, pp. 327–336.

Spinsanti, L., Ostermann, F.O., 2013. Automated geographic context analysis for volunteered information. Appl. Geogr. 43, 36–44. doi:10.1016/j.apgeog.2013.05.005

Sui, D., Elwood, S., Goodchild, M.F. (Eds.), 2012. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Springer, Berlin.