Enhancing Crowdsourced Classification on Human Settlements Utilizing Logistic Regression Aggregation and Intrinsic Context Factors

Benjamin Herfort
GIScience Chair Heidelberg University
Im Neuenheimer Feld
Heidelberg, Germany
herfort@uni-heidelberg.de

Alexander Zipf
GIScience Chair Heidelberg University
Im Neuenheimer Feld
Heidelberg, Germany
zipf@uni-heidelberg.de

Abstract

Among semi-automated methods and pre-processed data products, crowdsourcing is another tool which can help to collect information on human settlements and complement existing data, yet it's accuracy is debated. Whereas the potential of crowdsourced datasets for training of machine learning algorithms has been explored recently, only few work has been done towards utilizing machine learning techniques to enhance the crowdsourcing workflow itself. In this research we investigated a novel approach that incorporates logistic regression to aggregate crowdsourced classification on human settlements from the MapSwipe app. For a case study containing 941,589 mapping tasks, we analysed to what degree such an approach can improve data quality utilizing intrinsic context factors such as user agreement, user characteristics and spatial characteristics of the results. The results have shown that a logistic regression based aggregation of crowdsourced classifications produced significantly higher quality data than common approaches that use soft majority agreement. The findings pronounce that the integration of machine learning techniques into existing crowdsourcing workflows can become a key point for the future development of crowdsourcing applications. However, regarding the limited geographic scope of this research, further validation of the automated classification and its transferability need to be addressed in future investigations.

Keywords: data quality, crowdsourcing, machine learning, logistic regression, classifications.

1 Introduction

Beside the works that focus solely on crowdsourcing and the analysis of the corresponding data quality, many researchers have highlighted the potential of crowdsourcing to support automated information extraction and thus analysed how crowdsourcing and machine learning can be combined. This nexus will be the focus of the following.

One of the first large scale crowdsourcing approaches to support image classification tasks was the online game Peekaboom (von Ahn, Liu and Blum, 2006). Through the online game users helped to annotate information about the type of object that is present in an image, where each object is located, and how much of the image is necessary to recognize it. The data derived function as training samples for a computer vision algorithm.

In the field of earth observation Gueguen et al. (2017) present a system which was developed at Digital Globe for village boundary detection at 50-meter resolution. The system uses machine learning for identifying potential villages from very high-resolution satellite imagery and validates the generated polygons using a crowdsourcing classification. Chen and Zipf (2017) use data generated by MapSwipe volunteers to classify chunks of satellite imagery. Their study demonstrates that volunteered geographic information can be successfully incorporated for building detection for humanitarian mapping in rural African areas. OpenStreetMap (OSM) data has attracted the interest of several researchers as well, since the database contains a myriad of training samples for image interpretation and computer vision algorithms. Keller et al. (2016) generated a training sample from OSM to detect crosswalks on satellite imagery. Hagenauer and Helbich (2012) use a machine learning approach to model unmapped residential areas in OSM. Their approach uses OSM data for training purposes. The Terrapattern team (Levin *et al.*, 2016) provides a visual search tool for satellite imagery. Their approach utilizes a deep convolutional neural network using areas where satellite images have been labelled in OSM.

However, only few work has been done towards utilizing machine learning techniques to enhance the crowdsourced datasets intrinsically. It is still not fully understood how automated classifiers could help to aggregate crowdsourced classifications in respect to user agreement, user characteristics and spatial characteristics. Since aggregation of single classifications has a great influence on overall data quality, more elaborated techniques incorporating the intrinsic context factors are much-needed. This work will therefore focus on the following research question:

RQ: To what degree can automated classifiers considering intrinsic context factors (user agreement, user characteristics and spatial characteristics) enhance data quality of aggregated crowdsourced classification?

The research question identified will be addressed in a case study including four MapSwipe projects in Laos. The following sections of this work will further describe the methods applied and datasets used.

2 Datasets

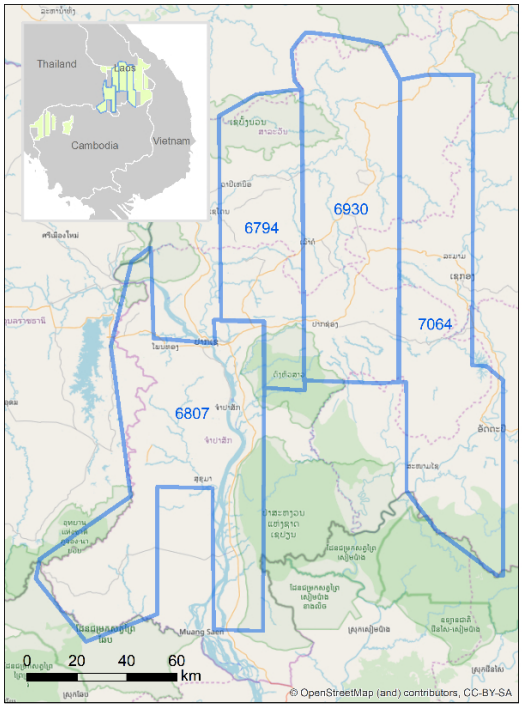
2.1 MapSwipe Dataset

This work focuses on crowdsourced data produced by volunteers using the MapSwipe app. A detailed perspective is chosen for a study region containing four projects in south west Laos with the following project IDs: 6807, 6794, 6930, 7064 (Figure 1). These projects are part of the Malaria Elimination Campaign organized by the Clinton Health Access Initiative and supported by the Humanitarian OpenStreetMap Team (HOT).

The MapSwipe crowdsourcing workflow is designed following an approach already presented by Albuquerque et al. (2016). Four concepts are important in the following: projects, groups, tasks and results. A more elaborated description of these concepts can be found in Herfort (2017).

Results contain information on the user classifications. However, only "Yes", "Maybe" and "Bad Imagery" classifications are stored as results. Whenever users indicate "No building" by just swiping to the next set of tasks, no data entry is created. Therefore, "No Building" classifications can only be modelled retrospectively. Initially, for user A all groups are selected, where this user submitted a result. For these groups all intersecting tasks are chosen in the second step. Finally, these tasks and the corresponding results are joined. All tasks where no classification result is obtained, will be marked as "No Building". This way of processing the data bears one limitation. Groups where user A classified all tasks as "No Building" cannot be considered, since they are not stored as results in the MapSwipe database. In total, 3,275,380 results by 1,534 users corresponding to 941,589 tasks are considered in this study.

Figure 1: Case Study MapSwipe Projects in Laos



2.2 OpenStreetMap Reference Dataset

The OSM reference dataset covers the extent of the selected MapSwipe projects and contains 324,152 individual buildings. The data was obtained from bbbike's planet.osm extracts in ESRI shapefile format.

To a great extent the OSM data was captured by HOT volunteers. The mapping efforts have been organised using the HOT Tasking Manager tool (Humanitarian OpenStreetMap Team, 2017). The area of interest corresponds to the following Tasking Manager project IDs: 3358, 3359, 3362, 3364, 3383, 3391, 3392, 3393, 3399 and 3400. All projects have been completely mapped and validated in the Tasking Manager. Thus, a first quality assurance was already applied. Since the Tasking Manager projects rely on aggregated and processed MapSwipe data, only built up areas that have been identified by the MapSwipe volunteers are considered for the detailed mapping in OSM. Due to this fact, this dataset may be more suited to assess the precision of the MapSwipe dataset towards detecting buildings rather than to assess its completeness and sensitivity. The OSM building data is intersected with the geometry of the MapSwipe tasks. For each MapSwipe task it is analysed whether the task contains at least one building.

3 Methods

A logistic regression model was utilized to assess how well tasks containing buildings (binary response) can be predicted. The model incorporates agreement characteristics (Scott's Pi, proportion of building classifications (building index), proportion of no building classifications (no building index)) as predictors. Furthermore, spatial characteristics (kernel density of no building classifications, kernel density of building classifications and kernel density of bad image classifications) were considered. For each task several different users with varying user characteristics contributed data. Therefore, the individual user characteristics were aggregated into single variables per task. In the study, user characteristics of each task were defined as the average user characteristics of all individual contributions of the same class ("no building", "building", "bad image"). For example, the average overall accuracy, no building precision and no building sensitivity were computed for each task using all no building classifications. Likewise, user characteristics were generated from building and bad image classifications. In the logistic regression model "no building average overall accuracy" (average of the overall accuracy for all users that classified as "no building" for this task), "building average building precision" (average of the building precision for all users that classified as "building" for this task) and "bad image average bad image precision" (average of the bad image precision for all users that classified as "bad image" for this task) were utilized. Missing values were imputed using the overall mean of each variable. A more elaborated description of these intrinsic indicators can be found in Herfort (2017). In the pre-processing variables were tested for independence and multicollinearity using a correlation matrix and by inspecting variance inflation factors (VIFs) (O'Brien, 2007).

In the second phase the tasks of the MapSwipe dataset were split up into training and testing samples. The fraction of the training sample was set to 0.3 which corresponds to circa 280,000 training samples. The samples were chosen randomly. Accordingly, about 660,000 tasks (70 %) of the dataset were used for testing.

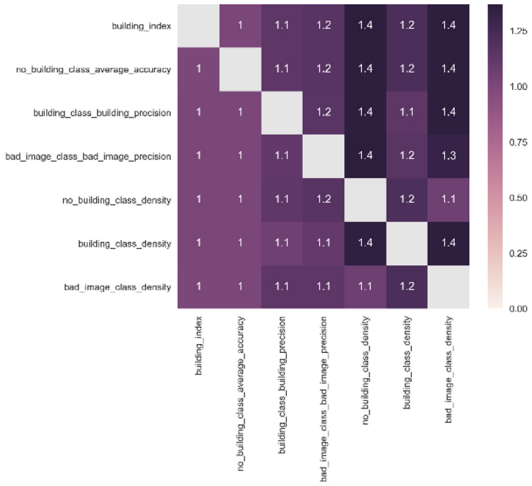
The performance of the logit-based aggregation was investigated in respect to overall accuracy, building precision, building sensitivity and building f1 score, which is the harmonic mean of building sensitivity and building precision. The data obtained from OSM functioned as a reference. We compared the results to a naïve aggregation method based on soft majority agreement. This method generates a classification from the several results for each task that have been submitted by different users by choosing the class that is present most often (e.g. when three out of five users classify as "no building", "no building" will be the aggregated class). If there are two classes with the same frequency, we classify as "building".

4 Results

In the first step, a logistic regression model was applied to test the impact of individual parameters. Initially, 15 different parameters describing agreement, user characteristics and spatial characteristics have been considered for the logistic regression analysis. After building the model and checking variables for multicollinearity and investigating variance inflation factors (VIFs) seven variables have been selected for the analysis. The correlation matrix plot reveals that for the chosen predictors no critical correlation between variables was observed (Figure 2). This was confirmed by the small VIFs close to 1.0 (Figure 3).

Figure 2: Correlation Matrix for Logistic Regression Input Variables

Figure 3: Variance Inflation Factors for Logistic Regression Input Variables



The results for the whole reference dataset containing 941,589 observations are presented in Table 1. The logistic regression model performed was statistically significant with $\chi^2(6) = 940{,}640$ and $p < 0.005$. The model explained 71.5 % of the variability in the crowdsourcing performance (Nagelkerke pseudo R^2). Increases in the building index and average building precision for building classifications were associated with a strong and significant increased probability of presence of buildings within the MapSwipe task. Less pronounced but still significant was the effect of building classification density. On the contrary, increases in the average no building average accuracy, average bad image precision, no building classification density and bad image classification density were associated with a significantly decreased likelihood of presence of buildings within the MapSwipe task.

Table 1: Results of the Logistic Regression Analysis

	Coeff.	StdEr	Sign.	Odds
Building Index	78.735	0.027	<0.005	2626.8337
Average Accuracy (No Building Results)	-84.139	0.076	<0.005	0.0002
Average Building Precision (Building Results)	62.469	0.051	<0.005	516.406
Average Bad Image Precision (Bad Image Results)	-14.723	0.059	<0.005	0.2294
No Building Class Density	-0.0148	0.001	<0.005	0.9853
Building Class Density	0.0919	0.001	<0.005	1.0962
Bad Image Class Density	-0.2360	0.005	<0.005	0.7897

Given the unbalanced distribution of building tasks in the dataset, it was no surprise that the logit-based aggregation method classified most tasks as "no building" (see Table 2). About 90 % of all tasks were assigned to this category. The same was observed for the soft majority aggregation (Table 3). The high proportion of correct no building classifications on the overall number of tasks was the main reason for the very high accuracy values obtained by both classifiers.

Table 2: Confusion Matrix for Logit Classifier

		Logit Classifier	
		No Bui	Bui
Ref	No Bui	588,483	1,514
	Bui	4,662	64,454

Table 3: Confusion Matrix for Soft Majority Aggregation

		Soft Majority Aggregation	
		No Bui	Bui
Ref	No Bui	584,735	5,262
	Bui	14,943	54,173

Using soft majority aggregation an accuracy of 96.7 % was reached, for the logit-based aggregation method an even higher value greater than 99 % was observed (Table 4). When looking at the building classifications the soft majority aggregation showed a considerable higher number of false positive classifications in comparison to the logit classifier. This was also reflected in the building precision scores (soft majority: 91.1 %, logit: 97.7 %).

The results for building precision and building sensitivity showed that logit-based aggregation outperforms soft majority aggregation in both directions and generated data with a better quality. This was expressed by the high value for f1 score. For soft majority aggregation, 84.3 % were obtained, whereas the logit derived a value of 95.4 %.

Table 4: Performance of Logit Classifier and Soft Majority Aggregation

	Soft Majority Agreement	Logit Classifier
Overall Accuracy	0.9693	0.9906
Building Sensitivity	0.7838	0.9325
Building Precision	0.9115	0.9770
Building F1 Score	0.8428	0.9543

5 Discussion and Conclusion

Machine learning based aggregation methods show potential to generate a high-quality settlement layer from crowdsourced MapSwipe data. The logistic regression model proved that the intrinsic characteristics of the dataset could explain the probability of correct building classifications. Nevertheless, the validity of the results of the logistic regression model need to be further evaluated towards a bias introduced by the imputation of missing values. Several authors (e.g. Donders et al. (2006), Greenland and Finkle (1995)) point out that simple techniques for handling missing data such as overall mean imputation used in this study can produce biased results. Future research should therefore consider more sophisticated replacement techniques for missing values such as multiple imputation.

Characteristics of the satellite imagery were not considered in this study. However, Chen and Zipf (2017) show recent advances of computer vision approaches for building detection from satellite imagery using neural networks. The potential of image analysis based on deconvolutional neural networks for human settlement mapping is also explored by Zhang et al. (2016). Including the characteristics of the satellite imagery could open further potential to improve the performance of the machine learning models, crowdsourcing workflow and resulting data quality.

The lack of reference data of sufficient quality limited the findings of this study. Since the reference dataset used in this study was derived from OSM, the data quality might vary given the large size of the examined area. The quality of OSM data has been investigated by many authors and spatial variations in data quality are well described (Ballatore and Zipf, 2015; Fonte *et al.*, 2015).

Although the OSM reference dataset was validated through the HOT mapping workflow, it cannot be guaranteed that all buildings are mapped, especially because MapSwipe data was already used to design the mapping projects. This can have implications regarding the obtained building sensitivity of both classifiers and needs to be evaluated further.

The logit-based aggregation outperformed the naïve aggregation method significantly regarding building precision and building sensitivity. However, the results describe the performance only for four selected MapSwipe projects. For the projects in Laos satellite imagery of very good quality was available, hence the quality of MapSwipe data in other parts of the world might be reduced. This will be also influenced by the experience of users involved. Given the global distribution of MapSwipe projects further validation of the automated classification and its transferability is needed.

The integration of machine learning methods into the aggregation of individual classifications has shown great potential to improve data quality. MapSwipe and other crowdsourcing applications should therefore build upon these initial findings. Thus, an integration of the explored machine learning techniques into the crowdsourcing workflow becomes a key point for the future development of crowdsourcing applications. This is not limited to the logistic regression analysis applied here, other methods such as support vector machines, regression trees should be tested in future investigations. Furthermore, also other crowdsourcing projects besides MapSwipe show the potential to incorporate machine learning techniques (e.g. for validating land use and land cover datasets).

Intelligent crowdsourcing approaches can dynamically derive data quality indicators to improve the task allocation process. For instance, for tasks reaching a high credibility no further classification should be obtained, whereas uncertain tasks should be repeated or validation should be prioritized. This could reduce the amount of required crowdsourced classifications while maintaining high quality. The setting bears great potential for features where fully automated techniques still fail to produce reasonable data quality. Slum mapping and slum type classification from satellite imagery might offer suitable challenges (Kuffer, Pfeffer and Sliuzas, 2016).

References

von Ahn, L., Liu, R. and Blum, M. (2006) 'Peekaboom: A Game for Locating Objects in Images', *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, p. 55. doi: 10.1145/1124772.1124782.

Albuquerque, J., Herfort, B. and Eckle, M. (2016) 'The Tasks of the Crowd: A Typology of Tasks in Geographic Information

Crowdsourcing and a Case Study in Humanitarian Mapping', *Remote Sensing*, 8(10), p. 859. doi: 10.3390/rs8100859.

Ballatore, A. and Zipf, A. (2015) 'A Conceptual Quality Framework for Volunteered Geographic Information', in Fabrikant, S. I., Raubal, M., Bertolotto, M., Davies, C., Freundschuh, S., and Bell, S. (eds) *Spatial Information Theory: 12th International Conference, COSIT 2015, Santa Fe, NM, USA, October 12-16, 2015, Proceedings*. Cham: Springer International Publishing, pp. 89–107. doi: 10.1007/978-3-319-23374-1_5.

Chen, J. and Zipf, A. (2017) 'DeepVGI: Deep Learning with Volunteered Geographic Information', *WWW '17 Companion: Proceedings of the 26th International Conference Companion on World Wide Web*, (1), pp. 771–772. doi: 10.1145/3041021.3054250.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T. and Moons, K. G. M. (2006) 'Review: A gentle introduction to imputation of missing values', *Journal of Clinical Epidemiology*, 59(10), pp. 1087–1091. doi: 10.1016/j.jclinepi.2006.01.014.

Fonte, C. C., Bastin, L., See, L., Foody, G. and Lupia, F. (2015) 'Usability of VGI for validation of land cover maps', *International Journal of Geographical Information Science*. Taylor & Francis, 29(7), pp. 1269–1291. doi: 10.1080/13658816.2015.1018266.

Greenland, S. and Finkle, W. D. (1995) 'A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses', *American Journal of Epidemiology*, 142(12), pp. 1255–1264. doi: 10.1093/oxfordjournals.aje.a117592.

Gueguen, L., Koenig, J., Reeder, C., Barksdale, T., Saints, J., Stamatiou, K., Collins, J. and Johnston, C. (2017) 'Mapping Human Settlements and Population at Country Scale from VHR Images', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2), pp. 524–538. doi: 10.1109/JSTARS.2016.2616120.

Hagenauer, J. and Helbich, M. (2012) 'Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks', *International Journal of Geographical Information Science*, 26(6), pp. 963–982. doi: 10.1080/13658816.2011.619501.

Herfort, B. (2017) *Understanding MapSwipe: Analysing Data Quality of Crowdsourced Classifications on Human Settlements*. Master Thesis, Heidelberg University. doi: 10.11588/heidok.00024257.

Humanitarian OpenStreetMap Team (2017) *HOT Tasking Manager*. Available at: https://tasks.hotosm.org/ (Accessed: 17 April 2018).

Keller, S., Bühler, S. and Kurath, S. (2016) 'Erkennung von Fußgängerstreifen aus Orthophotos', *AGIT Journal für Angewandte Geoinformatik*, pp. 162–166. doi: 10.14627/537622023.Dieser.

Kuffer, M., Pfeffer, K. and Sliuzas, R. (2016) 'Slums from Space — 15 Years of Slum Mapping Using Remote Sensing', *Remote Sensing*, 8(6), pp. 1–29. doi: 10.3390/rs8060455.

Levin, G., Newbury, D., McDonald, K., Alvarado, I., Tiwari, A. and Zaheer, M. (2016) *Terrapattern: Open-Ended, Visual Query-By-Example for Satellite Imagery using Deep Learning*. Available at: http://terrapattern.com.

O'Brien, R. M. (2007) 'A caution regarding rules of thumb for variance inflation factors', *Quality and Quantity*, 41(5), pp. 673–690. doi: 10.1007/s11135-006-9018-6.

Zhang, A., Liu, X., Tiecke, T. and Gros, A. (2016) 'Population Density Estimation with Deconvolutional Neural Networks', in *Workshop on Large Scale Computer Vision at NIPS 2016*.